

EMA MARIA DA SILVA BISCAIA

Licenciada em Matemática

MODELAÇÃO DA PROBABILIDADE DE PERSISTÊNCIA DE CADÁVERES DE AVES EM ESTUDOS DE MONITORIZAÇÃO DA MORTALIDADE EM LINHAS ELÉTRICAS

INFLUÊNCIA DE FATORES AMBIENTAIS
E VARIAÇÃO NO ESPAÇO

MESTRADO EM MATEMÁTICA E APLICAÇÕES

Universidade NOVA de Lisboa
Setembro, 2023

MODELAÇÃO DA PROBABILIDADE DE PERSISTÊNCIA DE CADÁVERES DE AVES EM ESTUDOS DE MONITORIZAÇÃO DA MORTALIDADE EM LINHAS ELÉTRICAS

INFLUÊNCIA DE FATORES AMBIENTAIS
E VARIAÇÃO NO ESPAÇO

EMA MARIA DA SILVA BISCAIA

Licenciada em Matemática

Orientadora: Regina Maria Baltazar Bispo

Professora Auxiliar, Universidade NOVA de Lisboa

Coorientadora: Joana Bernardino

*Investigadora, BIOPOLIS-CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos,
Universidade do Porto*

MESTRADO EM MATEMÁTICA E APLICAÇÕES

Universidade NOVA de Lisboa

Setembro, 2023

Modelação da probabilidade de persistência de cadáveres de aves em estudos de monitorização da mortalidade em linhas elétricas
Influência de fatores ambientais e variação no espaço

Copyright © Ema Maria da Silva Biscaia, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade NOVA de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objetivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Para os que voaram.

AGRADECIMENTOS

Esta dissertação teve um lugar especial no meu coração ao longo do ano que agora culmina com a sua entrega.

O meu primeiro agradecimento vai para as minhas duas incansáveis orientadoras Professora Regina e Joana, que me apoiaram nos momentos mais difíceis e festejaram comigo as vitórias que fomos alcançando. Às investigadoras Paula Moraga e Janet Van Niekerk que, para além de me inspirarem, me ajudaram a resolver muitos problemas.

Agradeço também à casa que me acolheu nestes 5 anos académicos, a FCT NOVA - Faculdade de Ciências e Tecnologia e o Departamento de Matemática que sempre tiveram um cantinho onde pude alcançar os meus objetivos. Ao NucM - Núcleo de Matemática, o meu companheiro de desafios, e à equipa da Direção que me acompanhou no cargo de Presidente, que tanto prezo por ter ocupado.

Não posso deixar de agradecer aos amigos que me acompanharam, que me tiraram todas as dúvidas loucas, que me ouviram e acalmaram. À minha incrível companheira deste percurso, Beatriz Jorge: não o teria palmilhado sem ti ao meu lado.

Por fim, o meu maior "Obrigado" para o meu maior alicerce nesta aventura: a minha querida família! Aos meus pais, Paula e Nuno, ao meu adorado irmão Henrique, ao meu querido namorado Vítor e às minhas tão estimadas avós, Lourdes e Graciosa, agradeço-vos infinitamente: não teria conseguido chegar aqui sem vocês, acreditem!

”

«*Take the time it takes, and it will take less time.*»

— *Pat Parelli*

(*Horseman*)

RESUMO

Em Portugal, as Linhas de Muito Alta Tensão, geridas pela REN – Redes Energéticas Nacionais, S.A., estão sujeitas a Planos de Monitorização cujo um dos objetivos é o de se estudar o impacto ambiental associado à mortalidade de aves por colisão com os cabos aéreos. Os planos incluem a quantificação de cadáveres de aves por baixo das linhas mas sabe-se que estes números diferem da mortalidade real, uma vez que os cadáveres podem ser removidos por predadores antes da sua contabilização. Assim, para considerar este fator de enviesamento, estes planos contemplam a realização de ensaios experimentais, designados por *testes de remoção*, que permitem obter dados sobre o tempo de persistência do cadáver no solo até à sua remoção, podendo estes ser dados censurados. A realização destes ensaios a cada nova linha elétrica implica, contudo, avultados custos financeiros e logísticos.

Neste âmbito, a "Cátedra REN em Biodiversidade" compilou informação sobre 30 ensaios, realizados em Linhas de Muito Alta Tensão em Portugal Continental, dando origem a dados geo-referenciados. Assim, neste estudo pretende-se construir um modelo que permita estimar, para qualquer ponto geográfico do território nacional, a probabilidade de persistência de um cadáver num determinado intervalo de tempo, obviando assim a realização de testes a cada novo projeto.

Para responder a este objetivo, foi usada a abordagem INLA (*Integrated Nested Laplace Approximation*) conjugada com a abordagem via equações diferenciais parciais estocásticas (*Stochastic Partial Differential Equations*) e com análise de sobrevivência, o que permitiu modelar a probabilidade de persistência de um cadáver considerando quer efeitos fixos (tamanho do cadáver e época do ano) quer efeitos aleatórios (localização geográfica e projeto).

Os resultados permitiram analisar a variação desta probabilidade para todo o território nacional e criar uma ferramenta para utilizadores comuns que estima a probabilidade de persistência, para uma localização específica, como função das covariáveis consideradas.

Palavras-chave: Análise de sobrevivência, Dados espaciais, INLA, Probabilidade de persistência

ABSTRACT

In Portugal, transmission power lines managed by REN - Redes Energéticas Nacionais, S.A., are subject to monitoring programmes whose one of the objectives is to assess the impact of bird mortality caused by collisions with overhead cables. These programmes include surveys to quantify bird carcasses under the lines, but these numbers are known to differ from actual mortality since carcasses can be removed by predators before they are counted. To account for this source of bias, these programmes also typically include the performance of field experiments, called *carcass removal trials*, which allow to obtain censored data on the persistence time of the carcass in the field until its removal. However, carrying out these trials on each new power line entails significant financial and logistical costs.

In this context, the "REN Chair in Biodiversity" compiled information on 30 trials, carried out on transmission lines in mainland Portugal, resulting in geo-referenced data. Therefore, in this study we intend to build a model that can estimate, for any geographical point of the national territory, the persistence probability of a carcass for a given time interval, thus eliminating the need for field trials in every new project.

To achieve this goal, the INLA (*Integrated Nested Laplace Approximation*) approach was used combined with the SPDE (*Stochastic Partial Differential Equations*) approach and survival analysis, which allowed to model the probability of persistence of a carcass considering both fixed effects (carcass size and season) and random effects (geographic location and project).

The results allowed to analyze the variation of this probability for the entire national territory and to create a tool for common users that estimates the persistence probability of a carcass, for a specific location, as a function of the covariates considered.

Keywords: INLA, Persistence probability, Spatial data, Survival analysis

ÍNDICE

Índice de Figuras	xvii
Índice de Tabelas	xix
Símbolos	xxi
1 Introdução	1
2 Estado da arte	3
2.1 Estimação da mortalidade de aves	3
2.2 Análise de Sobrevivência	5
2.3 Abordagem Bayesiana e INLA	8
3 Métodos	13
3.1 Análise de Sobrevivência	13
3.1.1 Censura	13
3.1.2 Funções de sobrevivência e de risco	15
3.1.3 Abordagem não paramétrica	16
3.1.4 Abordagem paramétrica	19
3.2 Abordagem Bayesiana e INLA	21
3.2.1 Integrated Nested Laplace Approximation	23
3.2.2 Modelos Espaciais	29
4 Aplicação ao caso de estudo	33
4.1 Dados relativos aos testes de remoção	33
4.2 Estimação das Funções de Sobrevivência	36
4.3 Modelação do Fator de Correção	39
4.3.1 Resultados para Portugal Continental	41
4.3.2 Resultados para localização específica	46
5 Conclusão	51

Referências Bibliográficas	55
Anexos	
I Estimador de Turnbull	59

ÍNDICE DE FIGURAS

3.1	Censura à direita, no contexto da dissertação.	14
3.2	Censura à esquerda, no contexto da dissertação.	14
3.3	Censura intervalar, no contexto da dissertação.	14
3.4	Valores da função base do ponto D.	30
4.1	Localização dos 35 projetos de linhas estudados e exemplos de troços monitorizados em três desses projetos.	33
4.2	Estimativas da função de sobrevivência, estratificada por estação do ano. . .	36
4.3	Estimativas da função de sobrevivência, estratificada por tamanho do cadáver. .	37
4.4	Localização dos 30 projetos das linhas onde foram realizados testes de remoção. .	39
4.5	Localização dos 33962 pontos para a previsão do fator de correção da remoção no território nacional.	41
4.6	Fator de correção médio para a estação "Primavera" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).	42
4.7	Fator de correção médio para a estação "Verão" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).	43
4.8	Fator de correção médio para a estação "Outono" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).	43
4.9	Fator de correção médio para a estação "Inverno" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).	44
4.10	Localização dos dois pontos específicos (Pombal - vermelho e Santarém - preto) para a previsão do fator de correção.	46
4.11	Curvas de sobrevivência associada a cada tamanho do cadáver para a estação "Primavera" (A), "Verão" (B), "Outono" (C) e "Inverno" (D), em Pombal. . .	49
4.12	Curvas de sobrevivência associada a cada tamanho do cadáver para a estação "Primavera" (A), "Verão" (B), "Outono" (C) e "Inverno" (D), em Foros de Salvaterra.	50

ÍNDICE DE TABELAS

3.1	Funções de risco instantâneo, de sobrevivência e densidade de probabilidade para as quatro distribuições.	20
4.1	Descrição das variáveis dos dados dos testes de remoção.	35
4.2	Resumo dos testes Log-Rank realizados para a comparação da sobrevivência entre as estações do ano.	38
4.3	Resumo dos testes Log-Rank realizados para a comparação da sobrevivência entre os tamanhos dos cadáveres.	38
4.4	Valores médios de DIC e WAIC para os quatro modelos espaciais AFT ajustados.	40
4.5	Valores médios do fator de correção e respetiva variabilidade.	45
4.6	Valores mínimos e máximos do coeficiente de variação do fator de correção.	46
4.7	Valores do fator de correção semanal médio e respetivos coeficientes de variação para Pombal e fator de correção semanal médio nacional.	47
4.8	Valores do fator de correção semanal médio e respetivos coeficientes de variação para Foros de Salvaterra e fator de correção semanal médio nacional.	48

SÍMBOLOS

A	Matriz de projeção do modelo SPDE (pp. 31, 40, 42)
C	Variável aleatória para o tempo de censura (pp. 14, 15)
c	Variável indicativa do tipo de censura (p. 15)
d	Número de falhas estimadas num determinado instante (pp. 16–18)
$f(t)$	Função densidade de probabilidade de T (pp. 15, 20)
G	Número de vértices na malha do modelo SPDE (pp. 30, 31)
$H(t)$	Função de risco acumulado (p. 15)
$h(t)$	Função de risco instantâneo (pp. 15, 19, 20)
j	Número de indivíduos em risco estimado num determinado instante (pp. 16–18)
k	Número de grupos de observações numa base de dados (pp. 6, 17, 18)
κ_{nm}	Peso da n -ésima observação no estimador Turnbull (pp. 17, 18, 60)
M	Número de instantes de tempo da grelha Ω (pp. 16–18)
N	Número de observações/localizações observadas numa base de dados (pp. 16–19, 21, 24–29, 31)
Ω	Grelha de instantes de tempo (pp. 16, 18, 60)
p	Probabilidade de falha num determinado instante (p. 17)
π	Função densidade de probabilidade definida no âmbito da teoria de Bayes (pp. 21–28)

η_i	Preditor linear de um modelo latente gaussiano associado à i -ésima observação (pp. 24, 31)
Q	Matriz de precisão de um campo aleatório gaussiano (pp. 24–27, 29)
r	Fator de correção da mortalidade associado à remoção de cadáveres do terreno (p. 4)
s	Conjunto das localizações observadas, dadas pelas coordenadas geográficas (pp. 29–31)
$S(t)$	Função de sobrevivência (pp. 4, 5, 15, 19, 20)
T	Variável aleatória para o tempo de sobrevivência (pp. 13–16, 19, 23, 31)
t	Valor da variável aleatória para o tempo de sobrevivência (pp. 13, 15)
T^*	Estatística de teste do teste Log-Rank adaptado (p. 18)
θ	Parâmetros desconhecidos de um modelo bayesiano (pp. 21, 22, 24–28)
U	Processo estocástico dos efeitos espaciais (p. 29)
V^*	Estimativa da variância da estatística de teste do teste Log-Rank adaptado (p. 18)
x	Vetor de efeitos latentes de um modelo latente gaussiano (pp. 24–28)

INTRODUÇÃO

Os impactes ambientais resultantes das atividades humanas são um tema cada vez mais recorrente na sociedade e precisam de ser estudados para poderem ser minimizados no futuro através do desenvolvimento de projetos mais sustentáveis. Neste sentido, em Portugal, o "Regime Jurídico da Avaliação de Impacte Ambiental (AIA)", regulamentado pelo Decreto-Lei n.º 151-B/2013, alterado e republicado pelo Decreto-Lei n.º 152-B/2017, estabelece que devem ser sujeitos a Estudo de Impacte Ambiental (EIA) e a monitorização após a sua construção, todos os projetos de Linhas de Muito Alta Tensão (≥ 110 kV) com mais de 10 km de extensão, ou todos os que atravessem áreas classificadas, como é o caso de áreas da Rede Natura 2000¹, independentemente da sua extensão. Assim, a REN, empresa que gere a rede de Linhas de Muito Alta Tensão em Portugal, realiza Avaliações de Impacte Ambiental e Programas de Monitorização nos seus diversos projetos espalhados pelo território nacional de modo a monitorizar os seus impactes nos valores ecológicos e, em particular, na avifauna. Entre os vários impactes monitorizados, a mortalidade de aves por colisão com os cabos aéreos é o impacte que suscita maior preocupação (Bernardino *et al.*, 2018; Martins *et al.*, 2023).

Nos Programas de Monitorização implementados pela REN, a quantificação da mortalidade de aves é, tradicionalmente, feita com base em prospeções regulares de cadáveres de aves ou de vestígios das mesmas, por baixo das linhas elétricas, com recurso a observadores humanos ou cães (CIBIO, 2020). No entanto, sabe-se que o número de cadáveres encontrados (mortalidade observada) não corresponde à verdadeira mortalidade, em virtude, por exemplo, da remoção desses cadáveres do terreno por predadores (Barrientos *et al.*, 2018). Para considerar este fator de enviesamento das prospeções de mortalidade no terreno, é necessário estimar a probabilidade de persistência de um cadáver no terreno, o que é feito realizando experiências de campo, vulgarmente denominadas *testes de remoção*, para o posterior cálculo de fatores de correção da mortalidade observada (Martins *et al.*,

¹"A Rede Natura 2000 é uma rede ecológica para o espaço comunitário da União Europeia (...) que tem como finalidade assegurar a conservação a longo prazo das espécies e dos habitats mais ameaçados da Europa, contribuindo para parar a perda de biodiversidade." «Instituto da Conservação da Natureza e Florestas - Rede Natura», 2023

2023). De forma geral, os testes de remoção consistem na distribuição aleatória de carcaças de aves de diferentes tamanhos pelo terreno, dentro da faixa de prospeção (debaixo da linha elétrica), em diferentes épocas do ano, sendo depois visitadas por um período de tempo pré-definido (por exemplo, visita ao 1º, 3º, 5º, 7º e 15º dia após a colocação) para determinar a sua persistência no terreno, ficando assim a conhecer-se para cada carcaça o último dia em que está presente e o primeiro dia em que já não se encontra no terreno (Bernardino *et al.*, 2022; Martins *et al.*, 2023). Até agora, a prática em Portugal tem sido implementar programas de monitorização de impactes que obrigam à realização de novos testes de remoção de cadáveres de aves sempre que uma nova linha elétrica da REN é construída (CIBIO, 2020; ICNB, 2010). Contudo, além da exigência logística e custos associados, a repetição destes testes de campo a cada novo projeto levanta também questões éticas, já que os mesmos implicam o sacrifício de um número considerável de animais.

A "Cátedra REN em Biodiversidade" (<https://www.catedraren-pt.com/>) é um acordo entre a REN, o CIBIO - Centro de Investigação em Biodiversidade e Recursos Genéticos da Universidade do Porto e a FCT - Fundação para a Ciência e Tecnologia que tem como objetivo desenvolver um programa de investigação científica e de transferência de conhecimentos, na área temática do impacto das linhas elétricas na biodiversidade. No âmbito deste programa, foram compilados os resultados de inúmeros testes de remoção realizados em projetos da REN desde 2003 (Bernardino *et al.*, 2022). Assim, questiona-se agora se estes dados poderão ser usados para estimar o fator de correção desta importante fonte de enviesamento nas estimativas de mortalidade de aves para futuras linhas elétricas em Portugal, evitando deste modo a realização de novos testes de campo a cada novo projeto.

Tendo em conta a vasta utilização de modelos estatísticos para a previsão de resultados ambientais, o objetivo desta dissertação é o da construção de um modelo preditivo do fator de correção associado à remoção de cadáveres de aves em linhas elétricas em Portugal, em função de algumas variáveis ambientais e tendo em consideração a localização geográfica do projeto. Em particular, pretende-se estudar a construção do modelo estimado no paradigma bayesiano pela abordagem INLA, com a inclusão de variáveis explicativas e da componente espacial.

Esta dissertação está organizada em quatro capítulos para além do Capítulo 1 (Introdução). No segundo Capítulo será feita a revisão da literatura, com a explicação do contexto e métodos utilizados por outros autores. De seguida, no Capítulo 3 serão apresentados os métodos estudados e utilizados nesta dissertação para criar o modelo em causa. No Capítulo 4 serão apresentados os dados e a sua caracterização e os vários passos metodológicos necessários para atingir os diferentes objetivos do estudo, para além de serem apresentados e discutidos os resultados obtidos. No último capítulo (Capítulo 5) apresenta-se a conclusão deste projeto e o trabalho futuro a desenvolver.

ESTADO DA ARTE

2.1 Estimação da mortalidade de aves

Esta dissertação insere-se no projeto "Cátedra REN em Biodiversidade", mencionado no Capítulo 1, que possui uma forte linha de investigação aplicada à melhoria do conhecimento sobre os impactos das Linhas de Muito Alta Tensão na biodiversidade, dando particular atenção à mortalidade direta de aves. Para melhor compreender esta problemática, foi inicialmente estudado o artigo de revisão de Bernardino *et al.* (2018), que teve o principal objetivo de compilar o conhecimento existente, a nível mundial, sobre a mortalidade de aves por colisão com os cabos. Os autores reviram 208 documentos e concluíram que o interesse por este tema começou nos anos 70 e que a maioria dos estudos, realizados na Europa e na América do Norte, focava-se em quantificar as fatalidades de aves por colisão. De seguida, apresentaram os fatores que influenciaram o risco de colisão, podendo estes eventos ser explicados pelas características das espécies (ex. morfologia, comportamento em voo), dos locais atravessados pelos projetos (ex. habitats, topografia) e das próprias linhas elétricas (ex. altura e diâmetro dos cabos). Por fim, o estudo identificou linhas prioritárias de investigação futura, entre elas, a necessidade de melhorar os métodos de estimação da mortalidade de aves causada por linhas elétricas.

O estudo de Barrientos *et al.* (2018) veio reforçar a ideia de que a mortalidade de aves observada nas prospeções no terreno, que se fazem para estimar as fatalidades causadas por infraestruturas, não corresponde à mortalidade real uma vez que os trabalhos de prospeção são afetados por dois importantes fatores de enviesamento: a probabilidade do observador não encontrar todos os cadáveres presentes no terreno e a probabilidade de os cadáveres desaparecerem antes de serem contabilizadas devido à sua remoção por predadores ou à sua decomposição. Assim sendo, os autores reviram 294 estudos com dados de testes de detetabilidade e de remoção, a nível mundial e considerando diferentes tipos de infraestruturas (linhas elétricas, parques eólicos, estradas, etc.), e fizeram uma análise através de Modelos Lineares com Efeitos Mistos das principais variáveis que afetam os referidos dois fatores de enviesamento. De salientar que, no caso da avaliação dos efeitos

na persistência dos cadáveres, a variável resposta utilizada correspondeu simplesmente à probabilidade de persistência ao final de 24h.

Bernardino *et al.* (2022) efetuou um estudo semelhante ao anterior, mas utilizando uma compilação de dados de testes de remoção de aves realizados em Portugal Continental em linhas elétricas da REN. Neste caso, os dados compilados correspondiam a tempos censurados da persistência dos cadáveres. Assim, e seguindo a proposta de Bispo *et al.* (2013) (ver Capítulo 2.2), os autores usaram Análise de Sobrevivência para o tratamento dos dados e ajustaram um modelo de Tempo de Vida Acelerado para testar a influência de variáveis ambientais, nomeadamente a estação do ano, o habitat e o tamanho do cadáver, nos tempos (censurados) de persistência dos cadáveres. Os autores concluíram que a maioria das variáveis ambientais comumente consideradas no desenho experimental dos testes de remoção influenciam, de facto, a persistência dos cadáveres no terreno. No entanto, indicaram que não serão suficientes para explicar a variação nos tempos de persistência, pelo que os modelos ajustados não serão os mais adequados para calcular valores de referência para as probabilidades de persistência, em locais específicos de futuros projetos de linhas elétricas.

Aprofundando os métodos de estimação da mortalidade real, Bernardino *et al.* (2013) apresentou uma vasta coletânea de estimadores existentes até à data. Todos os estimadores apresentados incluem fatores de correção da remoção na mortalidade observada durante as prospeções de cadáveres, embora os fatores possam ser incorporados com abordagens estatísticas distintas. Por exemplo, Jain *et al.* (2007) propôs um estimador que, numa simplificação da realidade, assume que o fator de correção da remoção corresponde à probabilidade de persistência de um cadáver uma vez decorrido, aproximadamente, metade do tempo entre as prospeções de cadáveres. Por exemplo, no caso de uma linha com prospeções com frequência semanal e assumindo que um evento de mortalidade tem igual probabilidade de ocorrer em qualquer um dos dias entre prospeções, o tempo decorrido entre o evento de colisão da ave com a linha e a deteção desse cadáver pelo observador na prospeção realizada ao 7º dia é, em média, de $(1 + 2 + 3 + 4 + 5 + 6)/6 = 3,5$ dias (aproximadamente quatro dias). Assim, em prospeções semanais, o estimador de Jain *et al.* (2007) assume que o fator de correção da remoção corresponde, simplesmente, à probabilidade de persistência dos cadáveres ao 4º dia.

Desde a revisão de Bernardino *et al.* (2013), foram propostos vários novos estimadores, sendo o GenEst, desenvolvido por Dalthorp *et al.* (2018), o estimador mais atual, porém mais complexo para estimar a mortalidade de aves. No que respeita ao ajuste do enviesamento causado pela remoção de cadáveres entre prospeções, o estimador GenEst adota o fator de correção proposto por Bispo *et al.* (2013), sendo o mesmo definido pela expressão:

$$r_k = \frac{1}{I} \int_0^I S(t)_k dt, \quad (2.1)$$

em que I representa o intervalo de tempo entre prospeções consecutivas e $S(t)_k$ corresponde à função de sobrevivência do k -ésimo grupo de observações.

2.2 Análise de Sobrevivência

Dada a natureza dos dados desta dissertação, neste capítulo foi realizada uma revisão das técnicas de Análise de Sobrevivência utilizadas na literatura, em particular em estudos científicos com aplicações a casos reais. Numa perspectiva não-paramétrica, os métodos mais tradicionais foram desenvolvidos para o tipo de censura mais comum que é a censura à direita. Contudo, os dados compilados e estudados neste projeto contêm censura intervalar, explicada no Capítulo 3. Assim tornou-se fundamental pesquisar as técnicas adequadas a este tipo de dados de sobrevivência.

Rodrigues *et al.* (2018) discutiu a importância de aplicar os métodos estatísticos corretos na estimação da sobrevivência em dados com censura intervalar, realizando um estudo de simulação, e avaliou o impacto de ignorar esses mecanismos em dados reais sobre cancro no Brasil. Para o estudo de simulação, os autores começaram por apresentar a estrutura do estimador de Turnbull, que tem em consideração a censura intervalar (Turnbull, 1976), descrito no Capítulo 3, afirmando que é um método análogo ao estimador de Kaplan-Meier, usado para dados com censura à direita, e que é igualmente de fácil interpretação, e geraram dados com este tipo de censura para quantificar o erro entre as três abordagens que aplicaram para a estimação da função de sobrevivência:

- A. usar o estimador de Turnbull, adequadamente;
- B. usar o estimador de Kaplan-Meier, assumindo que o evento ocorreu no instante final do intervalo de tempo;
- C. usar o estimador de Kaplan-Meier, assumindo que o evento ocorreu no ponto médio do intervalo de tempo em que foi registado.

Nos 75 cenários criados, fazendo variar o tamanho da amostra, a percentagem de censura nos dados e a amplitude dos intervalos de tempo, a medida de erro indicou sempre que o estimador de Turnbull tinha o melhor desempenho, nomeadamente à medida que a percentagem de censura aumentava, e que a abordagem do ponto médio tinha melhores resultados entre as duas abordagens com o estimador de Kaplan-Meier.

Além disso, na aplicação das três abordagens a duas bases de dados de pacientes do A.C. Camargo Cancer Center no Brasil, constatou-se, para ambas, que as abordagens B. e C. apresentaram sempre estimativas enviesadas da função de sobrevivência. Consequentemente, os autores defenderam a importância de usar métodos adequados uma vez que

(...) researchers and analysts tend to apply traditional survival methodologies because they are easier and well-known, or because not all statistical softwares have procedures for analyzing interval-censored data. (...) However, usual survival analysis methods overestimate the survival function and this can lead to erroneous conclusions.

Por sua vez, Sun (1996) afirmou que o estudo da comparação de funções de sobrevivência entre k grupos de observações não está devidamente desenvolvido para dados intervalares pelo que tinha como objetivo demonstrar a validade do teste Log-Rank generalizado, que propôs no artigo. Para isso, o autor aplicou-o a dados relativos a uma coorte de pacientes com SIDA, analisados anteriormente por outros investigadores. O teste mostrou-se satisfatório e confirmou analiticamente a diferença entre os dois grupos de pacientes que apenas tinha sido deduzida pela comparação gráfica das funções de sobrevivência pelos anteriores investigadores. O teste proposto, que é apresentado no Capítulo 3, é especialmente indicado para dados discretos de sobrevivência, isto é, para dados que são registados em instantes de tempo pré-definidos, como é o caso dos dados desta dissertação e de dados clínicos, por exemplo. O autor realçou a existência de um teste semelhante que, no entanto, foi desenvolvido para dados de sobrevivência contínuos e assume uma distribuição para o tempo de vida, ao contrário do teste proposto que é não-paramétrico. Além disso, para dados discretos, os métodos discretos como este novo teste são mais convenientes e preferíveis.

Uma alternativa aos modelos de sobrevivência não paramétricos são, por exemplo, os modelos totalmente paramétricos, em que se assume uma distribuição probabilística para os tempos de sobrevivência. Bispo *et al.* (2013) propôs a utilização de modelos de sobrevivência paramétricos para lidar com os dados resultantes dos testes de remoção de cadáveres de aves, no âmbito de estudos de estimação da mortalidade associada a diferentes infraestruturas. Utilizando uma base de dados de tempos de persistência de cadáveres em 10 parques eólicos portugueses, os autores do estudo compararam modelos de Tempo de Vida Acelerado com diferentes distribuições e estudaram o impacto que os diferentes modelos tiveram na estimativa da mortalidade de aves. As distribuições estudadas foram:

- Exponencial, por ser a mais comum no estudo da mortalidade em parques eólicos;
- Weibull, com o parâmetro α menor que 1, por representar um risco de remoção do cadáver decrescente ao longo do tempo, que é a situação mais plausível para o tipo de dados;
- Log-Logística e Log-Normal por acomodarem funções de risco que são inicialmente crescentes e depois decrescentes.

Os autores usaram este tipo de modelos paramétricos para poderem incluir as duas covariáveis dos dados, a estação do ano da prospeção e o tamanho do cadáver. No entanto,

o local dos parques eólicos não foi tido em conta como variável explicativa apesar de os autores mencionarem que outros investigadores realçaram o interesse e importância desse tipo de estudo, pelo que a análise foi realizada em separado para os 10 locais. Para estimarem a mortalidade de aves, os autores propuseram o fator de correção da Equação (2.1), indicando que a função de sobrevivência deveria ser estimada a partir do modelo com melhor ajuste aos dados. Para compararem os modelos anteriormente mencionados, usaram os valores de AIC bem como os gráficos de resíduos de Cox-Snell.

De acordo com os gráficos de Cox-Snell, todos os quatro modelos se ajustaram de forma adequada para os 10 parques eólicos. Avaliando os valores de AIC, para a maioria dos parques eólicos o modelo com a distribuição Log-Normal teve o melhor ajuste, seguido do modelo com a Log-Logística e por fim pela Weibull. A distribuição Exponencial teve sempre o pior ajuste aos dados. De forma a realçar o impacto da escolha do modelo mais adequado, os autores compararam os valores do fator de correção e concluíram que usando outros modelos que não o melhor ajustado, a mortalidade era estimada incorretamente, consideravelmente acima ou abaixo do esperado.

Por fim, os autores concluíram que a abordagem paramétrica era a mais vantajosa por comparação a uma abordagem semi-paramétrica, que as covariáveis eram significativas e afetavam as estimativas da mortalidade e que o estudo do ajuste dos modelos aumentava a credibilidade deste tipo de resultados.

Tobin e Dolbeer (1990) referiram que a aplicação de pesticidas em pomares tem um efeito preocupante na mortalidade de aves e para a avaliarem realizaram um estudo conceptualmente semelhante aos programas de monitorização implementados pela REN, através de prospeções em pomares de Nova Iorque para procurar os cadáveres, o que podia ser enviesado pela remoção por predadores ou pela falha na deteção por parte dos observadores. Assim, com o objetivo de monitorizar a taxa de persistência de cadáveres de aves em pomares e avaliar a eficiência das prospeções feitas no terreno, colocaram-se carcaças de aves em dois tipos de pomares que eram verificadas a cada 1-3 dias, durante 11-12 dias até desaparecerem ou se decomporem. Os autores usaram um modelo de Tempo de Vida Acelerado que ajustou uma distribuição Weibull ao tempo de persistência no terreno para estimar os tempos de sobrevivência, comparar os tipos de pomares e ainda os efeitos do estado do solo. O estudo permitiu corroborar os resultados da estimação de sobrevivência com o de outros investigadores e demonstrou que havia uma variação significativa dos tempos de sobrevivência entre os dois tipos de pomar mas o estado do solo não influenciava a sobrevivência. Foi ainda observado que a maioria dos cadáveres era removida nos primeiros dias após a colocação. No fim, os autores concluíram que prospeções cuidadosas nos primeiros dois dias após a aplicação dos pesticidas deveriam ser suficientes para detetar corretamente a mortalidade das aves.

Por outro lado, Lonsinger *et al.* (2016), para além de usar uma metodologia semelhante à dos dados desta dissertação, aplicou também um fator de correção. Neste artigo, foi inicialmente explicado que a monitorização de populações de carnívoros selvagens é feita com base em prospeções de excrementos ao longo de estradas e trilhos. Apesar de estas prospeções serem um método fiável para o controlo das populações, a remoção dos excrementos por fatores antropogénicos é algo problemático. Os autores tinham como objetivo explorar os fatores com maior influência na remoção dos excrementos e aplicar um fator de correção associado à taxa de persistência na avaliação da abundância relativa de várias espécies de carnívoros no Utah. Para isso, foram colocados 360 excrementos em locais geo-referenciados, em estradas e trilhos, que foram verificados ao longo de 42 dias, em dias específicos, de modo a registar o tempo até à sua remoção. Como tal, os dados possuíam censura intervalar e também censura à direita (ver Capítulo 3). Os autores ajustaram um modelo de Tempo de Vida Acelerado para analisar o efeito de variáveis como a espécie, a estação do ano, o tipo de estrada, o tipo de local e o trânsito na remoção dos excrementos e utilizaram a medida AIC para decidir entre as quatro distribuições (Exponencial, Weibull, Log-Logística e Log-Normal) para o tempo de vida. O modelo Exponencial teve o melhor ajuste, imediatamente seguido pelo modelo Weibull e as variáveis da espécie, tipo de estrada, tipo de local e trânsito mostraram-se significativas. Assim, os autores concluíram que os excrementos de animais mais pequenos, as estradas maiores e o maior volume de trânsito implicavam uma maior taxa de remoção dos excrementos. Além disso, após analisarem a abundância relativa de carnívoros em 15 troços de estradas, os autores definiram o fator de correção associado à remoção como a proporção de excrementos que persistem durante um período de tempo. Esta proporção foi prevista a partir das estimativas de máxima verosimilhança dos parâmetros do modelo, aplicadas à função de sobrevivência Exponencial. Ao aplicarem este fator de correção, os autores constataram que os resultados de abundância se alteraram significativamente.

2.3 Abordagem Bayesiana e INLA

Os métodos utilizados nesta dissertação baseiam-se no paradigma bayesiano da análise do tempo de vida, permitindo incluir conhecimento subjetivo e intuitivo sobre os dados. A estatística bayesiana tem evoluído particularmente nos últimos anos devido aos avanços computacionais que permitem ultrapassar as suas dificuldades de implementação. Neste sentido, foram estudados artigos que usam as mais recentes técnicas bayesianas no contexto da sobrevivência e inspiraram o estudo dos dados desta dissertação.

Martino *et al.* (2011) começou por afirmar que os métodos bayesianos permitem uma inferência *a posteriori* exata e completa de qualquer parâmetro ou previsão de uma variável de interesse, baseando-se habitualmente na técnica MCMC (Markov chain Monte Carlo) que requer um elevado esforço computacional, é exigente em termos de tempo e que se pode revelar demasiado complexa para o utilizador final da análise. Os autores afirmaram

que tem havido um crescente interesse por modelos bayesianos mais avançados, como por exemplo modelos com componente espacial, e que vários investigadores já tentaram estudar dados de sobrevivência espaciais, usando a técnica MCMC. Sabendo que é possível aplicar a estatística bayesiana a modelos de sobrevivência, este artigo teve como objetivo mostrar que vários modelos de sobrevivência podem ser vistos como modelos latentes Gaussianos, em que é bastante simples adicionar efeitos aleatórios ou componentes espaciais, e que a inferência bayesiana neste tipo de modelos pode ser realizada com uma ferramenta chamada INLA (*Integrated Nested Laplace Approximations*), cujas aproximações são mais rápidas e precisas que a anterior abordagem MCMC.

Os autores deste artigo começaram por analisar um caso comum de sobrevivência, nomeadamente dados de recorrência de infeção em pacientes com doença renal, que eram censurados à direita. Assumindo os pressupostos das três covariáveis seguirem uma distribuição Normal e de o efeito aleatório associado a cada paciente seguir uma distribuição Log-Normal (uma vez que há duas observações por doente), os autores justificaram que o modelo Weibull ajustado é um modelo latente Gaussiano e compararam os resultados do estudo utilizando a abordagem INLA e a MCMC. Inicialmente, fizeram o ajuste do modelo com distribuições *a priori* do parâmetro α da Weibull e do efeito aleatório adequadas ao caso em estudo e concluíram que, apesar de as estimativas serem praticamente indistinguíveis, o tempo de computação do INLA foi de apenas dois segundos, enquanto o do MCMC foi de cinco minutos. De seguida, usaram distribuições *a priori* pouco informativas, sendo este caso mais complexo na abordagem INLA, e constataram que, apesar de as estimativas apresentarem algumas discrepâncias, as medidas de avaliação analíticas e gráficas permitiram ignorá-las. Nesta situação, o INLA teve o mesmo tempo de computação, enquanto o MCMC aumentou para o dobro. Ao longo do artigo, Martino *et al.* (2011) apresentaram outros exemplos, nomeadamente de modelos semi-paramétricos com efeitos espaciais e de dados com censura intervalar, em que conseguiram justificar facilmente a utilização da abordagem INLA e onde obtiveram resultados satisfatórios, comparativamente com a abordagem MCMC. Por fim, concluíram que esta nova ferramenta de inferência estatística bayesiana é rápida, simples e intuitiva para o utilizador final, podendo ser aplicada em múltiplos casos de dados de sobrevivência.

Bivand e Gómez-Rubio (2021) estudaram os dados sobre a reabertura de negócios após a passagem do furacão Katrina em três ruas de Nova Orleães na perspetiva de sobrevivência, e utilizou as técnicas INLA e MCMC para comparar modelos não-espaciais e espaciais em relação ao estudo na perspetiva binária dos mesmos dados, realizado por outros investigadores. Na perspetiva de sobrevivência, os dados registavam o tempo até à reabertura da loja, entre o 41º e o 449º dia após a passagem do furacão, pelo que possuíam observações censuradas à esquerda e à direita (ver Capítulo 3), enquanto na perspetiva binária a variável resposta dos dados era apenas a indicação de reabertura ou não da loja durante esse período de tempo de observação.

Os autores ajustaram modelos Weibull de Tempo de Vida Acelerado, com três variáveis categóricas e, para o INLA, usaram as distribuições *a priori* pré-definidas (tal como no trabalho desenvolvido nesta dissertação). A componente espacial, dada pelas coordenadas das lojas, foi modelada através de uma estrutura de vizinhança entre os locais observados. Os modelos de sobrevivência não-espaciais ajustados incluíam ainda efeitos aleatórios independentes e identicamente distribuídos associados a cada rua. A abordagem INLA mostrou ser mais adequado para os modelos espaciais, enquanto as análises análogas feitas com a abordagem MCMC privilegiaram os modelos de sobrevivência, apenas com os efeitos aleatórios. Comparando os modelos de sobrevivência com os modelos ajustados aos dados na perspectiva binária, ajustados por outros investigadores, chegou-se à conclusão de que as diferenças na inferência não são relevantes, mas o estudo de sobrevivência neste tipo de dados será, ainda assim, vantajoso uma vez que respeita a natureza do registo do tempo até acontecer o evento de interesse. Entre a abordagem INLA e o MCMC, os modelos de sobrevivência propostos (espaciais e não-espaciais) apresentaram estimativas dos coeficientes das covariáveis semelhantes e os autores propuseram a sua avaliação, no futuro, através da eficácia preditiva dos mesmos.

As estruturas espaciais que a abordagem INLA permite estabelecer podem ser definidas de variadas formas. Assim, um outro exemplo é o estudo de Seppä *et al.* (2019), no qual foi avaliado o desempenho do método na variação regional entre distritos e entre municípios da sobrevivência ao cancro dos ovários e da mama, na Finlândia, entre 1960 e 2014. Os dados diziam respeito ao número de mortes por município, e apesar de não ser considerado nos modelos ajustados, as observações eram censuradas após cinco anos de acompanhamento uma vez que o risco de mortalidade se tornava muito baixo. Os autores ajustaram um modelo Poisson para o número de mortes observado, incluindo os distritos como efeitos aleatórios independentes e identicamente distribuídos e os municípios associados a uma estrutura de vizinhança. Esta estrutura espacial era diferente daquela usada no artigo anterior (sobre o furacão Katrina) e definia que municípios adjacentes no mesmo distrito são vizinhos. Os autores apresentaram a variação regional em termos da sobrevivência, o que requer a extração de amostras conjuntas das distribuições *a posteriori* de todos os parâmetros do modelo, algo em que a abordagem INLA foi bastante preciso. Além disso, a correlação espacial da sobrevivência entre municípios foi considerada importante e os autores afirmaram que a extensão do período de acompanhamento poderá ter impacto na variação regional. Novamente, esta abordagem foi bastante mais rápida do que o MCMC, pelo que os autores recomendaram-na para análises de rotina da variação regional da mortalidade e da sobrevivência.

Um dos artigos fulcrais desta dissertação foi o de Kang *et al.* (2018), que tinha como objetivo criar mapas contínuos da prevalência de malária em Madagáscar. Para isso, os autores usaram dados de três inquéritos sobre indicadores de malária (Malaria Indicator

Survey - MIS) que mediram a prevalência de infecções na ilha e algumas variáveis ambientais e socio-demográficas em 898 localizações geo-referenciadas, em cerca de 20 000 crianças, nos anos de 2011, 2013 e 2016. Seja pelo tipo de dados ou pelos métodos utilizados, explicados de seguida, este artigo serviu de guia ao estudo desenvolvido nesta dissertação.

Kang *et al.* (2018) ajustou modelos de regressão logística com efeitos aleatórios espaço-temporais, ou seja, considerando um processo latente espacial a variar com o tempo e usou a abordagem INLA para a inferência e posterior previsão. Os efeitos aleatórios foram modelados via equações diferenciais parciais estocásticas (Stochastic Partial Differential Equations - SPDE), que representa um campo Matérn espaço-temporal Gaussiano como um campo aleatório de Markov Gaussiano, via triangulação, que foi exatamente o processo utilizado nesta dissertação. Foi ainda utilizada a medida DIC (Deviance Information Criterion) para a comparação de modelos e uma grelha de previsão de 1km×1km de resolução espacial. Os autores obtiveram as previsões mensais médias da prevalência e a respetiva amplitude interquartis, como medida de incerteza, para cada mês dos três anos, que foram resumidas em mapas das previsões anuais médias da prevalência e da amplitude interquartis. As medidas de avaliação justificaram o bom ajuste do modelo aos dados observados e as medidas de avaliação da eficácia preditiva consideraram o modelo preciso. Além disso, foram calculados os coeficientes de variação para refletir a variabilidade relativa das previsões.

Os autores concluíram que as zonas mais ricas em registos dos MIS tiveram previsões com maior grau de confiança, isto é, com menor amplitude interquartil, enquanto as zonas com informação mais heterogênea tinham maior amplitude nas previsões. Verificou-se também que os mapas dos três anos apresentavam diferenças significativas, levando à análise temporal e espacial da tendência de prevalência, o que contribuiu, por fim, para a tomada de decisões com o objetivo de erradicar esta doença endémica da ilha.

Em resumo, esta pesquisa bibliográfica contemplou o estudo de artigos que lidam com o tempo de sobrevivência até acontecer determinado evento, dando o enquadramento daquilo que se tem desenvolvido no âmbito da motivação desta tese e que é explorado, de seguida, na apresentação da teoria de Análise de Sobrevivência. Além disso, reviram-se artigos que estudaram dados com as localizações desses eventos, abordando o tema dos Modelos Espaciais que será também explicado no Capítulo 3. Por fim, apresentaram-se artigos que conjugam estas duas teorias aplicadas através da abordagem INLA (*Integrated Nested Laplace Approximation*) conjugada com a abordagem via equações diferenciais parciais estocásticas (*Stochastic Partial Differential Equations*) e que também é apresentado no próximo capítulo.

MÉTODOS

Neste capítulo descreve-se a teoria de Análise de Sobrevida e os Modelos Espaciais aplicados através da abordagem INLA (*Integrated Nested Laplace Approximation*) conjugada com a abordagem via equações diferenciais parciais estocásticas (*Stochastic Partial Differential Equations*).

3.1 Análise de Sobrevida

O estudo de ocorrências ou eventos, envolvendo o tempo até ao seu acontecimento, é algo que tem vindo a ser desenvolvido em áreas tão diversas como a Economia, o Ambiente e principalmente a Medicina (Rocha e Papoila, 2009). A este tipo de estudo dá-se o nome de Análise de Sobrevida uma vez que é um conjunto de procedimentos de análise de dados em que a variável de interesse é o tempo até à ocorrência de um determinado evento, chamado de *falha* (Kleinbaum e Klein, 2012). Por outras palavras, e no contexto desta dissertação, a variável de interesse é o tempo de sobrevida de um cadáver de ave até acontecer a falha, que é a sua remoção do terreno por um predador. Outros exemplos de situações em que se pode aplicar este tipo de análise são os seguintes:

- tempo até um paciente contrair uma dada doença;
- período de desemprego de um contribuinte;
- tempo até um aparelho industrial avariar.

O tempo de sobrevida será denotado pela variável aleatória T e um instante específico desta variável será dado por t .

3.1.1 Censura

Por diversas razões, pode não se observar exatamente o instante em que ocorre o evento em estudo¹ e desse modo apenas se obtém informação parcial sobre o tempo

¹Em casos clínicos, um paciente pode não contrair a doença durante o tempo de observação ou, no âmbito desta dissertação, o cadáver de ave pode ser removido após o fim do teste, por exemplo.

de sobrevivência. Neste caso, diz-se que os dados são censurados e define-se a variável C como o tempo de censura. Além disso, a censura pode ser classificada em três tipos, segundo Kleinbaum e Klein (2012):

- censura à direita: acontece quando apenas se sabe que a falha ocorreu após um determinado instante, ou seja, o verdadeiro tempo de sobrevivência é maior ou igual ao tempo de censura observado, $T \geq C$ (Figura 3.1);

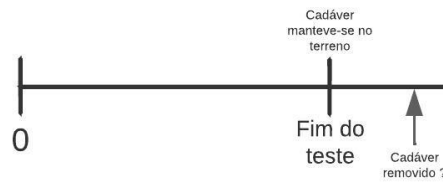


Figura 3.1: Censura à direita, no contexto da dissertação.

- censura à esquerda: acontece quando se sabe que a falha ocorreu antes de um determinado instante de tempo, isto é, o verdadeiro tempo de sobrevivência é menor ou igual ao tempo de censura, $T \leq C$ (Figura 3.2);

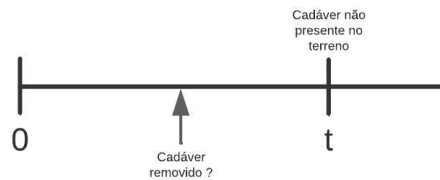


Figura 3.2: Censura à esquerda, no contexto da dissertação.

- censura intervalar: acontece quando se sabe que a falha ocorreu num intervalo de tempo mas não se sabe o instante exato, ou seja, o verdadeiro tempo de sobrevivência está compreendido num determinado intervalo de tempos de censura, $C_1 \leq T \leq C_2$ (Figura 3.3).

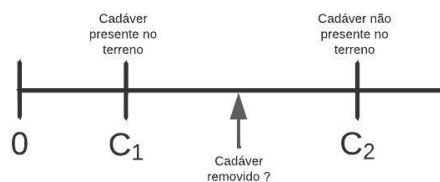


Figura 3.3: Censura intervalar, no contexto da dissertação.

Assim, pode definir-se uma variável c que indica o tipo de censura de uma dada observação:

$$c = \begin{cases} 0 & \text{se é censura à direita} \\ 1 & \text{se é observada a falha} \\ 2 & \text{se é censura à esquerda} \\ 3 & \text{se é censura intervalar} \end{cases} \quad (3.1)$$

Além disso, as censuras à esquerda e à direita são apenas casos particulares da censura intervalar, em que $C_1 = 0$ e $C_2 = +\infty$, respetivamente.

3.1.2 Funções de sobrevivência e de risco

De seguida definem-se algumas funções de extrema importância para este tipo de análise, de acordo com Rocha e Papoila (2009).

A função de sobrevivência $S(t)$ dá a probabilidade de um indivíduo, neste caso, cadáver, sobreviver além do instante t , ou seja, é dada por:

$$S(t) = P(T > t). \quad (3.2)$$

Esta função é monótona decrescente, contínua e apesar de, empiricamente, ser uma função por degraus, em teoria é uma curva em que $S(0) = 1$ e $S(+\infty) = \lim_{t \rightarrow +\infty} S(t) = 0$.

A função densidade de probabilidade de T é dada por:

$$f(t) = \lim_{dt \rightarrow 0^+} \frac{P(t \leq T < t + dt)}{dt}. \quad (3.3)$$

A função de risco instantâneo $h(t)$ mede o potencial de a falha ter ocorrido em t , dado que sobreviveu até t pelo que é dada por:

$$h(t) = \lim_{dt \rightarrow 0^+} \frac{P(t \leq T < t + dt | T \geq t)}{dt}. \quad (3.4)$$

Esta função tem a propriedade de ser não negativa e de não ter limite superior. Além disso, é uma função que traduz o risco instantâneo de falha pelo que é habitualmente mais útil e de fácil interpretação que a função de sobrevivência. Esta função pode apresentar diversas formas (monótona, constante ou não monótona) que ajudarão a caracterizar a distribuição do tempo de vida.

A função de risco acumulado $H(t)$ indica o risco de falha até t e é dada por:

$$H(t) = \int_0^t h(u) du, t \geq 0. \quad (3.5)$$

Decorrente das definições apresentadas, podem-se estabelecer as seguintes relações entre as funções:

$$f(t) = h(t)S(t) \quad (3.6)$$

$$S(t) = \exp \left[- \int_0^t h(u) du \right] \quad (3.7)$$

$$f(t) = -S'(t) \quad (3.8)$$

3.1.3 Abordagem não paramétrica

A estimação e análise da sobrevivência deste tipo de dados pode ser feita através de procedimentos paramétricos e não paramétricos. Os métodos não paramétricos caracterizam-se por não necessitarem de pressupostos específicos sobre a distribuição do tempo de sobrevivência T (Collett, 2015), enquanto os paramétricos, para além de especificarem essa distribuição, incluem variáveis na regressão que podem ajudar a explicar o tempo de sobrevivência (Kleinbaum e Klein, 2012).

A abordagem não paramétrica é inicialmente útil para uma análise exploratória dos dados, apresentando gráficos e estimativas empíricas das funções de sobrevivência (3.2) e de riscos (3.4), (3.5), que podem explicar potencialmente a variável resposta. É também importante na comparação dos tempos de sobrevivência para diferentes grupos das observações, que pode ser realizada informalmente pela análise dos gráficos de sobrevivência ou através de testes de hipóteses que avaliam as diferenças significativas da sobrevivência entre grupos (Collett, 2015). Assim, para fazer a estimação das curvas de sobrevivência utilizou-se um estimador específico para dados com censura intervalar, desenvolvido por Turnbull (1976), uma vez que apenas se sabe que o instante de remoção dos cadáveres de aves está compreendido entre dois momentos que correspondem aos dias pré-definidos para se realizar as prospeções ao terreno. Este estimador é um algoritmo análogo ao estimador de Kaplan-Meier, que é utilizado para dados censurados à direita e cuja fórmula é a seguinte:

$$\hat{S}_{KM}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{j_i} \right), \quad (3.9)$$

sendo, d_i o número de falhas ocorridas no instante $t_{(i)}$ e j_i o número de indivíduos em risco no instante $t_{(i)}$.

A implementação do estimador Turnbull será descrita de acordo com Giolo (2004), que apresenta uma versão mais simples e direta que a original, e o respetivo algoritmo encontra-se no Anexo I.

Para uma base de dados de N observações, considere-se $\Omega = \tau_0 < \tau_1 < \tau_2 < \dots < \tau_M$ uma grelha de M instantes de tempo que inclui todos os instantes iniciais L_i e finais R_i , com $i = 1, \dots, N$, dos intervalos de tempo observados durante o estudo do tempo

de sobrevivência. O estimador de Turbull é dado pela mesma fórmula que o estimador de Kaplan-Meier, apresentado na Equação (3.9), mas com definições diferentes para os valores de d e j , tal como se pode verificar no algoritmo seguinte:

Passo 1: Calcular uma estimativa inicial da função de sobrevivência, através do estimador de Kaplan-Meier original, dado pela equação (3.9);

Passo 2: Estimar o número de falhas ocorridas em τ_m como:

$$d_m = \sum_{n=1}^N \frac{\kappa_{nm} p_m}{\sum_{k=1}^M \kappa_{nk} p_k}, \quad (3.10)$$

em que para a n -ésima observação, se define o peso κ_{nm} como

$$\kappa_{nm} = \begin{cases} 1 & \text{se }]\tau_{m-1}; \tau_m[\subset]L_n; R_n] \\ 0 & \text{caso contrário.} \end{cases} \quad (3.11)$$

e a probabilidade da falha ter ocorrido em τ_m como:

$$p_m = \hat{S}_{KM}(\tau_{m-1}) - \hat{S}_{KM}(\tau_m), \quad m = 1, \dots, M; \quad (3.12)$$

De um modo geral, os pesos κ_{nm} indicam se a observação n , registada no intervalo $]L_n; R_n]$, poderá ter ocorrido no instante τ_m .

Passo 3: Calcular o número estimado de indivíduos em risco em τ_m como

$$j_m = \sum_{r=m}^M d_r, \quad m = 1, \dots, M; \quad (3.13)$$

Passo 4: Recalcular o estimador de Kaplan-Meier, tal como é dado por (3.9), usando os valores dos Passos 2 e 3. Se a estimativa recalculada da função de sobrevivência for relativamente próxima (considerando previamente uma diferença aceitável) à estimativa anterior para todos os instantes τ_m , parar o processo. Caso contrário, repetir os Passos 1 a 3, usando a estimativa atualizada de S , até se atingir a diferença entre as funções desejada.

Tendo as estimativas para as curvas de sobrevivência, pode-se agora compará-las e avaliar se há diferenças significativas no tempo de sobrevivência dos cadáveres para grupos de observações, por exemplo para as diferentes estações do ano ou para os tamanhos das aves.

O teste estatístico mais comum para esta comparação é o teste Log-Rank, originalmente proposto por Mantel (1966). No entanto, este teste foi pensado para dados com censura à direita pelo que foi necessário usar uma adaptação do referido teste, desenvolvida por Sun (1996). Nesta proposta para dados com censura intervalar, a hipótese nula $H_0 : S_1(t) = \dots = S_k(t)$ testa se há diferenças significativas entre as funções de sobrevivência de k grupos de observações. Em seguida será apresentada a estrutura do teste.

Considere-se a grelha de instantes Ω e os pesos κ_{nm} como definidos anteriormente. Denote-se $\hat{S}(\tau_i) = P(T > \tau_i)$, $i = 0, \dots, M$ como a estimativa da função de sobrevivência no instante τ_i , dada, por exemplo, pelo estimador de Turnbull. Considera-se esta estimativa da função de sobrevivência independente de quaisquer variáveis explicativas, ou seja, não tendo em conta os diferentes grupos em que as observações se podem dividir. Por outras palavras, esta é a função de sobrevivência geral dos dados. Depois, define-se o número de falhas e o número de indivíduos em risco estimados num determinado instante τ_m como:

$$d_m^* = \sum_{i=1}^N \left[\kappa_{im} (\hat{S}(\tau_{m-1}) - \hat{S}(\tau_m)) \right] \left/ \sum_{u=1}^{M+1} \kappa_{iu} (\hat{S}(\tau_{u-1}) - \hat{S}(\tau_u)) \right. \quad (3.14)$$

e

$$j_m^* = \sum_{r=m}^{M+1} \sum_{i=1}^N \left[\kappa_{ir} (\hat{S}(\tau_{r-1}) - \hat{S}(\tau_r)) \right] \left/ \sum_{u=1}^{M+1} \kappa_{iu} (\hat{S}(\tau_{u-1}) - \hat{S}(\tau_u)) \right. \quad (3.15)$$

Define-se, de forma análoga, o número de falhas e o número de indivíduos em risco estimados de cada grupo l , $l = 1, \dots, k$, no instante τ_m como:

$$d_{ml}^* = \sum_i^* \left[\kappa_{im} (\hat{S}(\tau_{m-1}) - \hat{S}(\tau_m)) \right] \left/ \sum_{u=1}^{M+1} \kappa_{iu} (\hat{S}(\tau_{u-1}) - \hat{S}(\tau_u)) \right. \quad (3.16)$$

e

$$j_{ml}^* = \sum_{r=m}^{M+1} \sum_i^* \left[\kappa_{ir} (\hat{S}(\tau_{r-1}) - \hat{S}(\tau_r)) \right] \left/ \sum_{u=1}^{M+1} \kappa_{iu} (\hat{S}(\tau_{u-1}) - \hat{S}(\tau_u)) \right. \quad (3.17)$$

em que $\sum_i^*$ representa a soma dos valores de todos os indivíduos no grupo l .

A partir destas definições, Sun (1996) baseia o teste à hipótese H_0 na estatística de teste $T^* = (T_1^*, \dots, T_k^*)^\top$ em que

$$T_l^* = \sum_{m=1}^M (d_{ml}^* - j_{ml}^* d_m^* / j_m^*), \quad l = 1, \dots, k \quad (3.18)$$

e considera uma distribuição χ_{k-1}^2 aproximada para $T^{*\top} V^{*-1} T^*$, em que V^* é a estimativa da variância de T^* que pode ser obtida de várias formas, tal como apresentado no artigo em questão.

Apesar deste teste de hipóteses avaliar a igualdade de k funções de sobrevivência, é também adequado aplicar a correção de Bonferroni dado que se está a proceder a comparações múltiplas entre os grupos de observações dos dados desta dissertação em particular (três grupos para o tamanho do cadáver - três comparações - e quatro grupos para a estação do ano - seis comparações). Deste modo, em vez de se utilizar um nível de significância geral α , corrige-se o mesmo para $\frac{\alpha}{\text{nº de comparações}}$ e toma-se a decisão de rejeitar ou não a hipótese nula que compara apenas as funções de sobrevivência de dois grupos com base neste novo nível de significância.

3.1.4 Abordagem paramétrica

Tal como foi descrito anteriormente, os métodos paramétricos, para além de integrar variáveis explicativas na regressão, assumem uma distribuição para o tempo de sobrevivência dos dados. No entanto, assume-se que o tempo de sobrevivência segue uma distribuição de uma determinada família, sendo que não se conhecem os seus parâmetros específicos. Nesta situação, os dados são usados para estimar os parâmetros que definem por completo a distribuição assumida (Kleinbaum e Klein, 2012).

Um dos modelos paramétricos mais importantes para descrever o tempo de sobrevivência é o de Tempo de Vida Acelerado (Accelerated Failure Time - AFT). O pressuposto deste tipo de modelos é que o efeito das covariáveis é multiplicativo em relação ao tempo de sobrevivência, ou seja, sendo $S_1(t)$ e $S_2(t)$ as funções de sobrevivência para duas determinadas combinações de covariáveis, nos modelos AFT pode-se afirmar que $S_1(t) = S_2(et)$, sendo e a constante de aceleração (Kleinbaum e Klein, 2012; Lawless, 2003). Nestes modelos assume-se uma transformação do tempo de vida T , nomeadamente $\log(T)$, podendo ser definidos como

$$\log(T_i) = \beta_0 + \sum_{r=1}^{n_\beta} \beta_r z_{ri} + \sigma \epsilon_i, \quad i = 1, \dots, N \quad (3.19)$$

para a i -ésima observação, sendo z_{ri} as covariáveis respetivas e $\sigma \epsilon_i$ o termo de erro (Collett, 2015).

Para esta dissertação, foram ajustados modelos AFT, seguindo os resultados do artigo de Bispo *et al.* (2013). Para além disso foram utilizados os Critérios de Informação de Desvio (DIC) e de Akaike-Watanabe (WAIC) para a comparação dos modelos, quando se assumem diferentes distribuições para T . A medida DIC é uma versão bayesiana do critério AIC, utilizando estimativas pontuais *a posteriori* do modelo, enquanto a medida WAIC usa a distribuição *a posteriori* dos parâmetros. Apesar de o critério DIC ser o mais utilizado no âmbito da teoria bayesiana, o critério WAIC apresenta melhor desempenho em contextos preditivos e é, de facto, um critério completamente bayesiano (Gelman *et al.*, 2014). Na comparação dos modelos, preferem-se os que têm menores valores DIC e WAIC.

Tendo assumido um tipo de distribuição para T , pode-se escrever a sua função densidade de probabilidade à custa dos parâmetros desconhecidos e, através das relações definidas em (3.6) e em (3.8), é possível obter as expressões analíticas das funções de sobrevivência $S(t)$ e de risco instantâneo $h(t)$ dos dados. De acordo com Kleinbaum e Klein (2012) e através das relações mencionadas, basta definir uma das três funções anteriores e torna-se trivial escrever as expressões das duas restantes.

Algumas das distribuições mais utilizadas para modelar o tempo de sobrevivência são a Exponencial, a Weibull, a Log-Logística ou a Log-Normal. Na Tabela 3.1 apresentam-se

as três principais funções associadas a estas distribuições, de acordo com Bispo *et al.* (2013) e considerando um modelo AFT.

Tabela 3.1: Funções de risco instantâneo, de sobrevivência e densidade de probabilidade para as quatro distribuições.

Função	Exp(λ)	Weibull(λ, α)	Log-Logística(ω, ρ)	Log-Normal(μ, σ)
$h(t)$	λ	$\lambda^\alpha \alpha t^{\alpha-1}$	$\frac{\omega \rho^\omega t^{\omega-1}}{1+(\rho t)^\omega}$	$f(t)/S(t)$
$S(t)$	$\exp(-\lambda t)$	$\exp(-(\lambda t)^\alpha)$	$\frac{1}{1+(\rho t)^\omega}$	$1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right)$
$f(t)$	$h(t)S(t)$	$h(t)S(t)$	$h(t)S(t)$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-0.5\left(\frac{\log t - \mu}{\sigma}\right)^2\right)$

Para um modelo Exponencial, de acordo com a definição de $h(t)$ na Tabela 3.1, observa-se que a função é constante, ou seja, o risco de ocorrer uma falha é igual em qualquer instante de tempo. Apesar de ser um modelo de fácil aplicação e com funções bastante simples, o facto de a função de risco ser constante não representa adequadamente a realidade uma vez que é raro que o risco de falha não se altere durante o tempo. Assim torna-se necessário usar outros modelos mais flexíveis e que representem melhor a situação do risco de ocorrer uma falha para o caso em estudo (Collett, 2015).

Segundo Kleinbaum e Klein (2012), para a distribuição Weibull que será a utilizada nos dados em estudo, o parâmetro λ é recalculado para cada combinação das variáveis explicativas dos dados enquanto o parâmetro α se mantém constante e independente dessas variações. Além disso, verifica-se que:

- se $\alpha = 1$, a função de risco é constante e o modelo passa a ser o Exponencial;
- se $\alpha > 1$, a função de risco é monótona crescente;
- se $0 < \alpha < 1$, a função de risco é monótona decrescente.

A adição deste parâmetro α à distribuição Weibull em comparação com a Exponencial, fornece a flexibilidade necessária para modelar diferentes tipos de tempos de sobrevivência, mantendo as expressões da função simples, pelo que é um dos modelos mais utilizados nesta área (Kleinbaum e Klein, 2012).

As distribuições Log-Logística e Log-Normal permitem ter funções de risco inicialmente crescentes e depois decrescentes, o que poderá modelar adequadamente o risco mais elevado de um cadáver ser removido logo após a morte da ave, devido ao cadáver se encontrar ainda relativamente fresco e ao seu cheiro, deixando ao longo do tempo de ser tão atrativo para os predadores (Bispo *et al.*, 2013).

3.2 Abordagem Bayesiana e INLA

Os princípios da estatística bayesiana vieram mudar o paradigma frequentista estabelecido uma vez que a sua abordagem permite incorporar conhecimento prévio subjetivo na estimação dos parâmetros dos modelos matemáticos. Enquanto na visão frequentista os parâmetros desconhecidos do modelo são considerados valores determinísticos, na visão bayesiana estes parâmetros são tratados como variáveis aleatórias que seguem distribuições conhecidas. É necessário definir a distribuição *a priori* destes parâmetros, isto é, expressar matematicamente o conhecimento que se tem sobre eles antes de qualquer realização da variável aleatória, para se obter a correspondente distribuição *a posteriori*, isto é, a distribuição com informação sobre os parâmetros após a observação dos dados, que é a base da inferência bayesiana para o estudo em causa.

Considere-se uma variável aleatória contínua Y , um conjunto de dados observados $\mathbf{y} = (y_1, \dots, y_N)$ e um conjunto de parâmetros de interesse $\boldsymbol{\theta}$. Um modelo bayesiano geral é dado pela seguinte formulação:

$$\begin{aligned} \mathbf{y}|\boldsymbol{\theta} &\sim \pi(\mathbf{y}|\boldsymbol{\theta}), \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}), \end{aligned} \quad (3.20)$$

sendo $\pi(\mathbf{y}|\boldsymbol{\theta})$ a função densidade que especifica a distribuição dos dados (também chamada de função de verosimilhança), considerando os parâmetros e $\pi(\boldsymbol{\theta})$ é a referida distribuição *a priori* dos parâmetros de interesse. Tendo o modelo definido, a obtenção da distribuição *a posteriori* dos parâmetros é baseada no teorema de Bayes, sendo dada por

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{y}, \boldsymbol{\theta})}{\pi(\mathbf{y})} = \frac{\pi(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})}{\pi(\mathbf{y})}, \quad (3.21)$$

em que $\pi(\boldsymbol{\theta}|\mathbf{y})$ é a distribuição dos parâmetros que se pretende obter e $\pi(\mathbf{y})$ é a função de verosimilhança marginal dos dados. A função de verosimilhança marginal pode ser considerada uma constante de normalização uma vez que não depende dos parâmetros do modelo, pelo que a Equação (3.21) pode ser reescrita como

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}) \quad (3.22)$$

Assim, a Equação (3.22) formaliza a atualização da informação prévia que se tem sobre os parâmetros (a distribuição *a priori* dos parâmetros) através da aprendizagem a partir dos dados observados (a função de verosimilhança dos dados, considerando os parâmetros) para se estimar a distribuição *a posteriori* dos parâmetros de interesse. Além disso, a distribuição *a posteriori* marginal de cada elemento do vetor de parâmetros $\boldsymbol{\theta}$ é definida por

$$\pi(\theta_i|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-i} \quad i = 1, \dots, n_{\boldsymbol{\theta}} \quad (3.23)$$

em que $\boldsymbol{\theta}_{-i}$ representa o conjunto de elementos $\boldsymbol{\theta}$ exceto o parâmetro θ_i (Gomez-Rubio, 2020).

A dificuldade dos métodos bayesianos é precisamente o cálculo de $\pi(\theta|y)$ uma vez que inclui integração de elevada dimensão e raramente apresenta uma fórmula definida (Moraga, 2019). De acordo com Wang *et al.* (2018) existem três maneiras de realizar a computação destas funções que são: métodos exatos, simulação e aproximação. Os métodos exatos encontram-se numa categoria muito restrita da análise bayesiana uma vez que usam fórmulas algébricas da distribuição *a posteriori* completamente definidas e isso apenas acontece quando se usam *prioris* conjugadas. A simulação tem como objetivo gerar amostras da distribuição *a posteriori*, nomeadamente através da técnica MCMC cuja distribuição estacionária é considerada a distribuição de interesse. Apesar do sucesso desta técnica, o seu tempo de computação é elevado e a aplicação em software é complexa. Por fim, os métodos de aproximação pretendem aproximar os integrais anteriormente mencionados, usando técnicas como a aproximação de Laplace, pelo que a abordagem INLA utilizada nesta dissertação se enquadra neste método de computação bayesiana.

Uma outra questão relevante na abordagem Bayesiana é a escolha adequada da distribuição *a priori* dos parâmetros. Estas distribuições podem classificar-se em quatro categorias: subjetivas, conjugadas, não informativas e de complexidade penalizada (Wang *et al.*, 2018). As *prioris* subjetivas são aquelas que advêm do conhecimento individual de um especialista ou de uma experiência anterior com dados semelhantes, definindo a distribuição dos parâmetros com essa informação prévia. As distribuições conjugadas são as mais convenientes em termos matemáticos uma vez que tanto a *priori* como a *posteriori* têm a mesma distribuição, o que permite ter uma fórmula explícita da *posteriori* em função dos parâmetros da *priori* e dos dados. Por exemplo, se for escolhida uma distribuição *a priori* Gama para o parâmetro da média de dados Poisson, sabe-se que a distribuição *a posteriori* será também uma Gama. No entanto, existe apenas um número limitado de modelos elementares, em que é possível fazer a escolha de *prioris* conjugadas. Avançando para as *prioris* não informativas, muitas vezes os investigadores não possuem informação robusta sobre os parâmetros pelo que a decisão mais segura é escolher distribuições que não têm grandes consequências nos resultados, como distribuições Uniformes, por exemplo. As *prioris* de complexidade penalizada são aqui mencionadas porque são as distribuições pré-definidas no package INLA para o parâmetro α da distribuição Weibull associada ao tempo de vida dos dados desta dissertação (ver secção 3.1.4).

As *prioris* de complexidade penalizada (Penalized Complexity - PC) foram propostas por Simpson *et al.* (2017) e estão definidas de forma a penalizar a complexidade introduzida por se desviar de um modelo mais simples para outro mais complexo, ajustados a dados Y . Segundo o artigo referido, estas distribuições são definidas com base na divergência de Kullback-Leibler (KLD) (Kullback e Leibler, 1951) que, para um parâmetro ξ , é dada por

$$\text{KLD}(\xi) = \text{KLD}(f(y)||g(y)) = \int_0^\infty f(y) \log \left(\frac{f(y)}{g(y)} \right) dy, \quad (3.24)$$

sendo $f(y)$ a função densidade do modelo mais complexo e $g(y)$ a função do modelo mais simples. A partir daqui define-se a distância entre os dois modelos como

$$d(\xi) = d(f(y)||g(y)) = \sqrt{2\text{KLD}(f(y)||g(y))} \quad (3.25)$$

e assume-se que $d(\xi) \sim \text{Exp}(\epsilon)$, sendo ϵ um hiperparâmetro pré-definido no package INLA pelo valor 5. Por fim a função da distribuição *a priori* de ξ é dada por

$$\pi(\xi) = \epsilon \times \exp(-\epsilon d(\xi)) \times |d'(\xi)|. \quad (3.26)$$

Para o caso particular, do parâmetro α do modelo de sobrevivência Weibull para a variável T , van Niekerk *et al.* (2021) derivou as expressões definidas anteriormente, tendo em conta que o modelo Weibull parte do modelo mais simples Exponencial, isto é, $f(y) = f(t) = \lambda^\alpha \alpha t^{\alpha-1} \exp(-(\lambda t)^\alpha)$ e $g(y) = g(t) = \lambda \exp(-\lambda t)$. Assim, obteve a divergência KLD

$$\text{KLD}(\alpha) = \alpha^{-1} \left(\gamma - \alpha(1 + \gamma) + \Gamma\left(\frac{1}{\alpha}\right) + \alpha \log(\alpha) \right), \quad (3.27)$$

em que $\gamma = 0,5772\dots$ representa a constante de Euler-Mascheroni e a função da distribuição *a priori* de α , tendo em conta o parâmetro ϵ da distância entre modelos, dada por

$$\begin{aligned} \pi(\alpha) = & \frac{\epsilon}{4} \exp \left[-\epsilon \sqrt{2\alpha^{-1}(\gamma - \alpha(1 + \gamma) + \Gamma(\alpha^{-1}) + \alpha \log(\alpha))} \right] \\ & \times [2\alpha^{-1}(\gamma - \alpha(1 + \gamma) + \Gamma(\alpha^{-1}) + \alpha \log(\alpha))]^{-1/2} \times \left| \frac{2}{\alpha^2}(\gamma - \alpha(1 + \gamma) \right. \\ & \left. + \Gamma(\alpha^{-1}) + \alpha \log(\alpha)) + \frac{2}{\alpha}(-(1 + \gamma) - \Gamma(\alpha^{-1})\varphi(\alpha^{-1})\alpha^{-2} + \log(\alpha) + 1) \right| \end{aligned} \quad (3.28)$$

em que φ é a função digama.

Tanto para os coeficientes dos efeitos fixos como para os efeitos aleatórios do modelo ajustado neste projeto, utilizaram-se as distribuições *a priori* pré-definidas do package INLA, uma vez que a escolha de melhores distribuições é algo que recai fora do âmbito deste estudo. Os coeficientes dos efeitos fixos seguem uma distribuição Normal de média 0 e precisão (inverso da variância) 0,001 e os efeitos aleatórios dos projetos são também gaussianos com média 0 e matriz de precisão (matriz de covariância inversa) $\delta\Sigma$, em que $\log(\delta)$ segue uma distribuição *a priori* Log-Gama com parâmetros 1 e 0,00005, ou seja, a distribuição *a priori* de δ é uma Gama (Gomez-Rubio, 2020).

3.2.1 Integrated Nested Laplace Approximation

Nesta secção serão apresentadas as três componentes chave da abordagem INLA, introduzido por Rue *et al.* (2009), que são: os modelos latentes gaussianos, os campos aleatórios gaussianos de Markov e a aproximação de Laplace. O INLA é uma abordagem determinística para aproximar a inferência bayesiana e foca-se em estimar as funções das distribuições *a posteriori* marginais de cada parâmetro, dadas pela Equação (3.23), em vez

da função da distribuição conjunta dada pela Equação (3.22) (Gomez-Rubio, 2020; Wang *et al.*, 2018).

3.2.1.1 Modelos latentes gaussianos

A abordagem INLA foi especialmente desenvolvido para modelos latentes gaussianos que têm um vasto horizonte de aplicação, nomeadamente modelos com componente espacial e efeitos aleatórios (modelos hierárquicos). O primeiro passo para a definição destes modelos é a identificação da distribuição da variável resposta Y , sendo depois estabelecida a seguinte formulação:

$$\begin{aligned} y_i | \mathbf{x}, \boldsymbol{\theta} &\sim \pi(y_i | x_i, \boldsymbol{\theta}), i = 1, \dots, N, \\ \mathbf{x} | \boldsymbol{\theta} &\sim \text{Normal}(\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{Q}(\boldsymbol{\theta})^{-1}), \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}), \end{aligned} \quad (3.29)$$

em que $\mathbf{x}$ representa o vetor de efeitos latentes gaussianos com média $\boldsymbol{\mu}(\boldsymbol{\theta})$ e matriz de precisão $\mathbf{Q}(\boldsymbol{\theta})$ e em que $\boldsymbol{\theta}$ é o conjunto de hiperparâmetros desconhecidos do modelo (Moraga, 2019).

Nestes modelos latentes gaussianos, a média μ_i da observação y_i é geralmente conectada com o preditor linear η_i através de uma função de ligação adequada (Gomez-Rubio, 2020). Em particular para o modelo de sobrevivência Weibull espacial ajustado neste estudo, o parâmetro λ da distribuição do tempo de persistência está ligado ao preditor linear através da função logarítmica, ou seja, $\log(\lambda_i) = \eta_i, i = 1, \dots, N$. O preditor linear é definido por

$$\eta_i = \beta_0 + \sum_{r=1}^{n_\beta} \beta_r z_{ri} + \sum_{h=1}^{n_f} f^{(h)}(z_{hi}^*), \quad i = 1, \dots, N, \quad (3.30)$$

em que β_r são os coeficientes associados às covariáveis z_{ri} e $f^{(h)}(\cdot)$ representa as funções de efeitos aleatórios em termos de algumas covariáveis z_{hi}^* , como sejam efeitos espaciais ou efeitos aleatórios independente e identicamente distribuídos (Blangiardo e Cameletti, 2015). Assim, podem-se definir os efeitos latentes gaussianos como

$$\mathbf{x} = (\beta_0, \{\beta_r\}, \{f^{(h)}\}) \quad (3.31)$$

que, condicionados aos parâmetros $\boldsymbol{\theta}$, seguem a distribuição Normal apresentada na Formulação (3.29). Neste vetor, também chamado de campo aleatório gaussiano, os elementos, isto é, β_0 , os coeficientes dos efeitos fixos e as funções dos efeitos aleatórios seguem uma distribuição *a priori* gaussiana. Além disso, Rue *et al.* (2009) assume que estes campos aleatórios admitem determinadas propriedades, explicadas de seguida, que os tornam em campos aleatórios gaussianos de Markov (Gaussian Markov random fields - GMRF), a segunda componente chave da abordagem INLA.

3.2.1.2 Campos aleatórios gaussianos de Markov

Numa descrição simplista, os campos aleatórios gaussianos de Markov são apenas vetores aleatórios que seguem uma Normal multivariada e satisfazem o pressuposto de independência condicional. Neste caso, o campo aleatório em questão será x , definido na Equação (3.31). De seguida, são apresentados alguns conceitos básicos e teoremas para compreender melhor a definição formal de um GMRF, de acordo com Rue e Held (2005).

Definição 3.2.1 (Independência) *Duas variáveis aleatórias X e Y dizem-se independentes se e só se, $\pi(X, Y) = \pi(X)\pi(Y)$. Neste caso, escreve-se que $X \perp Y$.*

Definição 3.2.2 (Independência Condicional - Propriedade de Markov) *Duas variáveis aleatórias X e Y dizem-se condicionalmente independentes, dada a variável Z se e só se, $\pi(X, Y | Z) = \pi(X | Z) \pi(Y | Z)$. Neste caso, escreve-se que $X \perp Y | Z$.*

Definição 3.2.3 (Grafo não orientado) *Um grafo não orientado é um tuplo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, em que $\mathcal{V}$ é o conjunto de nós e $\mathcal{E}$ é o conjunto de arestas que ligam os nós de $\mathcal{V}$.*

A partir da definição de grafo não orientado, estabelece-se que os nós i e j são vizinhos se e só se existe uma aresta que os une. Com estas três definições, os autores estabelecem informalmente a seguinte estrutura para um GMRF.

Definição 3.2.4 *Seja $x = (x_1, \dots, x_N)^T$ um vetor Normal com média $\mu(\theta)$ e matriz de precisão $Q(\theta)$. Defina-se um grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ em que $\mathcal{V} = \{1, \dots, N\}$ e $\mathcal{E}$ é tal que não existe nenhuma aresta entre o nó i e o nó j se e só se $x_i \perp x_j | x_{-ij}$. Então x é um campo aleatório gaussiano de Markov (GMRF) com respeito a $\mathcal{G}$.*

Desta definição informal de um GMRF deduz-se que a informação sobre a independência condicional entre dois elementos do campo aleatório deverá estar contida na matriz de precisão, pelo que Rue e Held (2005) apresentam o seguinte teorema.

Teorema 3.2.1 *Seja x um vetor Normal com média $\mu(\theta)$ e matriz de precisão $Q(\theta) > 0$. Então, para $i \neq j$,*

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

Este teorema permite concluir que dois elementos do campo aleatório são vizinhos se e só se o seu valor correspondente de precisão for diferente de 0. Garante ainda que os elementos diferentes de 0 de uma matriz de precisão determinam o grafo associado e vice-versa. Além disso, os autores afirmam que os elementos da diagonal da matriz $Q(\theta)$ são os valores da precisão condicionada de $x_i | x_{-i}$ enquanto os restantes elementos, com o devido cálculo, fornecem informação sobre a correlação entre x_i e x_j , dado x_{-ij} , pelo que a maioria dos elementos será igual a 0. Este facto confirma a esparsidade da matriz de precisão, o que é um fundamento essencial para a rapidez de computação de um GMRF

e, conseqüentemente, da abordagem INLA.

Por fim, apresenta-se a definição formal de um GMRF.

Definição 3.2.5 (Campo aleatório gaussiano de Markov) Um vetor aleatório $\mathbf{x} = (x_1, \dots, x_N)^\top$ é chamado de um campo aleatório gaussiano de Markov (GMRF) com respeito ao grafo $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ com média $\mu(\theta)$ e matriz de precisão $\mathbf{Q}(\theta) > 0$, se e só se a sua função densidade é da forma

$$\pi(\mathbf{x}) = (2\pi)^{-N/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top \mathbf{Q}(\mathbf{x} - \mu)\right) \quad (3.32)$$

e com

$$Q_{ij} \neq 0 \Leftrightarrow \{i, j\} \in \mathcal{E} \quad \forall i \neq j.$$

Está assim apresentada a estrutura do campo aleatório (3.31) que permite o ajuste do modelo bayesiano pela abordagem INLA.

3.2.1.3 Aproximações de Laplace e Inferência

A última componente essencial e que dá o nome à abordagem em questão é a aproximação de Laplace. Segundo Blangiardo e Cameletti (2015), suponha-se que se quer calcular o seguinte integral de uma função densidade de uma variável aleatória Y :

$$\int f(y) dy = \int \exp(\log f(y)) dy. \quad (3.33)$$

O objetivo do método de Laplace é aproximar o integral de uma dada função densidade ao integral de uma distribuição Normal, com uma expressão completa e definida. Começa-se por representar $\log f(y)$ através da expansão em série de Taylor no ponto $y = y_0$:

$$\log f(y) \approx \log f(y_0) + (y - y_0)(\log f(y_0))' + \frac{(y - y_0)^2}{2}(\log f(y_0))''. \quad (3.34)$$

Se y_0 for definido como sendo a moda $y^* = \operatorname{argmax}_y \log f(y)$ então $(\log f(y^*))' = 0$ pelo que a série de Taylor fica

$$\log f(y) \approx \log f(y^*) + \frac{(y - y^*)^2}{2}(\log f(y^*))''. \quad (3.35)$$

Assim, o integral da Equação (3.33) pode ser aproximado por:

$$\begin{aligned} \int f(y) dy &\approx \int \exp\left(\log f(y^*) + \frac{(y - y^*)^2}{2}(\log f(y^*))''\right) dy \\ &= \exp(\log f(y^*)) \int \exp\left(\frac{(y - y^*)^2}{2}(\log f(y^*))''\right) dy \end{aligned} \quad (3.36)$$

em que o integrando pode ser associado à densidade de uma distribuição Normal, tomando $\sigma^{2*} = -1/(\log f(y^*))''$:

$$\int f(y) dy \approx \exp(\log f(y^*)) \int \exp\left(-\frac{(y - y^*)^2}{2\sigma^{2*}}\right) dy \quad (3.37)$$

onde o integrando é o núcleo de uma distribuição Normal com média igual a y^* e variância σ^{2*} . Mais precisamente, o integral no intervalo $[a, b]$ é aproximado por

$$\int_a^b f(y)dy \approx f(y^*)\sqrt{2\pi\sigma^{2*}}(\phi(b) - \phi(a)), \quad (3.38)$$

sendo $\phi(\cdot)$ a função densidade cumulativa da Normal(y^*, σ^{2*}).

Este será o raciocínio aplicado na inferência bayesiana da função da distribuição *a posteriori* dos efeitos latentes, dada por

$$\pi(x, \theta | y) = \frac{\pi(y|x, \theta)\pi(x, \theta)}{\pi(y)} \propto \pi(y|x, \theta)\pi(x, \theta), \quad (3.39)$$

sendo $\pi(y)$ uma constante normalizadora que pode ser ignorada (Gomez-Rubio, 2020). A função $\pi(y|x, \theta)$ representa a verosimilhança dos dados, que por serem independentes dos efeitos latentes e dos parâmetros, pode ser reescrita como

$$\pi(y|x, \theta) = \prod_{i=1}^N \pi(y_i|x_i, \theta). \quad (3.40)$$

A função da distribuição conjunta dos efeitos latentes e dos parâmetros $\pi(x, \theta)$ pode ser reformulada para

$$\pi(x, \theta) = \pi(x|\theta) \times \underbrace{\pi(\theta)}_{\text{Priori dos hiperparâmetros}}. \quad (3.41)$$

Dado que x é um GMRF, com distribuição Normal definida na Formulação (3.29), a função da sua distribuição *a posteriori* pode ser reescrita, com base na Equação (3.32) como

$$\pi(x|\theta) \propto |Q(\theta)|^{1/2} \exp\left(-\frac{1}{2}(x - \mu)^\top Q(\theta)(x - \mu)\right). \quad (3.42)$$

Assim, conjugando as Equações (3.40), (3.41) e (3.42), pode-se reformular a Equação (3.39) para

$$\begin{aligned} \pi(x, \theta | y) &\propto \pi(\theta)|Q(\theta)|^{1/2} \exp\left(-\frac{1}{2}(x - \mu)^\top Q(\theta)(x - \mu)\right) \prod_{i=1}^N \pi(y_i|x_i, \theta) = \\ &= \pi(\theta)|Q(\theta)|^{1/2} \exp\left(-\frac{1}{2}(x - \mu)^\top Q(\theta)(x - \mu) + \sum_{i=1}^N \log(\pi(y_i|x_i, \theta))\right) \end{aligned} \quad (3.43)$$

Tal como mencionado em Gomez-Rubio (2020), onde também se derivaram as expressões aqui apresentadas a partir da Equação (3.39), na abordagem INLA não se estima a função da distribuição *a posteriori* conjunta definida na Equação (3.43) mas propõe-se estimar as funções das distribuições marginais dos efeitos latentes e dos parâmetros desconhecidos da seguinte forma:

$$\pi(x_i | y) = \int \pi(x_i | \theta, y) \pi(\theta | y) d\theta, i = 1, \dots, N; \quad (3.44)$$

$$\pi(\theta_r|\mathbf{y}) = \int \pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-r}, r = 1, \dots, n_{\boldsymbol{\theta}}. \quad (3.45)$$

O problema da abordagem INLA é precisamente o cálculo das duas equações anteriores e é aqui que entra o método de aproximação dos integrais, nomeadamente o da aproximação de Laplace. Rue *et al.* (2009) propõe a reformulação do problema em vários sub-problemas mais fáceis de lidar e aplica a aproximação de Laplace definida no início deste subcapítulo apenas em funções densidade que são quase gaussianas. Assim, a estratégia proposta passa pela aproximação de $\pi(\boldsymbol{\theta}|\mathbf{y})$ e de $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})^2$. Como resultado, as funções das distribuições *a posteriori* aproximadas de interesse, retornadas pelo INLA, têm as seguintes expressões:

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_{r=1}^{n_{\boldsymbol{\theta}}} \tilde{\pi}(x_i|\boldsymbol{\theta}^{(r)}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}^{(r)}|\mathbf{y}) \Delta\boldsymbol{\theta}^{(r)} \quad i = 1, \dots, N; \quad (3.46)$$

$$\tilde{\pi}(\theta_r|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}_{-r} \quad r = 1, \dots, n_{\boldsymbol{\theta}}. \quad (3.47)$$

em que $\{\boldsymbol{\theta}^{(r)}\}$ representa um conjunto de valores de $\boldsymbol{\theta}$ que são usados na integração numérica e que têm respetivamente associado um peso de integração dado por $\Delta\boldsymbol{\theta}^{(r)}$.

Começando pela aproximação de $\pi(x_i|\boldsymbol{\theta}, \mathbf{y})$, Rue *et al.* (2009) propõe três tipos de aproximações: gaussiana, de Laplace e de Laplace simplificada, sendo esta última a aproximação pré-definida no package INLA. A aproximação de Laplace simplificada, representada por $\tilde{\pi}_{\text{SLA}}(x_i|\boldsymbol{\theta}, \mathbf{y})$ é obtida pela expansão em série de Taylor até à 3ª ordem do denominador e do numerador da aproximação de Laplace que é dada por

$$\tilde{\pi}_{\text{LA}}(x_i|\boldsymbol{\theta}, \mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{\tilde{\pi}_{\text{G}}(\mathbf{x}_{-i}|\mathbf{x}_i, \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}_{-i}=\mathbf{x}_{-i}^*(x_i, \boldsymbol{\theta})}. \quad (3.48)$$

Cita-se novamente Rue *et al.* (2009) para uma descrição mais detalhada dos métodos de aproximação mencionados.

A aproximação de $\pi(\boldsymbol{\theta}|\mathbf{y})$ é feita com o método de Laplace, donde se obtém que

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{\tilde{\pi}_{\text{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}. \quad (3.49)$$

Nas duas equações anteriores ((3.48) e (3.49)), a função $\tilde{\pi}_{\text{G}}(\cdot)$ é uma aproximação gaussiana, como definida em Rue *et al.* (2009), avaliada no seu ponto modal.

Estão assim definidas as componentes chave para o sucesso e rapidez de computação da abordagem INLA.

²Denota-se $\tilde{\pi}(\cdot)$ como a aproximação da respetiva função.

3.2.2 Modelos Espaciais

Nesta secção serão apresentados os conceitos espaciais que complementam a abordagem INLA e permitem ajustar o modelo que responde ao objetivo desta dissertação.

Os Modelos Espaciais distinguem-se dos demais por conseguirem incorporar a variação espacial que muitos dados e informações têm vindo a apresentar recentemente. Estes dados, para além de outras categorias, podem classificar-se como dados referentes a pontos ou geoestatísticos (Natário e Carvalho, 2008) que podem, simplesmente, ser vistos como observações do processo estocástico $\{U(\mathbf{s}) : \mathbf{s} \in D\}$, em que $\mathbf{s} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_N\}$ é o conjunto das localizações observadas e precisas (identificadas pelas suas coordenadas geográficas) e D representa uma área de todo o domínio $\mathbb{R}^2$ (Cressie, 1993). Os dados relativos à remoção dos cadáveres de aves desta dissertação inserem-se nesta categoria porque são observações registadas em localizações exatas (apesar de serem aleatorizadas, como é explicado no Capítulo 4), no domínio contínuo que é Portugal.

Os dados geoestatísticos têm uma estrutura espacial subjacente que pode ser modelada por um GMRF, ou seja, a distribuição de $U(\mathbf{s})$ será uma Normal multivariada com média 0 e uma determinada matriz de covariância Q (Moraga, 2019). É a matriz de covariância que especifica a dependência estrutural do domínio contínuo dos dados e pode ser construída a partir da família de funções de covariância Matérn, que é uma família de funções flexível e a mais utilizada para dados geoestatísticos. De acordo com Krainski *et al.* (2019), a função de covariância Matérn para duas localizações $\mathbf{s}_i$ e $\mathbf{s}_j$ é dada por

$$\text{Cov}(U(\mathbf{s}_i), U(\mathbf{s}_j)) = \frac{\sigma^2 2^{1-v}}{\Gamma(v)} (b \|\mathbf{s}_i - \mathbf{s}_j\|)^v \times B_v(b \|\mathbf{s}_i - \mathbf{s}_j\|) \quad (3.50)$$

em que σ^2 é a variância marginal do processo estocástico, $B_v(\cdot)$ é a função de Bessel de segunda espécie e ordem $v > 0$, $b > 0$ relaciona-se com o alcance ρ , que é a distância para a qual a correlação entre dois pontos é aproximadamente 0 e $\|\cdot\|$ é a distância Euclideana entre as duas localizações.

Após a definição completa da distribuição do GMRF, Lindgren *et al.* (2011) propõe a modelação do campo gaussiano através da abordagem SPDE, isto é, considerando uma equação diferencial parcial estocástica cuja solução é precisamente um GMRF com covariância Matérn. Tal como demonstrado por Whittle (1963), um campo nessas condições é a solução da equação diferencial

$$(b^2 - \Delta)^{a/2}(\phi u(\mathbf{s})) = \mathbf{W}(\mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^2, \quad (3.51)$$

em que Δ representa o operador laplaciano do campo gaussiano e $\mathbf{W}(\mathbf{s})$ é um processo estocástico gaussiano em que $\mathbf{W}(\mathbf{s}_i)$ é independente de $\mathbf{W}(\mathbf{s}_j)$. De acordo com Lindgren e Rue (2015), os parâmetros v e σ^2 da função Matérn, da Equação (3.50), relacionam-se com

esta equação diferencial através de:

$$v = a - 1, \quad (3.52)$$

$$\sigma^2 = \frac{\Gamma(v)}{\Gamma(a)4\pi b^{2v}\rho^2}. \quad (3.53)$$

Note-se que no package INLA, o valor de a está pré-definido para 2.

Com esta abordagem é então possível ajustar a componente espacial do modelo que é o objetivo desta dissertação. Estes autores sugerem, de seguida, uma aproximação da solução da Equação (3.51) através do Método dos Elementos Finitos. Este método consiste na divisão do domínio espacial D num conjunto de triângulos que podem ser irregulares e não se intersejam, sendo que dois triângulos têm no máximo uma aresta ou um vértice em comum. Os vértices dos triângulos chamam-se nós e criam uma malha com G vértices sobre o domínio. A aproximação do campo gaussiano contínuo a um campo discreto é assim dada por

$$u(\mathbf{s}) = \sum_{i=1}^G \psi_i(\mathbf{s})w_i \quad (3.54)$$

em que $\psi_i(\cdot)$ são funções base e w_i são pesos Normalmente distribuídos com média 0. As funções base $\psi_i(\cdot)$ serão úteis para transformar a aproximação $u(\mathbf{s})$ dos nós da malha para outros pontos de interesse, nomeadamente, localizações não observadas, e são, de seguida, explicadas considerando a Figura 3.4.

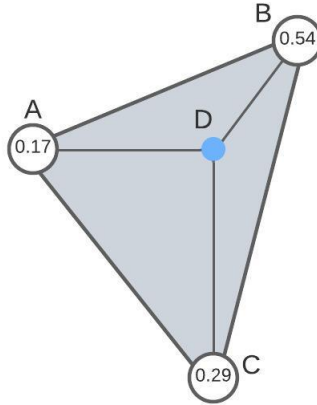


Figura 3.4: Valores da função base do ponto D.

Na Figura 3.4 observa-se um triângulo maior, o ponto D a azul dentro desse triângulo e os respetivos sub-triângulos formados. Os números nos vértices A, B e C representam a área relativa (a sua soma é 1) dos triângulos mais pequenos que lhes são diagonalmente opostos. Portanto, estes números, também chamados de coordenadas barocêntricas, representam a função base do ponto D, em cada um dos vértices do triângulo maior, ou seja, $\psi_A(D)$, $\psi_B(D)$ e $\psi_C(D)$. O INLA lê esta informação em forma de matriz, pelo que se cria uma

matriz de projeção A com um número de linhas igual ao número de localizações em estudo, N , e um número de colunas igual ao número de nós da malha, G . Assim, numa dada linha i , isto é, na localização s_i , poderemos ter até três valores não nulos (Krainski *et al.*, 2019; Moraga, 2019):

1. se a localização estiver dentro de um dos triângulos da malha, os valores de cada uma das três colunas correspondentes a cada um dos três vértices desse triângulo serão os das coordenadas barocêntricas, ou seja, serão proporcionais às áreas dos três sub-triângulos formados por essa localização;
2. se a localização estiver precisamente sobre um dos nós da malha (vértice de um triângulo), na coluna corresponde a esse nó constará um 1 e os restantes valores dessa linha serão 0;
3. se a localização estiver numa aresta, a situação é semelhante à 1. mas haverá apenas duas colunas com valores diferentes de zero, correspondentes aos vértices que criam a aresta.

Por fim, com todos os elementos apresentados ao longo deste capítulo é possível descrever o modelo completo que foi ajustado nesta dissertação. Considerem-se as variáveis explicativas do tamanho dos cadáveres e da estação do ano, a variável do projeto de cada observação e as coordenadas que definem o campo espacial, explicadas no próximo capítulo. Além disso, assume-se que o tempo de sobrevivência T segue uma distribuição Weibull(λ, α) e que é ajustado por um modelo AFT, logo é transformado em $\log(T)$. Deste modo, o parâmetro λ , definido em função das observações, liga-se ao preditor linear da i -ésima observação da seguinte forma:

$$\log(\lambda_i) = \eta_i = \beta_0 + \sum_{r=1}^{n_\beta} \beta_r z_{ri} + f(z_i^*) + A_i \cdot u(s_i), \quad i = 1, \dots, N, \quad (3.55)$$

em que z_{ri} são as variáveis explicativas (tamanho do cadáver e estação do ano); $f(z_i^*)$ é a função do efeito aleatório associado ao projeto da observação, z_i^* ; A_i é a linha da matriz de projeção das coordenadas da i -ésima observação da base de dados e $u(s_i)$ é o efeito espacial, modelado por um GMRF com matriz de covariância Matérn avaliado na localização s_i da observação em causa (Krainski *et al.*, 2019; Lindgren e Rue, 2015).

APLICAÇÃO AO CASO DE ESTUDO

4.1 Dados relativos aos testes de remoção

A informação disponível para o presente estudo teve como ponto de partida os dados compilados por Bernardino *et al.* (2022), relativos à monitorização de 35 projetos de Linhas de Muito Alta Tensão em Portugal Continental (Figura 4.1), sendo que nem todos contemplaram a realização de testes de remoção.

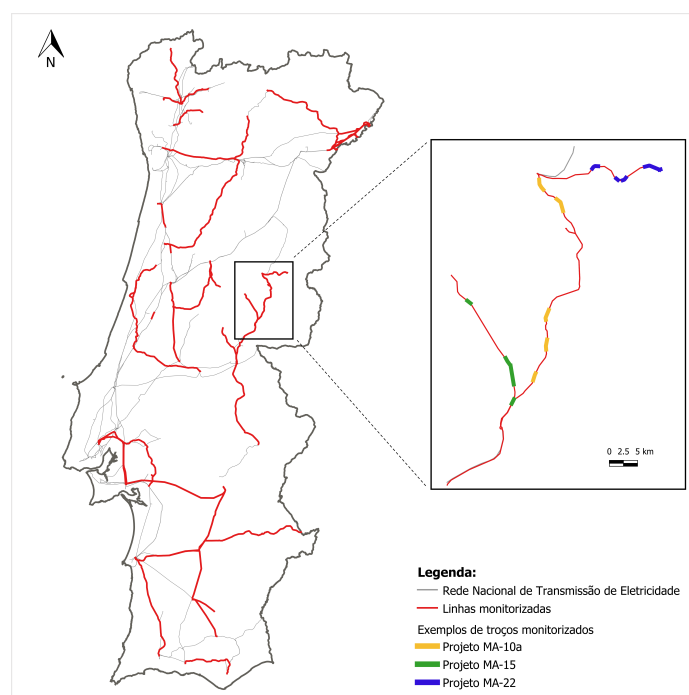


Figura 4.1: Localização dos 35 projetos de linhas estudados e exemplos de troços monitorizados em três desses projetos.

De acordo com os manuais de boas práticas sobre a análise dos impactes nas Linhas de Muito Alta Tensão sobre a avifauna (CIBIO, 2020; ICNB, 2010), as zonas sensíveis devem ser monitorizadas na sua totalidade, devendo a restante extensão da linha ser amostrada pelo menos em 20%. Ambos os manuais indicam que as prospeções de cadáveres de aves

deverão ocorrer, pelo menos, em quatro períodos distintos do ano, num total de quatro visitas consecutivas e com um intervalo de 7 dias, em troços selecionados representativos dos diferentes habitats atravessados pela linha. No entanto, é importante referir novamente que entre campanhas de prospeção pode ocorrer a remoção dos cadáveres por necrófagos, pelo que ambos os manuais recomendam então a realização dos testes de remoção para, assim, se conhecer a probabilidade de um cadáver persistir no terreno até ao dia da prospeção.

Os procedimentos para os testes de remoção variam do 1º manual (ICNB, 2010) para o 2º (CIBIO, 2020), sendo que apenas este último apresenta orientações precisas sobre como realizá-los. As variáveis a serem consideradas devem ser a época do ano e o tamanho do cadáver, pelo que se terão de realizar estes testes em cada uma das quatro estações. Deverão ser distribuídos 20 a 30 cadáveres de cada tamanho (pequeno, médio e grande) em locais aleatórios nos dois eixos da faixa de prospeção e distanciados entre si, no mínimo, 100 metros. Estes locais deverão ser representativos dos principais habitats presentes nos troços selecionados e identificados por coordenadas geográficas para serem posteriormente visitados. Após a colocação, os cadáveres deverão ser visitados diariamente até ao 4º dia, inclusive, e no 7º, 14º e 21º dias para verificação da permanência ou não no terreno.

Contudo, os dados reunidos para o presente estudo remontam aos anos entre 2003 e 2015, durante os quais, na ausência de orientações metodológicas mais precisas, havia uma grande liberdade na forma como se implementavam os testes de remoção de cadáveres. Assim, uma das limitações deste estudo é o facto de os dados disponíveis terem uma combinação de procedimentos experimentais muito diferentes e poderem ser divergentes em relação ao que foi anteriormente descrito, ao nível, por exemplo, das combinações de variáveis testadas (e categorias dentro destas) ou da dimensão da amostra (número de carcaças colocadas por combinação de variáveis testadas).

Além destas limitações na metodologia dos testes, uma primeira verificação dos dados permitiu concluir que os resultados dos testes são reportados de forma global, isto é, os dados não especificam as coordenadas dos cadáveres colocados no terreno, nem os troços da linha (dentro de cada projeto) onde decorreram os testes. Os dados são apenas referentes à linha elétrica monitorizada (à qual foi atribuída um código único de Projeto) pelo que para que fosse possível incorporar a componente espacial, consideraram-se coordenadas geográficas aleatórias para a localização dos cadáveres dentro dos troços de linha elétrica monitorizados em cada projeto. Apesar desta situação ser uma solução artificial, permite extrapolar os resultados das regiões monitorizadas para todo o território nacional, sendo que para um exato processo de modelação seria importante ter os dados com maior detalhe espacial, nomeadamente a coordenada geográfica precisa de cada cadáver colocado no terreno e testes realizados de forma homogénea por toda a rede de linhas elétricas em Portugal.

A matriz de dados inicial relativa aos testes de remoção tem 3494 observações e cinco variáveis, descritas na Tabela 4.1. No pré-processamento destes dados removeram-se apenas quatro observações que não tinham o tamanho do cadáver definido, ficando-se com uma base de dados final que inclui 3490 observações de 30 projetos diferentes. Criaram-se também três novas variáveis (*Status*, *x-LON* e *y-LAT*) que serão explicadas de seguida.

Tabela 4.1: Descrição das variáveis dos dados dos testes de remoção.

Nome	Descrição	Tipo
<i>Cod-Proj</i>	Referência alfanumérica que identifica univocamente o projeto	Nominal
<i>Season</i>	Estação do ano: "Primavera", "Verão", "Outono", "Inverno"	Categórica
<i>Size</i>	Tamanho do cadáver utilizado: "Grande", "Médio", "Pequeno"	Categórica
<i>R1</i>	Último dia após a colocação em que a carcaça esteve presente	Discreta
<i>R2</i>	Primeiro dia após a colocação em que a carcaça não esteve presente	Discreta

A informação sobre a persistência dos cadáveres no terreno está contida nas variáveis *R1* e *R2* da Tabela 4.1 e, para a completar, foi criada a variável numérica *Status* que indica o tipo de censura associada: o valor 0 indica censura à direita, o valor 2 indica censura à esquerda e o valor 3 indica censura intervalar. Como referido acima, a cada observação foi atribuída uma localização aleatória, extraída da *shapefile* disponibilizada pela "Cátedra REN em Biodiversidade" que contém os elementos geográficos correspondentes aos troços de linha elétrica monitorizados em cada projeto. Desse modo, acrescentaram-se duas variáveis contínuas à matriz inicial da Tabela 4.1: *x-LON* que dá a longitude da coordenada e *y-LAT* que dá a latitude da coordenada do cadáver, ambas em graus. No Capítulo 4.3 será explicada a obtenção artificial destas coordenadas, que fez com que os dados passassem a estar geo-referenciados.

Uma vez que a componente espacial associada a estas localizações engloba vários elementos que não podem ser expressos por outras variáveis, usaram-se Modelos Espaciais para a previsão do fator de correção associado à remoção dos cadáveres do terreno. Além disso, estes dados também descrevem o tempo até acontecer a remoção do cadáver, isto é, o tempo de sobrevivência de um cadáver no terreno pelo que foi também utilizada Análise de Sobrevivência para este estudo.

A análise dos dados foi realizada no software R (R Core Team, 2013), em particular usando os seguintes packages:

- *survival* (Terry M. Therneau e Patricia M. Grambsch, 2000; Therneau, 2023) para o estudo da Análise de Sobrevivência;

- interval (Fay e Shaw, 2010) para os testes de hipóteses da Análise de Sobrevida;
- INLA (Lindgren e Rue, 2015; Lindgren *et al.*, 2011; «R-INLA Project», 2020; Rue *et al.*, 2009; Rue *et al.*, 2017) para o ajuste de modelos;
- sf (Pebesma, 2018; Pebesma e Bivand, 2023) para a importação de dados geográficos;
- ggplot2 e patchwork (Pedersen, 2023; Wickham, 2016) para a visualização de dados;
- raster (Hijmans, 2023) para manipulação de dados geográficos.

4.2 Estimação das Funções de Sobrevida

Inicialmente foram aplicados os métodos não paramétricos descritos no Capítulo 3 aos dados dos testes de remoção, sem a componente espacial, isto é, sem as variáveis referentes às coordenadas aleatórias, para avaliar as diferenças nas funções de sobrevivência entre os grupos de observações.

Assim, aplicando o algoritmo de Turnbull, avaliaram-se graficamente as estimativas das funções de sobrevivência, estratificadas por estação do ano (Figura 4.2) e por tamanho do cadáver (Figura 4.3).

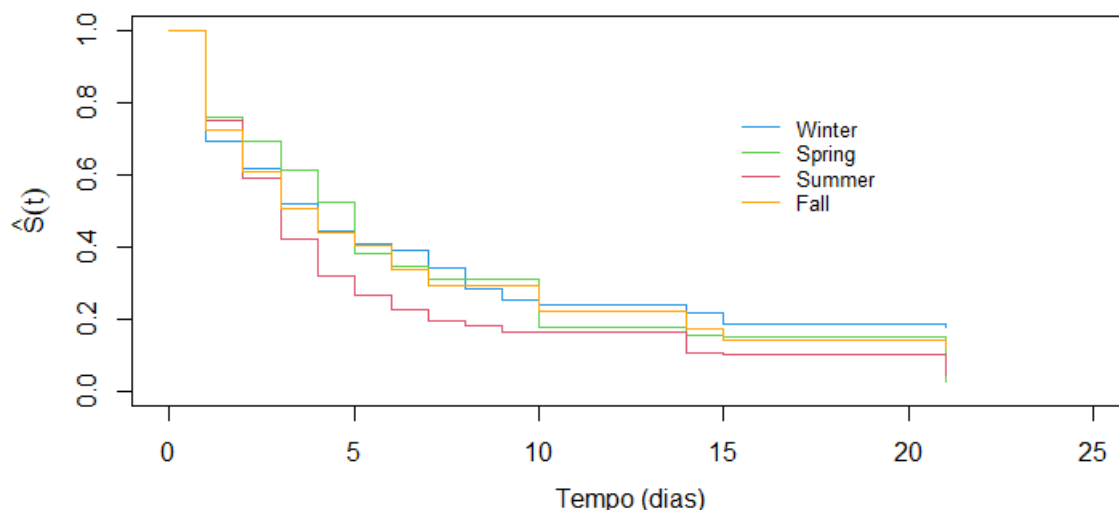


Figura 4.2: Estimativas da função de sobrevivência, estratificada por estação do ano.

Analisando o cruzamento das quatro curvas de sobrevivência da Figura 4.2, pode-se presumir que haverá diferenças relevantes entre a curva da estação "Verão" e as restantes, uma vez que esta está quase sempre abaixo das outras, sendo mais evidente, por exemplo, entre os dias 3 e 10. O instante máximo observado nos dados estratificados foi sempre o dia 30 da verificação de cadáveres nos testes pelo que as quatro curvas apresentam o último degrau no último dia em que há informação sobre a probabilidade de sobrevivência,

de acordo com o algoritmo, isto é, no dia 21 após a sua colocação no terreno. Assim, a probabilidade de sobrevivência do cadáver no terreno entre o dia 21 e o dia 30 é constante e igual a $\hat{S}(21)$, para cada uma das estações do ano correspondentes.

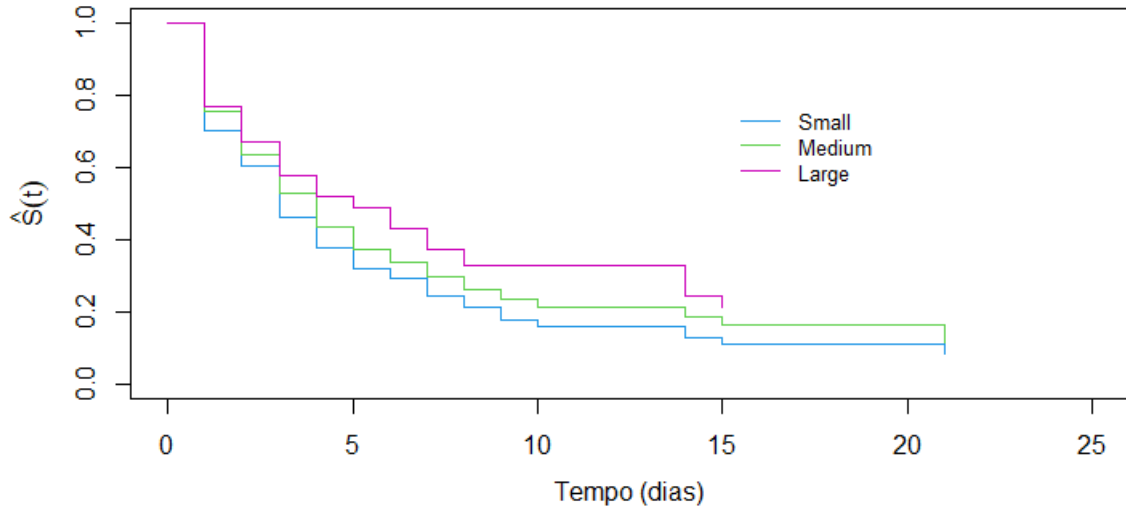


Figura 4.3: Estimativas da função de sobrevivência, estratificada por tamanho do cadáver.

Analisando as três curvas de sobrevivência da Figura 4.3, pode-se presumir, por exemplo, que haverá diferenças relevantes entre a curva do tamanho "Grande", que está claramente sempre acima das restantes curvas, sendo mais evidente a partir do dia 4. O instante máximo observado nos dados filtrados pelo tamanho "Grande" foi o dia 21, pelo que, pelo raciocínio anteriormente explicado, a sua curva termina no dia 15 e a probabilidade de sobrevivência de um cadáver grande no terreno é constante entre os dias 15 e 21 e igual a $\hat{S}_{Large}(15)$. Para os dados estratificados pelos outros dois níveis de tamanho, "Pequeno" e "Médio", o instante máximo foi observado no dia 30, pelo que as curvas terminam no dia 21 que é o último instante de tempo de onde é possível extrair informação sobre a probabilidade de sobrevivência. Assim, para cadáveres pequenos e médios, a probabilidade de sobrevivência no terreno é constante entre os dias 21 e 30 e igual a $\hat{S}(21)$, para cada um dos tamanhos correspondentes.

Analiticamente, as comparações entre as estimativas das funções de sobrevivência podem ser realizadas através do teste Log-Rank descrito no Capítulo 3 e aplicando a correção de Bonferroni. Assim, para as seis comparações das curvas de sobrevivência das estações do ano utilizou-se um nível de significância igual a $0,05/6 = 0,008(3)$ e obtiveram-se os resultados resumidos na Tabela 4.2.

Tabela 4.2: Resumo dos testes Log-Rank realizados para a comparação da sobrevivência entre as estações do ano.

Comparação	Hipótese nula	<i>p-value</i>	Decisão
"Inverno" - "Primavera"	$H_0 : S_{Winter}(t) = S_{Spring}(t)$	0,1188	Não rejeitar
"Inverno" - "Verão"	$H_0 : S_{Winter}(t) = S_{Summer}(t)$	$1,22 \times 10^{-7}$	Rejeitar
"Inverno" - "Outono"	$H_0 : S_{Winter}(t) = S_{Fall}(t)$	0,2105	Não rejeitar
"Primavera" - "Verão"	$H_0 : S_{Spring}(t) = S_{Summer}(t)$	$4,12 \times 10^{-4}$	Rejeitar
"Primavera" - "Outono"	$H_0 : S_{Spring}(t) = S_{Fall}(t)$	0,9825	Não rejeitar
"Verão" - "Outono"	$H_0 : S_{Summer}(t) = S_{Fall}(t)$	$1,47 \times 10^{-3}$	Rejeitar

Como se pode verificar pela última coluna da Tabela 4.2, confirmam-se as diferenças significativas entre a curva de sobrevivência da estação "Verão" e as restantes, uma vez que se rejeitam as respetivas hipóteses nulas. Assim, pode-se afirmar que durante o verão os cadáveres têm menor probabilidade de persistirem no terreno do que nas outras estações do ano.

Para as três comparações das curvas de sobrevivência dos diferentes tamanhos dos cadáveres utilizou-se, agora, um nível de significância de $0,05/3 = 0,0167$ e os resultados estão resumidos na Tabela 4.3.

Tabela 4.3: Resumo dos testes Log-Rank realizados para a comparação da sobrevivência entre os tamanhos dos cadáveres.

Comparação	Hipótese nula	<i>p-value</i>	Decisão
"Pequeno" - "Médio"	$H_0 : S_{Small}(t) = S_{Medium}(t)$	$6,059 \times 10^{-6}$	Rejeitar
"Pequeno" - "Grande"	$H_0 : S_{Small}(t) = S_{Large}(t)$	$4,441 \times 10^{-16}$	Rejeitar
"Médio" - "Grande"	$H_0 : S_{Medium}(t) = S_{Large}(t)$	$6,89 \times 10^{-5}$	Rejeitar

Neste caso, pela última coluna da Tabela 4.3, verifica-se que há diferenças significativas entre a curva do tamanho "Grande" e as duas restantes bem como entre as curvas de sobrevivência dos dois menores tamanhos dos cadáveres. Confirma-se, assim, que a probabilidade de persistência de cadáveres pequenos, médios e grandes é, de facto, diferente ao longo do tempo, sendo que os maiores cadáveres têm maiores probabilidades e os menores têm, coincidentemente, menores probabilidades de persistência.

Esta análise permite concluir, empiricamente, que ambas as variáveis explicativas têm relevância na variável resposta, que é o tempo de persistência do cadáver no terreno e alteram, de facto, as suas curvas de sobrevivência.

4.3 Modelação do Fator de Correção

A *shapefile* que contém os elementos geográficos dos troços de linha monitorizados em cada projeto e onde foram realizados os testes de remoção foi disponibilizada pela coorientadora deste projeto. Na Figura 4.4 são visíveis os diferentes troços de linha associados a cada projeto.

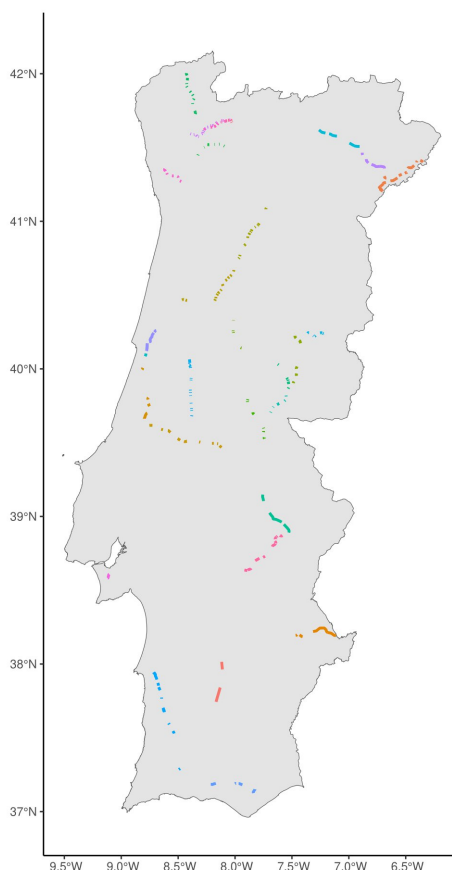


Figura 4.4: Localização dos 30 projetos das linhas onde foram realizados testes de remoção.

A partir desta *shapefile* foram obtidos pares de coordenadas (x -LON, y -LAT) para cada uma das linhas da base de dados dos testes de remoção, através da amostragem aleatória de pontos dentro dos troços do projeto associado a cada observação. Estas localizações aleatórias foram tomadas como os locais onde foram colocados os cadáveres de aves nos testes de remoção, uma vez que os dados originais não continham essa informação. De forma a tentar abranger ao máximo os possíveis cenários das localizações exatas dos cadáveres, realizaram-se 100 iterações do processo de amostragem aleatória, que se conjugaram numa única base de dados, constituída pela matriz inicial da Tabela 4.1, juntamente com a coluna *Status* e mais duas colunas correspondentes à latitude e longitude das coordenadas amostradas aleatoriamente. Dado que a base de dados contém as 100 iterações já referidas (tendo no total 349 000 linhas), este estudo baseou-se no ajuste de 100 modelos com o procedimento explicado de seguida e com a obtenção final dos valores médios do fator de

correção da remoção (para prospeções com frequência semanal) para as 12 combinações de estação/tamanho e os respetivos valores de variância como medida de precisão.

Tendo a base de dados pronta para a modelação, com as componentes espacial e de sobrevivência, realizou-se primeiramente a seleção do melhor modelo espacial de Tempo de Vida Acelerado, ajustando os modelos Exponencial, Weibull, Log-Logística e Log-Normal a cada uma das 100 iterações da base de dados, e escolhendo o modelo com menor medida DIC e WAIC, em média. Na Tabela 4.4 encontram-se os valores médios das medidas mencionadas.

Tabela 4.4: Valores médios de DIC e WAIC para os quatro modelos espaciais AFT ajustados.

Modelo	DIC	WAIC
Exponencial	9479,4	9499,6
Weibull	9193,3	9195,4
Log-Logística	9241,9	9250,1
Log-Normal	9254,4	9259,6

Pode-se concluir que o modelo Weibull apresenta os menores valores em média de DIC e WAIC, pelo que é considerado o melhor ajuste aos dados e foi o modelo escolhido para continuar o estudo desta dissertação.

Os modelos ajustados para cada iteração contêm os seguintes elementos:

- a variável resposta de sobrevivência definida através de $R1$, $R2$ e $Status$;
- os efeitos fixos associados às variáveis explicativas categóricas $Season$ e $Size$ (isto é, estação do ano e tamanho do cadáver, respetivamente), com "Outono" e "Grande" como os respetivos níveis de referência;
- os efeitos aleatórios associados aos projetos das linhas, dados por $Cod-Proj$;
- o efeito espacial definido pelas coordenadas de cada iteração.

O efeito espacial é traduzido pelo modelo SPDE criado e a informação geográfica das coordenadas obtidas no território nacional foi traduzida numa matriz de projeção A tal como explicado no Capítulo 3, em que cada localização dos cadáveres está perfeitamente sinalizada na malha de nós criada para Portugal Continental.

Para as distribuições *a priori* dos vários parâmetros do modelo foram usadas as distribuições pré-definidas do package INLA, tal como descritas no Capítulo 3. Os efeitos aleatórios dos projetos foram definidos como independentes e identicamente distribuídos e os efeitos fixos associados às covariáveis foram introduzidos no modelo como fatores.

O tempo de sobrevivência segue uma distribuição Weibull, cujo parâmetro de escala λ é calculado usando o preditor linear e o parâmetro α segue uma distribuição *a priori* de complexidade penalizada.

A partir do modelo ajustado, foram extraídos os coeficientes dos efeitos fixos, do campo espacial e o valor do parâmetro α , essenciais para se definirem as funções de sobrevivência¹ para cada uma das combinações dos cenários de estação do ano e tamanho do cadáver. Por fim, calculou-se o fator de correção do enviesamento causado pela remoção de cadáveres usando a fórmula da Equação (2.1), assumindo, a título de exemplo, uma situação de prospeções semanais, isto é, $I = 7$, sem perda de generalidade.

4.3.1 Resultados para Portugal Continental

As condições gerais dos modelos espaciais de sobrevivência ajustados às 100 iterações da base de dados foram já apresentadas, pelo que neste capítulo serão apresentados os resultados da previsão do fator de correção da remoção obtidos para todo o território nacional, em função das 12 combinações de covariáveis.

De forma a obter um suporte visual da variação do fator em todo o país, foi necessário criar as localizações em Portugal para onde a previsão do fator de correção seria feita. Para isso, recolheu-se um objeto *raster* com informação da altitude de Portugal no software R, donde se extraíram as localizações, em formato de coordenadas, para onde a altitude tinha sido registada. Obtiveram-se 33962 pontos distribuídos uniformemente pelo território continental, tal como se pode ver na Figura 4.5.



Figura 4.5: Localização dos 33962 pontos para a previsão do fator de correção da remoção no território nacional.

A partir daqui, foi criado um ciclo para ajustar os 100 modelos e extrair os coeficientes

¹A expressão da função de sobrevivência é a que se apresenta na Tabela 3.1, para o modelo Weibull.

já mencionados. Em particular, os coeficientes do campo espacial ajustado conjugaram-se com a nova matriz A que projeta os pontos de previsão na malha de Portugal, permitindo obter as funções de sobrevivência associadas a cada ponto e o respetivo fator de correção. Este ciclo demorou cerca de 6 horas e 15 minutos até terminar e foram obtidos os mapas para os valores médios do fator de correção da remoção (prospeções semanais), representados nas Figuras 4.6, 4.7, 4.8 e 4.9.

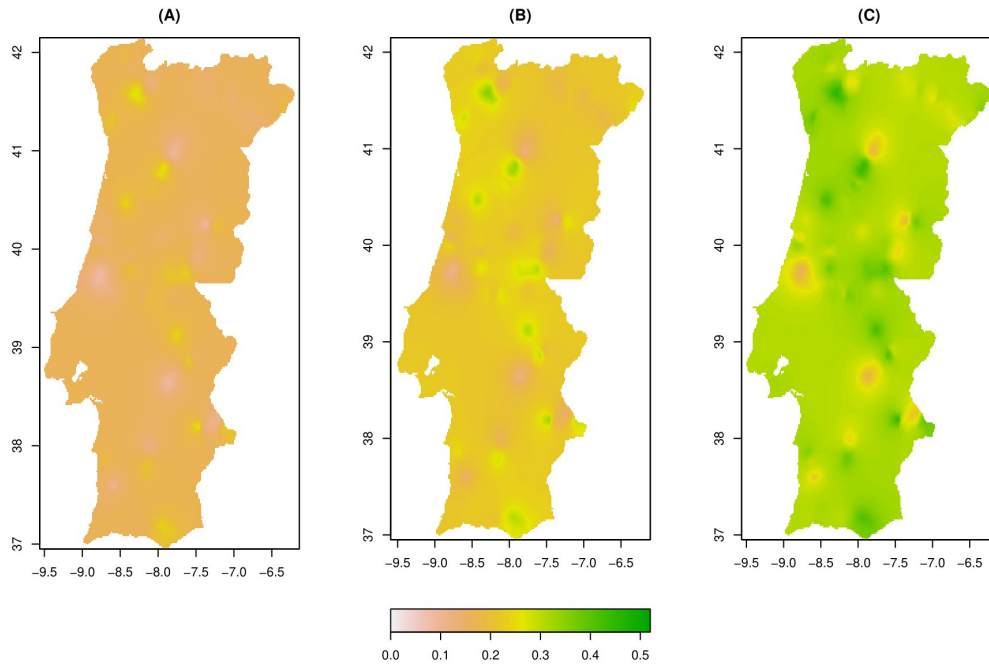


Figura 4.6: Fator de correção médio para a estação "Primavera" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).

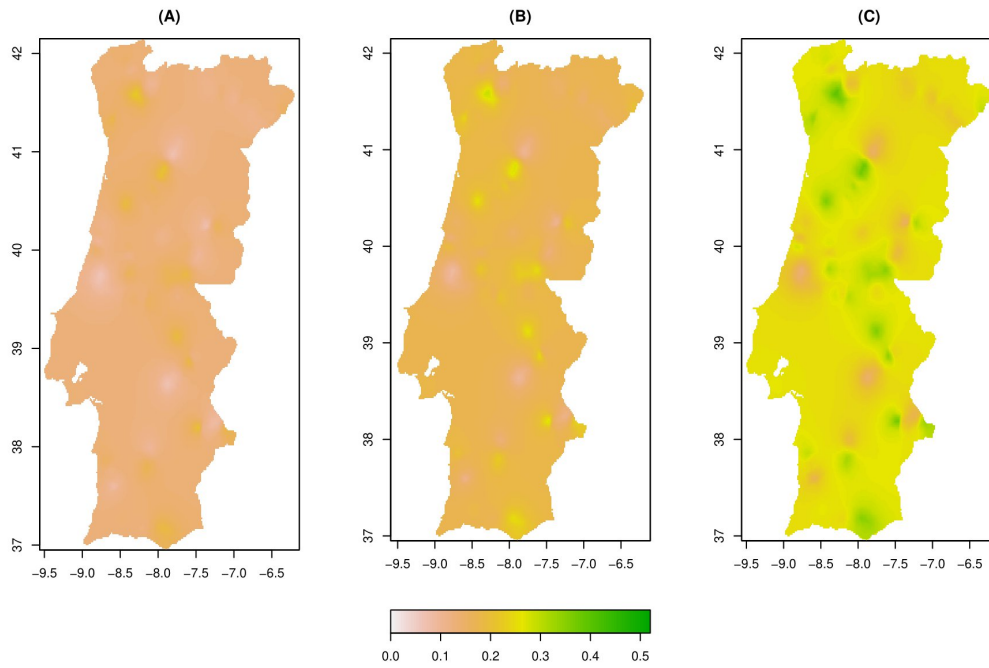


Figura 4.7: Fator de correção médio para a estação "Verão" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).

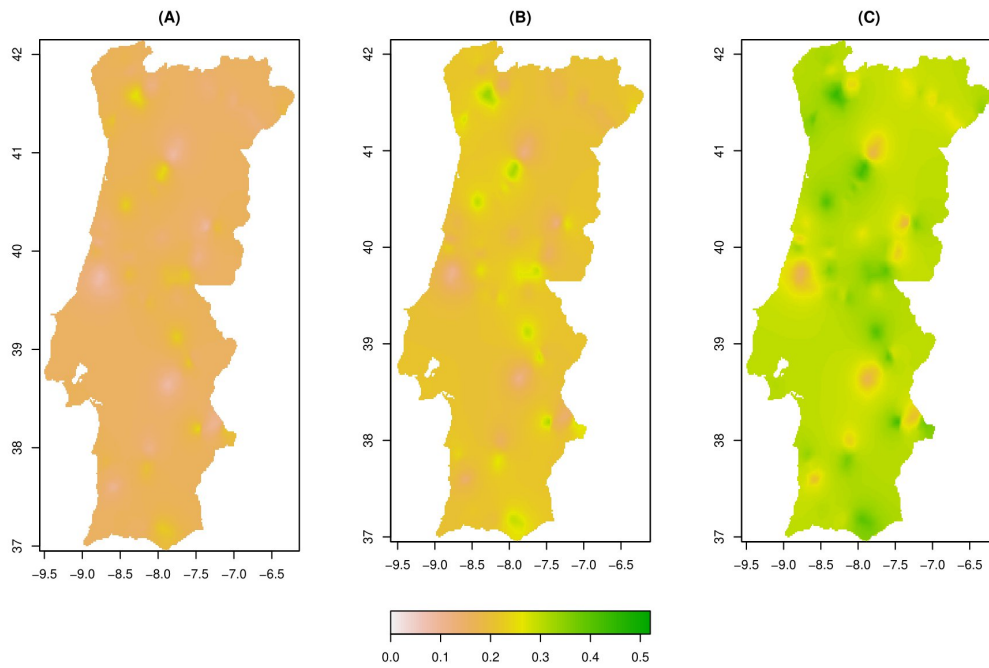


Figura 4.8: Fator de correção médio para a estação "Outono" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).

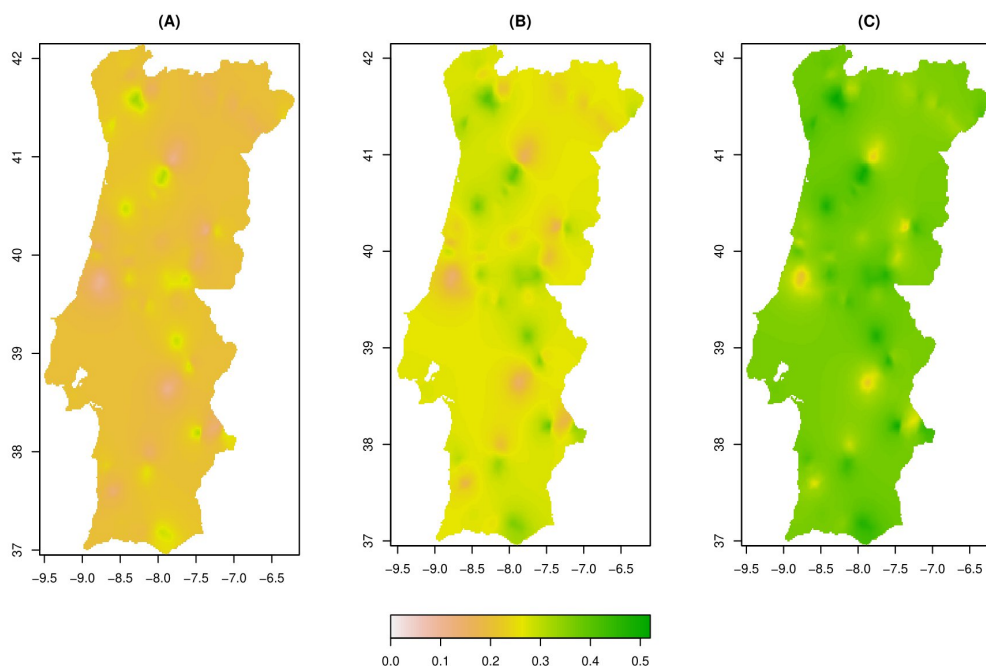


Figura 4.9: Fator de correção médio para a estação "Inverno" e tamanho "Pequeno" (A), "Médio" (B) e "Grande" (C).

Nas Figuras 4.6, 4.7, 4.8 e 4.9 pode-se observar um padrão geral de aumento do fator de correção à medida que o tamanho do cadáver aumenta. Além disso, existem localizações que se destacam recorrentemente, isto é, há zonas nos mapas melhor definidas porque estão próximas de locais onde há informação prévia sobre os testes de remoção realizados que foi incluída no modelo. Assim, estas zonas apresentam uma previsão do fator de correção melhor caracterizada (melhor diferenciação das cores, correspondentes a diferentes valores do fator de correção) enquanto para o restante mapa se pode observar uma única cor correspondente a valores do fator muito semelhantes entre si. Para essas regiões, o modelo apenas consegue prever valores medianos do fator, uma vez que a informação que lhe é passada próxima desses locais não é suficientemente informativa nem homogeneamente espalhada.

É também de notar que em todos os conjuntos de mapas, o fator de correção médio tem valores compreendidos entre 0 e 0,52, sendo que atinge valores máximos na estação "Inverno", no mapa (C), e valores mínimos na estação "Verão", no mapa (A).

Além disto, pode-se comentar que, na sua generalidade, as previsões da estação "Primavera" apresentam valores distribuídos por todo o espectro da escala; as previsões da estação "Verão", exceto as zonas mais destacadas, têm valores até cerca de 0,25; as previsões

da estação "Outono" são semelhantes às da 1ª estação do ano e as da estação "Inverno", à exceção das zonas mais destacadas, têm valores do fator a começar em, aproximadamente, 0,2.

A informação dos quatro mapas apresentados pode ser resumida na Tabela 4.5 em que se apresentam as médias do fator de correção nacional para cada uma das 12 combinações e os respetivos valores de desvio padrão e coeficiente de variação.

Tabela 4.5: Valores médios do fator de correção e respetiva variabilidade.

Combinação	FC médio nacional	DP	Coef. Variação
"Primavera" - "Pequeno"	0,164	0,0171	10,46%
"Primavera" - "Médio"	0,220	0,0212	9,66%
"Primavera" - "Grande"	0,313	0,0258	8,24%
"Verão" - "Pequeno"	0,128	0,0139	10,85%
"Verão" - "Médio"	0,175	0,0180	10,29%
"Verão" - "Grande"	0,256	0,0233	9,10%
"Outono" - "Pequeno"	0,155	0,0164	10,56%
"Outono" - "Médio"	0,209	0,0205	9,81%
"Outono" - "Grande"	0,300	0,0253	8,44%
"Inverno" - "Pequeno"	0,202	0,0201	9,93%
"Inverno" - "Médio"	0,267	0,0239	8,95%
"Inverno" - "Grande"	0,368	0,0273	7,41%

As tendências observadas nos mapas mantêm-se nesta tabela e pode-se observar que o desvio padrão não é muito elevado, o que está espelhado nos valores do coeficiente de variação entre 7% e 10,6%. Estes valores relativamente baixos permitem afirmar que não existe grande variabilidade entre os valores do fator de correção e a sua respetiva média, para cada uma das 12 combinações. No entanto, os maiores valores do coeficiente de variação (mais próximos de 10%) para o tamanho "Pequeno", em todas as estações, indicam que há uma maior amplitude dos respetivos valores do fator de correção, em todo o território.

Como medida de erro, foram também calculados os valores da variância associada aos mapas do fator de correção médio, para cada um dos 33962 pontos de previsão e para as 12 combinações. Na Tabela 4.6 apresentam-se os valores mínimos e máximos do coeficiente de variação, para melhor interpretação do seu significado.

Tabela 4.6: Valores mínimos e máximos do coeficiente de variação do fator de correção.

Tamanho	"Primavera"		"Verão"		"Outono"		"Inverno"	
	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo	Mínimo	Máximo
"Pequeno"	1,73%	21,15%	2,02%	22,69%	0,93%	21,63%	2,07%	20,10%
"Médio"	1,58%	19,88%	1,89%	21,23%	0,95%	19,87%	1,88%	19,03%
"Grande"	1,89%	17,86%	1,73%	19,15%	0,58%	17,98%	1,39%	16,57%

Da Tabela 4.6 entende-se que mesmo os valores máximos dos coeficientes de variação são valores relativamente baixos, indicando que os valores dos fatores estão em concordância com a sua média. A maior variação dos valores do coeficiente ocorre, no geral, para o tamanho "Pequeno" indicando que há uma maior dispersão dos valores do fator para este tamanho em torno dos valores médios dos mapas apresentados. A variação máxima ocorre para a estação "Outono" e tamanho "Pequeno" e a variação mínima acontece na estação "Inverno", tamanho "Grande", levando a concluir que os valores do fator de correção, para a combinação "Outono" - "Pequeno", têm uma maior dispersão em torno do seu valor médio do que para a combinação "Inverno" - "Grande".

4.3.2 Resultados para localização específica

Neste capítulo serão apresentados os resultados para duas localizações específicas, indicadas na Figura 4.10. Estas localizações foram escolhidas aleatoriamente e considerando apenas a sua proximidade ou não a locais com dados dos testes.

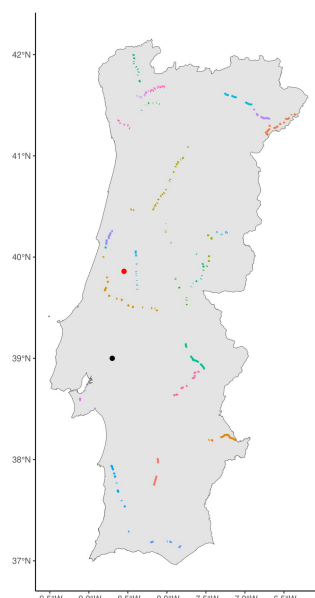


Figura 4.10: Localização dos dois pontos específicos (Pombal - vermelho e Santarém - preto) para a previsão do fator de correção.

Uma delas situa-se próximo de Pombal, indicada pelo ponto vermelho (8,5491°W, 39,8592°N), próxima de zonas em que foram realizados testes de remoção e cujos dados foram registados na matriz inicial da Tabela 4.1, e outra fica próximo de Foros de Salva-terra, Santarém, indicada pelo ponto preto (8,7000°W, 39,0000°N), afastada de zonas com informação dos dados.

Novamente, a abordagem INLA foi aplicada (nas condições gerais já explicadas no início do Capítulo 4.3) a dois ciclos de 100 iterações (um ciclo para cada localização específica), sendo que os coeficientes do campo espacial foram agora conjugados com a matriz de projecção de cada uma das duas localizações, na malha de Portugal. O facto de se estar a prever apenas para um par de coordenadas diminui o tempo de computação para cerca de 1 hora e 30 minutos, o que tornou possível a extração de outros resultados para além do fator de correção médio e respetiva variância para cada uma das combinações de variáveis explicativas, nomeadamente os valores médios da função de sobrevivência para um horizonte temporal de 30 dias.

Assim, os valores médios do fator de correção semanal e os respetivos valores do coeficiente de variação obtidos para Pombal são apresentados na Tabela 4.7, que, para efeitos de comparação, contém também os valores do fator de correção médio nacional, calculado a partir dos resultados para Portugal da secção anterior.

Tabela 4.7: Valores do fator de correção semanal médio e respetivos coeficientes de variação para Pombal e fator de correção semanal médio nacional.

Combinação	FC médio	Coef. Variação	FC médio nacional
"Primavera" - "Pequeno"	0,152	5,56%	0,164
"Primavera" - "Médio"	0,205	5,32%	0,220
"Primavera" - "Grande"	0,295	4,57%	0,313
"Verão" - "Pequeno"	0,119	6,34%	0,128
"Verão" - "Médio"	0,163	6,2%	0,175
"Verão" - "Grande"	0,240	5,54%	0,256
"Outono" - "Pequeno"	0,144	5,44%	0,155
"Outono" - "Médio"	0,195	5,27%	0,209
"Outono" - "Grande"	0,283	4,51%	0,300
"Inverno" - "Pequeno"	0,189	5,45%	0,202
"Inverno" - "Médio"	0,251	5,08%	0,267
"Inverno" - "Grande"	0,350	4,14%	0,368

Na Tabela 4.7 vê-se claramente que o fator de correção aumenta novamente à medida que o tamanho do cadáver aumenta. No geral, a estação "Inverno" apresenta as maiores

previsões do fator de correção e a estação "Verão" apresenta as menores, estando de acordo com aquilo que foi observado nos resultados para Portugal. O fator de correção está compreendido entre os valores de 0,119 e 0,350, sendo esta uma amplitude menor do que o que se apresenta nos mapas de Portugal. Comentando os valores do coeficiente de variação, observa-se que são novamente baixos, indicando que os valores do fator calculados em cada ajuste dos 100 modelos não estão muito longe da sua média. As maiores variações dos valores do fator de correção acontecem para o tamanho "Pequeno", novamente, devido aos maiores valores do coeficiente de variação. Comparando a primeira coluna com a última, observa-se que os fatores de correção estimados para Pombal não são próximos da média nacional, o que evidencia que o modelo INLA consegue, de facto, fazer uma previsão mais diferenciada, ou seja, mais autónoma da média geral do país, para locais que se situam mais perto de locais para onde havia originalmente dados sobre a remoção de cadáveres.

Para a localização em Foros de Salvaterra, o procedimento foi análogo e obtiveram-se os resultados para o fator de correção semanal médio e os respetivos valores do coeficiente de variação apresentados na Tabela 4.8, que inclui também os valores do fator de correção médio nacional.

Tabela 4.8: Valores do fator de correção semanal médio e respetivos coeficientes de variação para Foros de Salvaterra e fator de correção semanal médio nacional.

Combinação	FC médio	Coef. Variação	FC médio nacional
"Primavera" - "Pequeno"	0,161	2%	0,164
"Primavera" - "Médio"	0,217	1,89%	0,220
"Primavera" - "Grande"	0,310	1,57%	0,313
"Verão" - "Pequeno"	0,126	2,61%	0,128
"Verão" - "Médio"	0,172	2,52%	0,175
"Verão" - "Grande"	0,254	2,19%	0,256
"Outono" - "Pequeno"	0,153	1,58%	0,155
"Outono" - "Médio"	0,207	1,58%	0,209
"Outono" - "Grande"	0,297	1,1%	0,300
"Inverno" - "Pequeno"	0,200	2,37%	0,202
"Inverno" - "Médio"	0,264	2,16%	0,267
"Inverno" - "Grande"	0,365	1,57%	0,368

Na Tabela 4.8 mantém-se a tendência de aumento do fator de correção à medida que o tamanho do cadáver em teste aumenta, assim como o facto de os valores médios do fator, no geral, serem mais elevados para a estação "Inverno" e mais baixos para a estação "Verão". Comparando com a Tabela 4.7, o fator de correção tem valores mais elevados,

apesar de se manter entre 0,120 e 0,365, o que é uma escala semelhante à anterior. Nesta tabela verifica-se claramente que os valores do fator de correção para Foros de Salvaterra estão muito mais próximos e praticamente indistinguíveis, nalgumas combinações, do fator de correção médio para o território nacional, ao contrário do que acontece para a localização em Pombal. Isto indica que para esta zona de Santarém, em que não há informação sobre os testes de remoção, o modelo aproxima as estimativas à média dos dados de remoção originais disponíveis, ou seja, os valores do fator de correção para Foros de Salvaterra não estão tão diferenciados da média nacional como para Pombal, onde se consegue fazer uma melhor estimativa. Esta situação também é visível nos valores do coeficiente de variação uma vez que ao serem mais baixos do que os valores da Tabela 4.7 indicam que os valores calculados nos 100 modelos ajustados rondaram sempre a sua média, apresentada na 1ª coluna.

Tal como referido, foi possível calcular os valores das funções de sobrevivência para cada uma das 12 combinações de estação do ano e tamanho e representar as suas curvas para Pombal e Foros de Salvaterra (Figura 4.11 e Figura 4.12).

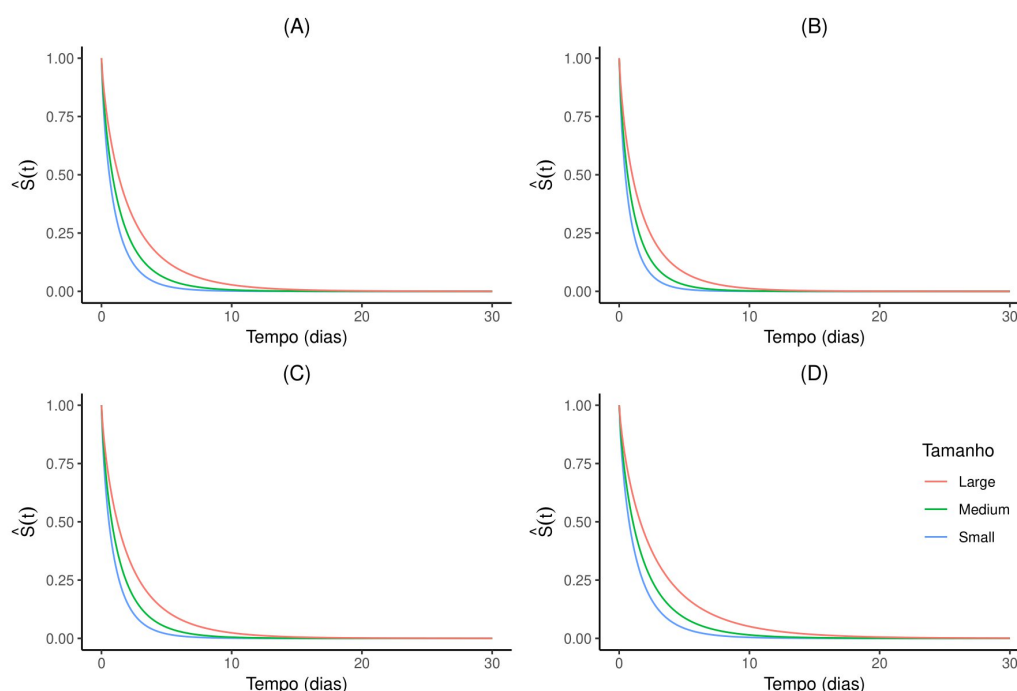


Figura 4.11: Curvas de sobrevivência associada a cada tamanho do cadáver para a estação "Primavera" (A), "Verão" (B), "Outono" (C) e "Inverno" (D), em Pombal.

Em Pombal (Figura 4.11) observa-se que a sobrevivência dos cadáveres, isto é, a sua persistência no terreno decai mais rapidamente na estação "Verão", sendo praticamente igual a 0 logo a partir do 7º dia para todos os tamanhos, e mais lentamente na estação "Inverno", em que só ao 10º dia é que isso acontece. De acordo com aquilo que tem

tido analisado e também com a estimativa da sobrevivência empírica, da Figura 4.3, a sobrevivência do tamanho "Pequeno" é sempre inferior às restantes.

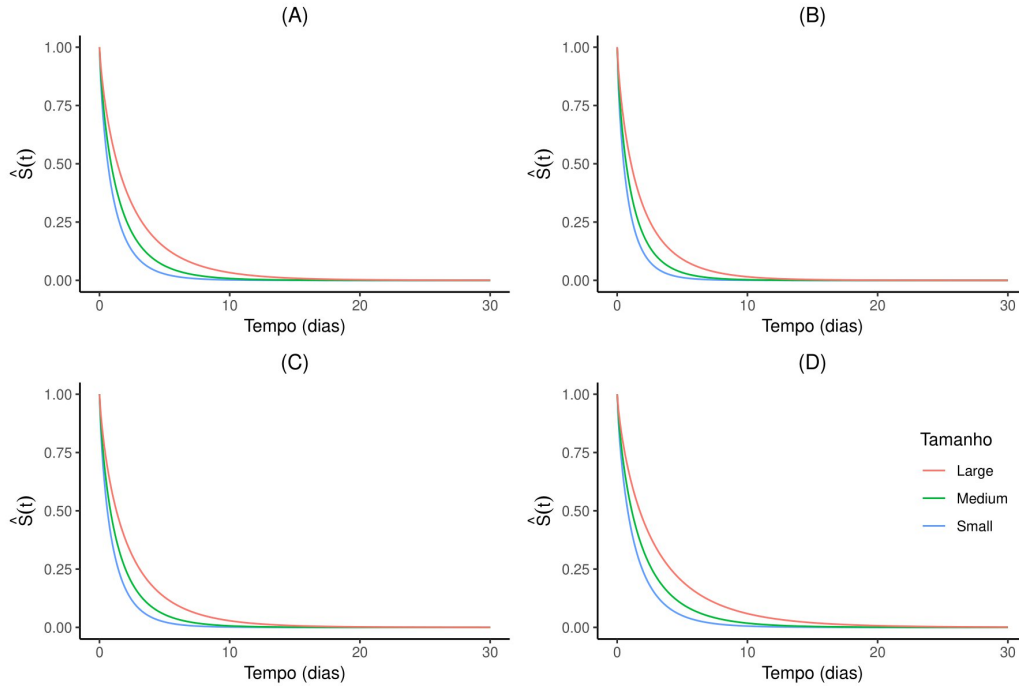


Figura 4.12: Curvas de sobrevivência associada a cada tamanho do cadáver para a estação "Primavera" (A), "Verão" (B), "Outono" (C) e "Inverno" (D), em Foros de Salvaterra.

Em Foros de Salvaterra (Figura 4.12) observam-se as tendências semelhantes a Pombal, dado que novamente para a estação "Verão", a sobrevivência decai para perto de 0 logo ao 8º dia, para todos os tamanhos, mas para a estação "Inverno" isso só acontece, após o 10º dia. Apesar de ser uma diferença muito ligeira, repara-se que as curvas de sobrevivência desta localização estão um pouco acima das de Pombal, indicando que um cadáver em Foros de Salvaterra persistirá mais tempo no terreno.

CONCLUSÃO

A mortalidade de aves associada à colisão com os cabos aéreos das Linhas de Muito Alta Tensão, em Portugal, é um tema preocupante que tem vindo a ser avaliado através da realização de Programas de Monitorização, ao abrigo da legislação nacional de avaliação de impacto ambiental. Apesar de nestes programas se quantificar os cadáveres observados, sabe-se que muitos deles são removidos por predadores antes do seu registo pelo que é fundamental estimar a probabilidade de persistência de um cadáver no terreno, através de um fator de correção, que considere esta fonte de enviesamento na estimação da mortalidade real. Para isso, realizam-se *testes de remoção* em vários troços das linhas elétricas alvo de monitorização, que permitem obter dados sobre o tempo de persistência dos cadáveres, podendo estes dados ser censurados. Assim, o objetivo último definido para a presente dissertação prende-se com a criação de um modelo preditivo do fator de correção da remoção, estimado no paradigma bayesiano através da abordagem INLA, em função de algumas variáveis ambientais e a localização geográfica do projeto em Portugal Continental, por forma a evitar a realização de novos testes de campo a cada novo projeto.

Neste sentido, foi revista uma coletânea de artigos referente aos avanços e aplicações que se têm realizado na área da estimação da mortalidade e dos modelos de sobrevivência espaciais. A mortalidade de aves neste contexto foi estudada por alguns autores na perspetiva de se entender quais as variáveis explicativas, ambientais ou de outro tipo, que podiam ajudar a explicá-la e foi também apresentado o fator de correção usado nesta dissertação e proposto usando a abordagem de Análise de Sobrevivência. Na área de Análise de Sobrevivência, há um grande desenvolvimento de métodos para dados com censura intervalar que se mostraram relevantes e diferenciadores no estudo deste tipo de dados e diversos autores usaram modelos de Tempo de Vida Acelerado e fatores de correção em contextos semelhantes ao deste trabalho, com conclusões igualmente alinhadas. Por fim, para os métodos bayesianos, há uma grande exploração das vantagens do INLA em relação ao MCMC em modelos de sobrevivência com componente espacial, que se mostram bastante satisfatórias e com grande evolução no tópico de previsão de informação espacial.

Os dados trabalhados ao longo desta dissertação são registos dos tempos de sobrevivência até acontecer a remoção do cadáver do terreno que, devido às suas características, são considerados dados com censura intervalar. Neste âmbito foi realizada uma primeira análise exploratória não paramétrica, onde se aplicaram métodos adequados a este tipo de censura como o estimador de Turnbull ou o teste Log-Rank generalizado.

Depois, uma vez que as localizações dos cadáveres não eram conhecidas nos dados originais, foram realizadas 100 iterações de amostragem aleatória de pontos geográficos dentro dos troços do projeto associado a cada observação para criar o máximo de cenários possíveis de onde os cadáveres poderiam ter sido colocados e, com isso, ajustar um modelo de sobrevivência espacial a cada iteração. Assim, os dados foram modelados através de um modelo de Tempo de Vida Acelerado, em função das variáveis "Tamanho" e "Estação do ano", com efeitos aleatórios referentes ao projeto de linha elétrica associado a cada observação e com a componente espacial definida pelas coordenadas geográficas (aleatoriamente geradas) dos cadáveres e pela abordagem SPDE. Através da abordagem INLA, foi ajustado um modelo para cada iteração, isto é, foram ajustados 100 modelos no total, a partir dos quais se extraíram os coeficientes necessários para definir a função de sobrevivência que permitia calcular, a título ilustrativo, o fator de correção da remoção, assumindo uma situação em que as prospeções ocorrem com frequência semanal. No final, calculou-se a média dos valores do fator obtidos e a respetiva variância como medida de erro, tanto para todo o território nacional, como para duas localizações específicas, para as diferentes combinações das covariáveis.

Nos resultados da análise não paramétrica, confirmou-se que o tamanho e a estação do ano têm influência significativa na probabilidade de persistência de um cadáver no terreno, não só pelas conclusões deduzidas a partir dos gráficos das curvas de sobrevivência empírica como pela garantia dada pelos testes de hipóteses.

Nos resultados paramétricos, confirmou-se em primeiro lugar que o melhor modelo de sobrevivência espacial ajustado foi o Weibull em comparação com as restantes três distribuições testadas para o tempo de sobrevivência (Exponencial, Log-Logística e Log-Normal). Para Portugal, verificou-se que o modelo consegue fazer uma melhor caracterização do fator de correção em zonas onde há informação dos testes de remoção, sendo que para as zonas com menos conhecimento em termos dos dados o modelo atribui um valor mediano do fator. Para os valores do fator de correção médio nacional, a variabilidade não é muito expressiva. Além disso, verificou-se que o fator de correção é maior para cadáveres maiores e que o fator varia entre as estações do ano. Seguidamente, para cada uma das localizações específicas, observou-se que em Pombal, que está próximo de zonas com bastantes dados de remoção, o modelo consegue prever os valores médios do fator de correção autonomamente dos valores da média nacional enquanto para Foros de Salvaterra, que fica numa zona sem quaisquer dados, os valores são muito próximos

das médias nacionais. Esta situação era intuitivamente esperada e foi, de facto, verificada. Reparou-se também que os valores dos coeficientes de variação para as médias do fator de correção da remoção para ambas as localizações são consideravelmente baixos. Para cada uma das localizações específicas, foram também obtidas as curvas de sobrevivência num horizonte temporal de 30 dias que verificaram as tendências observadas nas curvas empíricas: a probabilidade de persistência de um cadáver é mais baixa para cadáveres pequenos e no verão decai mais rapidamente do que nas outras estações do ano.

Em termos de resultados metodológicos, a adição das várias componentes do modelo no software R foi bastante intuitiva e há uma vasta escolha na definição dessas componentes (SPDE, efeitos aleatórios simples, passeios aleatórios, entre outros) e nas distribuições *a priori* usadas. Além disso, relativamente à abordagem MCMC, o tempo de computação do INLA foi bastante reduzido para as mais de 40 milhões de iterações que eram pedidas. Contudo, ainda houve algumas desvantagens: o processo preditivo foi complexo devido à existência dos efeitos aleatórios relacionados com os projetos da linha elétrica de cada observação; o modelo é sensível a mudanças na malha de nós e nas distribuições *a priori* utilizadas, sendo que para este projeto se usou uma malha homogênea e as distribuições pré-definidas no software; e há uma falta de ferramentas para validar o modelo ajustado, que é um tema assumidamente em desenvolvimento pelos autores do INLA.

O próximo passo no contexto desta tese será a criação de um package no software R, destinado aos investigadores que estudam a correção da mortalidade através do fator de correção do enviesamento associado à remoção, onde se irá implementar toda a metodologia desenvolvida neste âmbito e que marca o início de um estudo promissor para a estimação do fator de correção da remoção. Pretende-se com este package que o utilizador final indique uma localização em Portugal Continental e o intervalo de tempo da sua prospeção em dias (ex. semanal - 7, quinzenal - 15, mensal - 30) e obtenha valores do fator de correção da remoção (com a respetiva medida da variabilidade) e a curva de sobrevivência associada a essa localização, em função de diferentes combinações de tamanho do cadáver (pequeno, médio e grande) e estação do ano pretendidos. Será ainda facultado um mapa da variação do fator em Portugal, também para o tamanho e estação pretendidos. Pode-se afirmar que esta plataforma interativa é o objetivo final do estudo que começa com esta dissertação.

Para além disto, existe ainda muito trabalho que pode ser desenvolvido neste contexto nomeadamente em tópicos como análise de sensibilidade às distribuições *a priori* utilizadas e à malha de nós criada, exploração de outros modelos para além do INLA que podem ser utilizados para diminuir o tempo de computação, estudo mais aprofundado da influência dos efeitos aleatórios associados ao código de projeto e dos métodos de previsão para localizações não observadas. Claramente que a metodologia aqui aplicada está limitada à qualidade dos dados fornecidos pela REN que, para além de serem escassos

relativamente à quantidade e dispersão pelo território nacional, têm a problemática de não terem as localizações originais, algo que foi remediado pela amostragem aleatória das coordenadas. No entanto, com estes procedimentos criados há a possibilidade de a base de dados vir a ser alimentada de forma continuada com mais informação e, por isso, forma-se a expectativa de gradualmente as previsões do modelo virem a ser melhoradas, nomeadamente para zonas que agora não têm dados.

Concluindo, esta dissertação estabelece uma metodologia de estudo aos dados dos testes de remoção da REN, algo que nunca tinha sido feito e que foi abordado com um método inovador e bastante recente. O trabalho aqui desenvolvido, com dados e características desafiantes, propõe uma linha de raciocínio inicial para que os recursos da REN sejam rentabilizados e utilizados numa perspetiva sustentável da estimação da mortalidade de aves em Linhas de Muito Alta Tensão.

REFERÊNCIAS BIBLIOGRÁFICAS

- Barrientos, R., Martins, R. C., Ascensão, F., D'Amico, M., Moreira, F., & Borda-de-Água, L. (2018). A review of searcher efficiency and carcass persistence in infrastructure-driven mortality assessment studies. *Biological Conservation*, 222, 146–153. <https://doi.org/10.1016/j.biocon.2018.04.014> (ver pp. 1, 3)
- Bernardino, J., Bevanger, K., Barrientos, R., Dwyer, J., Marques, A., Martins, R., Shaw, J., Silva, J., & Moreira, F. (2018). Bird collisions with power lines: State of the art and priority areas for research. *Biological Conservation*, 222, 1–13 (ver pp. 1, 3).
- Bernardino, J., Bispo, R., Costa, H., & Mascarenhas, M. (2013). Estimating bird and bat fatality at wind farms: a practical overview of estimators, their assumptions and limitations. *New Zealand Journal of Zoology*, 40(1), 63–74. <https://doi.org/10.1080/03014223.2012.758155> (ver p. 4)
- Bernardino, J., Martins, R. C., Bispo, R., Marques, A. T., Mascarenhas, M., Silva, R., & Moreira, F. (2022). Ecological and methodological drivers of persistence and detection of bird fatalities at power lines: Insights from multi-project monitoring data. *Environmental Impact Assessment Review*, 93, 106707. <https://doi.org/10.1016/j.eiar.2021.106707> (ver pp. 2, 4, 33)
- Bispo, R., Bernardino, J., Marques, T. A., & Pestana, D. (2013). Modeling carcass removal time for avian mortality assessment in wind farms using survival analysis. *Environmental and Ecological Statistics*, 20, 147–165. <https://doi.org/10.1007/s10651-012-0212-5> (ver pp. 4, 6, 19, 20)
- Bivand, R. S., & Gómez-Rubio, V. (2021). Spatial survival modelling of business re-opening after Katrina: Survival modelling compared to spatial probit modelling of re-opening within 3, 6 or 12 months. *Statistical Modelling*, 21(1-2), 137–160. <https://doi.org/10.1177/1471082X20967158> (ver p. 9)
- Blangiardo, M., & Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley. (Ver pp. 24, 26).
- CIBIO. (2020). *Manual para a monitorização de impactes de linhas de muito alta tensão sobre a avifauna e avaliação da eficácia das medidas de mitigação*. Cátedra REN em Biodiversidade. (Ver pp. 1, 2, 33, 34).

- Collett, D. (2015). *Modelling Survival Data in Medical Research* (Third). CRC Press. (Ver pp. 16, 19, 20).
- Cressie, N. A. C. (1993). *Statistics for Spatial Data, Revised Edition*. Wiley. (Ver p. 29).
- Dalthorp, D., Madsen, L., Huso, M., Rabie, P., Wolpert, R., Studyvin, J., Simonis, J., & Mintz, J. (2018). GenEst statistical models - A generalized estimator of mortality. <https://doi.org/DOI:10.3133/tm7A2> (ver p. 4)
- Fay, M. P., & Shaw, P. A. (2010). Exact and asymptotic weighted logrank tests for interval censored data: The interval R Package. *Journal of Statistical Software*, 36(2), 1–34. <https://doi.org/10.18637/jss.v036.i02> (ver p. 36)
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24, 997–1016. <https://doi.org/10.1007/s11222-013-9416-2> (ver p. 19)
- Giolo, S. R. (2004). Turnbull's nonparametric estimator for interval-censored data. *Department of Statistics, Federal University of Paraná*, 1–10 (ver pp. 16, 59).
- Gomez-Rubio, V. (2020). *Bayesian inference with INLA* (First). Chapman; Hall/CRC. (Ver pp. 21, 23, 24, 27).
- Hijmans, R. J. (2023). *raster: Geographic Data Analysis and Modeling* [R package version 3.6-23]. <https://CRAN.R-project.org/package=raster>. (Ver p. 36)
- ICNB. (2010). *Manual de apoio à análise de projetos relativos à instalação de linhas aéreas de distribuição e transportes de energia elétrica*. Instituto da Conservação da Natureza e Biodiversidade. (Ver pp. 2, 33, 34).
- Instituto da Conservação da Natureza e Florestas - Rede Natura. (2023). Obtido 2023-01-27, de <https://www.icnf.pt/biodiversidade/natura2000/redenatura>. (Ver p. 1)
- Jain, A., Kerlinger, P., Curry, R., & Slobodnik, L. (2007). Annual Report for the Maple Ridge Wind Power Project Post-construction Bird and Bat Fatality Study - 2006 (ver p. 4).
- Kang, S. Y., Battle, K. E., Gibson, H. S., Ratsimbaoa, A., Randrianarivelosia, M., Ramboarina, S., Zimmerman, P. A., Weiss, D. J., Cameron, E., Gething, P. W., & Howes, R. E. (2018). Spatio-temporal mapping of Madagascar's Malaria Indicator Survey results to assess *Plasmodium falciparum* endemicity trends between 2011 and 2016. *BMC Medicine*, 16(71). <https://doi.org/10.1186/s12916-018-1060-4> (ver pp. 10, 11)
- Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text* (Third). Springer New York, NY. (Ver pp. 13, 14, 16, 19, 20).
- Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., & Rue, H. (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA* (First). Chapman; Hall/CRC. (Ver pp. 29, 31).
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86. <https://doi.org/10.1214/aoms/1177729694> (ver p. 22)
- Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime data* (Second). John Wiley Sons. (Ver p. 19).
- Lindgren, F., & Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63(19), 1–25. <http://www.jstatsoft.org/v63/i19/> (ver pp. 29, 31, 36)

- Lindgren, F., Rue, H., & Lindström, J. (2011). An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach (with discussion). *Journal of the Royal Statistical Society B*, 73(4), 423–498 (ver pp. 29, 36).
- Lonsinger, R. C., Gese, E. M., Knight, R. N., Johnson, T. R., & Waits, L. P. (2016). Quantifying and correcting for scat removal in noninvasive carnivore scat surveys. *Wildlife Biology*, 22(2), 45–54 (ver p. 8).
- Lourenço, J. M. (2021). *The NOVAthesis L^AT_EX Template User's Manual*. NOVA University Lisbon. <https://github.com/joamlourenco/novathesis/raw/master/template.pdf>. (Ver p. iii)
- Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3), 163–170 (ver p. 17).
- Martino, S., Akerkar, R., & Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics*, 38(3), 514–528. <https://doi.org/10.1111/j.1467-9469.2010.00715.x> (ver pp. 8, 9)
- Martins, R. C., Bernardino, J., & Moreira, F. (2023). A review of post-construction monitoring practices used in the evaluation of transmission power line impacts on birds and mitigation effectiveness, with proposals for guideline improvement. *Environmental Impact Assessment Review*, 100, 107068. <https://doi.org/10.1016/j.eiar.2023.107068> (ver pp. 1, 2)
- Moraga, P. (2019). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny* (First). Chapman; Hall/CRC Biostatistics Series. (Ver pp. 22, 24, 29, 31).
- Natário, I. C., & Carvalho, M. L. (2008). *Análise de Dados Espaciais: Congresso Anual da Sociedade Portuguesa de Estatística*. Sociedade Portuguesa de Estatística. (Ver p. 29).
- Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1), 439–446. <https://doi.org/10.32614/RJ-2018-009> (ver p. 36)
- Pebesma, E., & Bivand, R. (2023). *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. <https://doi.org/10.1201/9780429459016>. (Ver p. 36)
- Pedersen, T. L. (2023). *patchwork: The Composer of Plots* [R package version 1.1.3]. <https://CRAN.R-project.org/package=patchwork>. (Ver p. 36)
- R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org/>. (Ver p. 35)
- R-INLA Project. (2020). Obtido 2023-08-09, de <https://www.r-inla.org>. (Ver p. 36)
- Rocha, C., & Papoila, A. L. (2009). *Análise de Sobrevivência*. Sociedade Portuguesa de Estatística. (Ver pp. 13, 15).
- Rodrigues, A. S., Calsavara, V. F., Silva, F. I. B., Alves, F. A., & Vivas, A. P. M. (2018). Use of interval-censored survival data as an alternative to Kaplan-Meier survival curves: studies of oral lesion occurrence in liver transplants and cancer recurrence. *Applied Cancer Research*, 38(16). <https://doi.org/10.1186/s41241-018-0067-7> (ver p. 5)
- Rue, H., & Held, L. (2005). *Gaussian Markov Random Fields - Theory and Applications* (First). Chapman; Hall/CRC. (Ver p. 25).

- Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2), 319–392. <https://doi.org/10.1111/j.1467-9868.2008.00700.x> (ver pp. 23, 24, 28, 36)
- Rue, H., Riebler, A. I., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Reviews of Statistics and Its Applications*, 4(1), 395–421. <https://doi.org/10.1146/annurev-statistics-060116-054045> (ver p. 36)
- Seppä, K., Rue, H., Hakulinen, T., Läärä, E., Sillanpää, M. J., & Pitkaniemi, J. (2019). Estimating multilevel regional variation in excess mortality of cancer patients using integrated nested Laplace approximation. *Statistics in Medicine*, 38(5), 778–791. <https://doi.org/10.1002/sim.8010> (ver p. 10)
- Simpson, D., Rue, H., Riebler, A., Martins, T. G., & Sørbye, S. H. (2017). Penalising model Component Complexity : a principled, practical approach to constructing priors. *Statistical Science*, 32(1), 1–28. <https://doi.org/10.1214/16-STS576> (ver p. 22)
- Sun, J. (1996). A Non-Parametric Test For Interval-Censored Failure Time Data With Application to AIDS Studies. *Statistics in Medicine*, 15(13), 1387–1395. [https://doi.org/10.1002/\(SICI\)1097-0258\(19960715\)15:13<1387::AID-SIM268>3.0.CO;2-R](https://doi.org/10.1002/(SICI)1097-0258(19960715)15:13<1387::AID-SIM268>3.0.CO;2-R) (ver pp. 6, 17, 18)
- Terry M. Therneau & Patricia M. Grambsch. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer. (Ver p. 35).
- Therneau, T. M. (2023). *A Package for Survival Analysis in R* [R package version 3.5-5]. <https://CRAN.R-project.org/package=survival>. (Ver p. 35)
- Tobin, M. E., & Dolbeer, R. A. (1990). Disappearance and recoverability of songbird carcasses in fruit orchards. *Journal of Field Ornithology*, 61(2), 237–242. <http://www.jstor.org/stable/4513536> (ver p. 7)
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3), 290–295 (ver pp. 5, 16).
- van Niekerk, J., Bakka, H., & Rue, H. (2021). A principled distance-based prior for the shape of the Weibull model. *Statistics Probability Letters*, 174, 109098. <https://doi.org/10.1016/j.spl.2021.109098> (ver p. 23)
- Wang, X., Yue, Y. R., & Faraway, J. (2018). *Bayesian Regression Modeling with INLA* (First). Chapman; Hall/CRC. (Ver pp. 22, 24).
- Whittle, P. (1963). Stochastic-processes in several dimensions. *Bulletin of the International Statistical Institute*, 40(2), 974–994 (ver p. 29).
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>. (Ver p. 36)

ESTIMADOR DE TURNBULL

Neste anexo encontra-se o algoritmo desenvolvido por Giolo (2004) para a implementação do estimador Turnbull no software RStudio. Foram feitas ligeiras modificações ao código original para que se tornasse adequado ao dados em estudo nesta dissertação: a função `cria.tau`, que tinha como único argumento a base de dados completa, passou a ter os dois argumentos "left" e "right", assim como a função `cria.kappa` que passou a ter os mesmos argumentos que representam os instantes iniciais e finais dos intervalos de tempo registados na base de dados.

```
cria.tau <- function(left , right){
  l <- left
  r <- right
  tau <- sort(unique(c(l,r[is.finite(r)])))
  return(tau)}

S.ini <- function(tau){
  m<-length(tau)
  ekm<-survfit(Surv(tau[1:m-1],rep(1,m-1))~1)
  So<-c(1,ekm$surv)
  p <- -diff(So)
  return(p)}

cria.kappa <- function(left , right , tau){
  tau12 <- cbind(tau[-length(tau)],tau[-1])
  interv <- function(x,inf,sup) ifelse(x[1]>=inf & x[2]<=sup,1,0)
  kappa <- apply(tau12,1,interv ,inf=left ,sup=right)
  id.lin.zero <- which(apply(kappa==0, 1, all))
  if(length(id.lin.zero)>0) kappa <- kappa[-id.lin.zero , ]
  return(kappa)}
```

```
Turnbull <- function(p, kappa, data.right, eps=1e-3, iter.max=200,
```

```
verbose=FALSE){
  n<-nrow(kappa)
  m<-ncol(kappa)
  Q<-matrix(1,m)
  iter <- 0
  repeat {
    iter <- iter + 1
    diff<- (Q-p)
    maxdiff<-max(abs(as.vector(diff)))
    if (verbose)
      print(maxdiff)
    if (maxdiff<eps | iter >= iter.max)
      break
    Q<-p
    C<-kappa%*%p
    p<-p*((t(kappa)%*%(1/C))/n)
  }
  cat("Iterations_=", iter, "\n")
  cat("Max_difference_=", maxdiff, "\n")
  cat("Convergence_criteria: Max_difference_<_1e-3", "\n")
  dimnames(p)<-list(NULL, c("P_Estimate"))
  surv<-round(c(1,1-cumsum(p)), digits=5)
  right <- data.right
  if(any(!(is.finite(right)))){
    t <- max(right[is.finite(right)])
    return(list(time=tau[tau<t], surv=surv[tau<t]))
  }
  else
    return(list(time=tau, surv=surv))}
```

A função `cria.tau` cria a grelha de instantes Ω , de acordo com os instantes de tempo presentes na base de dados, ou seja, "left" irá corresponder aos valores da variável $R1$ e "right" irá corresponder aos valores da variável $R2$.

A função `S.ini` calcula a estimativa inicial para a função que sobrevivência que vai sendo atualizada pelo algoritmo, tal como descrito no capítulo 3. De acordo com este código, a estimativa inicial é simplesmente calculada através do estimador de Kaplan-Meier, implementado na função `survfit`.

A função `cria.kappa` cria a matriz de pesos κ_{nm} , usados posteriormente no algoritmo. Por fim, na função `Turnbull` gera-se o processo iterativo do algoritmo aplicando os valores

das funções definidas anteriormente e retornando os intervalos de tempo e o respectivo valor de probabilidade de sobrevivência.



Modelação da produtividade em função da localização e da utilização do espaço

2023

Abilidade de pensar e agir de forma eficaz no mundo

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço

Modelação da produtividade em função da localização e da utilização do espaço