



UNIVERSIDADE DE ÉVORA

Mestrado em Matemática e Aplicações

**SELECÇÃO DE MODELOS EM
AMOSTRAGEM POR DISTÂNCIAS**

Maria de Fátima Romão Morgado

Dissertação apresentada para obtenção do grau de mestre
em Matemática e Aplicações

Orientador: Professor Doutor Russell Gerardo Alpizar-Jara

Setembro 2008



UNIVERSIDADE DE ÉVORA

Mestrado em Matemática e Aplicações

**SELECÇÃO DE MODELOS EM
AMOSTRAGEM POR DISTÂNCIAS**

Maria de Fátima Romão Morgado

Dissertação apresentada para obtenção do grau de mestre
em Matemática e Aplicações



169 268

Orientador: **Professor Doutor Russell Gerardo Alpizar-Jara**

Setembro 2008

AGRADECIMENTOS

A realização desta dissertação durante este ano foi para mim uma árdua tarefa apenas concretizável devido à ajuda de algumas pessoas a quem não posso deixar de agradecer.

Em primeiro lugar quero agradecer ao meu orientador, Professor Doutor Russell Alpizar-Jara, pela enorme compreensão que sempre demonstrou, especialmente nos momentos em que me senti um pouco desanimada. Quero também agradecer-lhe por se mostrar sempre disponível e pelas suas explicações e sugestões dadas ao longo do ano.

Agradeço também ao Mestre José Saias, por me ter dado uma grande ajuda na programação com o MS-DOS e com o Cygwin, essencial para a realização das simulações.

Agradeço ainda ao Mestre João Filipe Monteiro, por me ter facultado alguma bibliografia importante para o trabalho.

Agradeço aos elementos do Júri as sugestões dadas durante a discussão da dissertação e que ajudaram a melhorar alguns aspectos do texto.

Por fim quero agradecer aos meus pais, Brígida e António, por se mostrarem sempre compreensivos pela minha “quase ausência” no seio familiar, durante o ano.

A todos o meu profundo MUITO OBRIGADA.

No decurso desta investigação, foi apresentada no *XXIVth International Biometric Conference*, que decorreu em Dublin de 13 a 18 Julho de 2008, uma comunicação (poster) intitulada “*Model selection in distance sampling*”.

Foi também apresentada no XVI Congresso SPE, que decorreu em Vila Real de 1 a 4 de Outubro de 2008, uma comunicação oral com o título “Seleccção de modelos em amostragem por distâncias”.

SUMÁRIO

A selecção do modelo que se ajuste adequadamente a um conjunto de dados é uma parte fulcral no processo da inferência estatística. O critério mais comum para seleccionar uma função de detectabilidade, que se ajuste a um conjunto de dados de distâncias, baseado na informação K-L, denomina-se critério de informação de Akaike (AIC). Para avaliar a performance da rotina da selecção de modelos implementada no programa DISTANCE, realizaram-se simulações Monte Carlo para gerar dados de duas funções densidade de probabilidade frequentemente usadas. Aplicando esta metodologia, observou-se um enviesamento substancial na selecção do modelo, sendo os estimadores dos parâmetros baseados no modelo seleccionado frequentemente enviesados. Utilizaram-se duas abordagens na estimação dos parâmetros: selecção de modelos tradicional e ponderação de modelos. Para ilustrar a metodologia da selecção de modelos usaram-se dados provenientes de um estudo controlado.

Palavras-chave: Amostragem por distâncias, critério de informação de Akaike, ponderação de modelos, selecção de modelos.

MODEL SELECTION IN DISTANCE SAMPLING

ABSTRACT

The selection of a model that adequately fits a data set is an important part in the statistical inference process. The most commonly used criterion to select a detectability function that fits a distance sampling data set, is based on the K-L information, and it is known as the Akaike information criterion (AIC). In order to evaluate the performance of the model selection routine implemented in software DISTANCE, we carried out Monte Carlo simulations generating data from two frequently used probability density functions. We found substantial model selection biases when using this methodology. We further observed that parameter estimators based on the selected model were often biased. Parameter estimation was conducted under two approaches: traditional model selection and model averaging. We used data from a controlled field experiment to illustrate the model selection methodology.

Key words: Akaike information criterion, distance sampling, model averaging, model selection.

ÍNDICE GERAL

AGRADECIMENTOS.....	i
SUMÁRIO.....	iii
ABSTRACT.....	iv
ÍNDICE DE FIGURAS.....	vii
ÍNDICE DE QUADROS.....	ix
NOTAÇÃO E ABREVIATURAS.....	xi
1. INTRODUÇÃO.....	1
2. ALGUMAS ABORDAGENS À INFERÊNCIA ESTATÍSTICA E CRITÉRIOS DE SELECÇÃO DE MODELOS.....	5
2.1 Abordagens estatísticas.....	5
2.1.1 Abordagem clássica.....	5
2.1.2 Abordagem bayesiana.....	6
2.1.3 Teoria da informação.....	6
2.2 Selecção de modelos.....	8
2.2.1 Testes de hipóteses.....	10
2.2.2 Selecção de modelos baseada na informação K-L.....	11
2.2.2.1 Critério de informação de Akaike – AIC.....	12
2.2.2.2 Critério de Informação de segunda ordem – AIC _C	14
2.2.2.3 Critério de Informação modificado para dados sobredispersos – QAIC e QAIC _C	14
2.2.2.4 Diferenças de Akaike, Δ_i	15
2.2.3 Outros critérios de informação.....	16
3. INFERÊNCIA BASEADA EM VÁRIOS MODELOS - MMI.....	17
3.1 Pesos de Akaike.....	18
3.2 Estimação de parâmetros usando a ponderação de modelos.....	18
3.3 Incerteza na selecção de modelos.....	19
3.4 “Conjunto” de confiança para o melhor modelo.....	20
4. AMOSTRAGEM POR DISTÂNCIAS.....	22
4.1 Função de detectabilidade.....	22
4.2 Transectos lineares.....	24

4.3 Estimadores de D , N e $var(\hat{D})$	26
5. APLICAÇÃO DA PONDERAÇÃO DE MODELOS EM TRANSECTOS LINEARES.....	30
5.1 Análise dos resultados	31
6. RESULTADOS DAS SIMULAÇÕES	35
6.1 Resultados obtidos na simulação da <i>half-normal</i>	37
6.1.1 Enviesamento do AIC e das estimativas	37
6.1.2 Ponderação de modelos versus selecção de modelos tradicional	41
6.2 Resultados obtidos na simulação da <i>hazard-rate</i>	43
6.2.1 Enviesamento do AIC e das estimativas	43
6.2.2 Ponderação de modelos versus selecção de modelos tradicional	47
6.3 Comparação entre os resultados dos modelos <i>half-normal</i> e <i>hazard-rate</i>	49
6.4 Comparação entre os resultados obtidos com o AIC e com o AIC _C	50
7. CONCLUSÃO.....	55
APÊNDICE A	58
A.1 Dedução da fórmula da variância condicional.....	58
A.2 Dedução da variância de \hat{D}	59
APÊNDICE B.....	61
APÊNDICE C.....	66
C.1 Etapas realizadas desde a simulação até à análise dos dados	66
C.2 Algumas rotinas de programação utilizadas.....	67
C.3 Ficheiro de comandos do DISTANCE que são executados com o MCDS.exe....	71
BIBLIOGRAFIA.....	73

ÍNDICE DE FIGURAS

2.2.1. Princípio da parcimónia: quadrado do enviesamento (linha contínua) e variância (linha tracejada) versus número de parâmetros estimados do modelo.	8
3.4.1. Relação entre as diferenças de Akaike e a verosimilhança do modelo (evidência relativa do modelo).	21
4.2.1. Transectos lineares com a mesma direcção e localização aleatória na área em estudo.	24
4.2.2. Representação esquemática de como são medidas as distâncias.	25
4.3.1. Probabilidade de detectar um objecto dentro da área examinada.	27
4.3.2. Função densidade de probabilidade, $f(x)$, sobre o histograma das frequências das distâncias perpendiculares.	28
6.1.1.1 Percentagem de amostras em que cada modelo foi seleccionado para cada um dos quatro cenários (função simulada: <i>half-normal</i>).	38
6.1.1.2 Distribuição do enviesamento de $\hat{f}(0)$ para as amostras em que o modelo seleccionado produz estimativas mais enviesadas que os restantes (função simulada: <i>half-normal</i>).	41
6.1.2.1 Comparação das estimativas de $f(0)$ dadas pela ponderação de modelos e pela selecção de modelos tradicional (função simulada: <i>half-normal</i>).	42
6.2.1.1 Percentagem de amostras em que cada modelo foi seleccionado para cada um dos quatro cenários (função simulada: <i>hazard-rate</i>).	44
6.2.1.2 Distribuição do enviesamento de $\hat{f}(0)$ para as amostras em que o modelo seleccionado produz estimativas mais enviesadas que os restantes (função simulada: <i>hazard-rate</i>).	47
6.2.2.1 Comparação das estimativas de $f(0)$ dadas pela ponderação de modelos e pela selecção de modelos tradicional (função simulada: <i>hazard-rate</i>).	48
B.1 Modelo ajustado aos dados sobre o histograma para o qual se obtém um melhor ajustamento.	61
B.2 Distribuições das percentagens enviesamento de $\hat{f}(0)$ considerando vários cenários, no caso da simulação da função densidade de probabilidade <i>half-normal</i>	64

B.3 Distribuições das percentagens enviesamento de $\hat{f}(0)$ considerando vários cenários, no caso da simulação da função densidade de probabilidade *hazard-rate*.65

ÍNDICE DE QUADROS

4.4.1 Funções chave e expansões em série disponíveis no DISTANCE	23
5.1.1 Modelo seleccionado para cada função chave.	31
5.1.2 Estimativas de D e N para cada modelo	31
5.1.3 Modelos analisados e respectivos AIC, Δ_i e w_i (inclui o modelo EN).	32
5.1.4 Modelos analisados e respectivos AIC, Δ_i e w_i (não inclui o modelo EN).	32
5.1.5 Estimativas dos parâmetros para a selecção de um único modelo e para a ponderação de modelos, considerando dois cenários.	34
6.1.1.1 Selecção do modelo <i>half-normal</i> nos quatro cenários estudados	37
6.1.1.2 Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série (função simulada: <i>half-normal</i>).	39
6.1.1.3 Percentagem de amostras em que o modelo seleccionado fornece estimativas de $f(0)$ mais enviesadas face aos restantes modelos (função simulada: <i>half-normal</i>).	40
6.1.2.1 Média e desvio-padrão da percentagem de enviesamento de $\hat{f}(0)$ obtida nos quatro cenários, para a selecção de modelos tradicional e para a ponderação de modelos (função simulada: <i>half-normal</i>).	43
6.2.1.1 Selecção do modelo <i>hazard-rate</i> nos quatro cenários estudados	43
6.2.1.2 Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série (função simulada: <i>hazard-rate</i>).	45
6.2.1.3 Percentagem de amostras em que o modelo seleccionado fornece estimativas mais enviesadas face aos restantes modelos (função simulada: <i>hazard-rate</i>).	46
6.2.2.1 Média e desvio-padrão da percentagem de enviesamento de $\hat{f}(0)$ obtida nos quatro cenários, para a selecção de modelos tradicional e para a ponderação de modelos (função simulada: <i>hazard-rate</i>).	49
6.4.1 Comparação entre os termos de penalização dos critérios AIC e AIC _C	51
6.4.2 Percentagem de amostras para as quais cada modelo é seleccionado, usando o AIC e o AIC _C , em cada cenário analisado (função simulada: <i>half-normal</i>). ..	51
6.4.3 Percentagem de amostras com enviesamento de $\hat{f}(0)$ entre -20% e 20%, para cada um dos cenários analisados, considerando a selecção de modelos	

tradicional e a ponderação de modelos (função simulada: <i>half-normal</i>).	52
6.4.4 Percentagem de amostras para as quais cada modelo é seleccionado, usando o AIC e o AIC _C , em cada cenário analisado. (função simulada: <i>hazard-rate</i>).	53
6.4.5 Percentagem de amostras com enviesamento de $\hat{f}(0)$ entre -20% e 20%, para cada um dos cenários analisados, considerando a selecção de modelos tradicional e a ponderação de modelos (função simulada: <i>hazard-rate</i>).	54
B.1. Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série respeitante à simulação da função densidade de probabilidade <i>half-normal</i>	62
B.2. Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série respeitante à simulação da função densidade de probabilidade <i>hazard-rate</i>	63

NOTAÇÃO E ABREVIATURAS

A – Área ocupada pela população em estudo.

a – Área coberta.

AIC – Critério de informação de Akaike.

AIC_C – Critério de informação de Akaike de segunda ordem.

AIC_{min} – Critério de informação de Akaike correspondente ao modelo que apresenta o menor valor para o AIC, de entre um conjunto de R modelos considerados. Dados os modelos f_{2_1}, \dots, f_{2_R} , se o melhor modelo for f_{2_3} então $min=3$.

CV – Coeficiente de variação.

D – Densidade populacional.

\hat{D} – Estimador de D .

\mathcal{D} – Distância ou informação Kullback-Leibler.

Δ_i – Diferenças de Akaike. Determinam-se através da diferença entre o valor de um AIC e o valor do menor AIC que se encontra no conjunto dos R modelos considerados. $\Delta_i = AIC_i - AIC_{min}$ $i = 1, \dots, R$.

EN – Modelo exponencial negativo.

$E(\hat{\theta})$ – Esperança matemática do estimador $\hat{\theta}$.

$f(0)$ – Valor da função densidade de probabilidade no ponto de abcissa zero (a distância zero corresponde aos objectos que se encontram sobre a linha central do transecto).

$\hat{f}(0)$ – Estimador de $f(0)$.

$f(x)$ – Função densidade de probabilidade dos dados das distâncias perpendiculares, sabendo que o objecto foi detectado.

$f_1(\underline{x})$ – Notação utilizada para designar a função “verdadeira”. $f_1(\underline{x})$ é uma função densidade de probabilidade.

$f_2(\underline{x})$ – Notação utilizada para designar o modelo aproximado da função verdadeira. $f_2(\underline{x})$ é uma função densidade de probabilidade.

$g(0)$ – Probabilidade de detectar um objecto sobre a linha central do transecto (como pressuposto assume-se $g(0) = 1$).

$g(x)$ – Probabilidade de detectar um objecto sabendo que se encontra a uma distância x da linha. $g(x)$ é a função de detecção e em geral não é uma função densidade de probabilidade.

HAZ – Modelo *hazard-rate*.

HN – Modelo *half-normal*.

k – Número de parâmetros do modelo aproximado.

L – Comprimento do transecto linear ou, no caso de serem usados mais do que um transecto, soma dos comprimentos dos vários transectos considerados.

$\mathcal{L}(\underline{\theta}|\underline{x}, f_2)$ – Função de verosimilhança dos parâmetros do modelo, dada os dados \underline{x} e o modelo f_2 .

K-L – Kullback-Leibler.

μ – Área abaixo da curva que representa a função de detecção $g(x)$.

$\hat{\mu}$ – Estimador de μ

n – Tamanho da amostra.

N – Tamanho da população.

\hat{N} – Estimador de N .

ω_i – Pesos de Akaike. Medida para interpretar a verosimilhança relativa de cada modelo.

P_a – Proporção de objectos detectados na área observada.

R – Número de modelos candidatos a melhor modelo que se encontram no conjunto.

$\underline{\theta}$ – Representa um parâmetro genérico. Pode ser um escalar ou um vector.

$\hat{\underline{\theta}}$ – Estimador do parâmetro $\underline{\theta}$. Pode ser um escalar ou um vector.

$\underline{\theta}_0$ – Valor óptimo do parâmetro para um certo modelo, dada uma amostra de tamanho fixo, mas ignorando resultados da estimação. Pode ser um escalar ou um vector.

Θ – Espaço paramétrico.

UNIF – Modelo uniforme.

w – Largura da área observada para cada lado da linha.

$\underline{x}, \underline{y}$ – É uma variável aleatória. Pode ser um escalar ou um vector

x_i – Distâncias perpendiculares, entre o objecto detectado e a linha central do transecto linear.

1. INTRODUÇÃO

Várias são as razões apresentadas pelos biólogos para monitorizarem as populações selvagens. Entre elas podem destacar-se a necessidade de definir a diversidade biológica de uma determinada área; a intenção de introduzir ou reintroduzir espécies num certo território; a possibilidade de uma espécie estar a ser explorada ou em risco de extinção; a potencialidade de algumas espécies poderem vir a tornar-se numa praga que afecte a agricultura ou mesmo alguns recursos naturais; os efeitos dos avanços do Homem em zonas mais ou menos protegidas; etc...

Para que a estimação da densidade ou abundância de populações animais em determinado habitat, seja a mais próxima da realidade possível, é imprescindível que as ferramentas colocadas à disposição dos biólogos forneçam estimativas com razoável grau de precisão. Um dos métodos, por eles utilizados, é a amostragem por distâncias.

Nesta dissertação abordar-se-á um dos métodos de amostragem por distâncias - transectos lineares. O observador percorre uma linha central e mede as distâncias perpendiculares entre a linha percorrida e o animal detectado. A estimação da densidade populacional ou da abundância baseia-se numa função de detectabilidade, tipicamente um modelo não linear que se ajuste o melhor possível ao histograma das distâncias recolhidas pelo observador. Os parâmetros do modelo são depois estimados usando o método da máxima verosimilhança implementado no software DISTANCE, versão 5.0 release 2 (Thomas *et al.*, 2006), talvez o mais utilizado por biólogos e investigadores de outras áreas. Este software permite também seleccionar um modelo, de entre vários modelos que oferecem bons ajustamentos, usando critérios baseados na teoria da informação. Assim, um “bom” modelo deverá ajustar-se bem aos dados e fornecer estimativas próximas do valor verdadeiro.

A selecção de modelos assume, portanto, um papel fundamental na realização da inferência estatística acerca da densidade de populações animais. Os testes de hipótese, os critérios bayesianos ou critérios baseados na teoria da informação são métodos frequentemente utilizados para seleccionar o modelo que servirá de base para a inferência. Enquanto nos testes de hipóteses se elabora uma hipótese nula que pode ser rejeitada com base num certo valor p arbitrário, na selecção de modelos baseada na

teoria da informação pode fazer-se inferência considerando várias hipóteses e seleccionar o “melhor modelo” ou os “melhores modelos” de entre um conjunto de vários modelos escolhidos *a priori* e que poderão descrever os dados em estudo.

A abordagem da selecção de modelos baseada na teoria da informação, segundo a literatura, tem vindo a ser cada vez mais utilizada em detrimento do tradicional método dos testes de hipóteses (Guthery *et al.*, 2005). A ascensão deste método de análise estatística, não tem sido pacífica, verificando-se na literatura algumas críticas por parte dos defensores da abordagem dos testes de hipóteses. Por outro lado, têm surgido também críticas relativas à metodologia dos testes de hipóteses principalmente na última década (Stephens *et al.*, 2005; Lukacs *et al.*, 2007). O número de artigos publicados, em várias áreas (ecologia, medicina, economia, estatística, ...), que questionam a utilidade dos testes de hipóteses tem crescido (Anderson *et al.*, 2000). Stephens *et al.* (2005), indicam também algumas falhas da selecção de modelos baseada na teoria da informação e afirmam que ambas as metodologias podem ser utilizadas em conjunto. Contrários a esta opinião são Burnham e Anderson (2002) afirmando que por assentarem em princípios diferentes não se devem misturar aquelas duas metodologias de selecção de modelos. Também Johnson e Omland (2004) são da mesma opinião que Burnham e Anderson (2002), ao declararem que os biólogos devem decidir quando é que devem utilizar os testes de hipóteses e quando é que é mais apropriado utilizarem a selecção de modelos baseada na teoria da informação.

Tendo por finalidade a escolha de um modelo parcimonioso, ou seja, que ajuste bem os dados e com poucos parâmetros, a selecção de modelos baseada na teoria da informação tem ao seu dispor vários critérios passíveis de efectuar aquela escolha. Um dos critérios mais utilizado na selecção de modelos é o critério de informação de Akaike (AIC) e seus derivados, baseado na informação de Kullback-Leibler e na teoria da máxima verosimilhança (Stephens *et al.*, 2007). Este é expresso por $AIC = -2\ln[\mathcal{L}(\hat{\theta}|\underline{x})] + 2k$, onde $\mathcal{L}(\hat{\theta}|\underline{x})$ é a função de verosimilhança avaliada na estimativa de máxima verosimilhança dos parâmetros do modelo, $\hat{\theta}$, e k é o número de parâmetros do modelo. O AIC mede a bondade do ajustamento do modelo aos dados, através da verosimilhança e inclui o termo ($2k$) que “penaliza” modelos com muitos parâmetros, obedecendo assim ao princípio da parcimónia. O modelo escolhido de entre um conjunto de modelos candidatos é o que apresenta um menor AIC, pois é o que

exibe uma menor “perda de informação”. É fundamental que os modelos seleccionados *a priori* sejam criteriosamente escolhidos pois se não estiverem muito próximos dos dados produzirão estimativas muito afastadas do valor verdadeiro.

Por vezes, quando se determina o AIC, verifica-se que os vários modelos considerados apresentam valores de AIC semelhantes, e portanto a escolha de um modelo pode tornar-se difícil. Neste caso, em vez de estimar os parâmetros de interesse usando apenas um modelo pode considerar-se o conjunto de modelos cujos AIC são semelhantes para efectuar a estimação. Este método designa-se ponderação de modelos e baseia-se nos pesos de Akaike, que representam a evidência empírica de cada modelo, para estimar os parâmetros (Buckland *et al.*, 1997).

Para realizar uma inferência credível é necessário que a escolha do modelo seja feita de forma acertada, ou seja, que o modelo se ajuste o melhor possível aos dados recolhidos presumivelmente de forma aleatória. Sendo assim, um dos objectivos desta dissertação consiste em avaliar se o software DISTANCE, frequentemente utilizado por investigadores de várias áreas, escolhe o “melhor modelo”, com base no AIC. Pretende-se portanto, investigar se o critério de informação de Akaike é enviesado na escolha do modelo, tal como foi referido em Stefanski *et al.* (manuscrito em revisão).

O segundo objectivo desta dissertação, consiste em caracterizar as situações (se existirem) em que a ponderação de modelos fornece estimativas dos parâmetros mais próximas do valor verdadeiro do que quando se selecciona um único modelo.

De forma a poder encontrar resposta para essas questões recorre-se a dados simulados com recurso ao software R. Com esses dados estimam-se os parâmetros de interesse usando a ponderação de modelos e o software DISTANCE, que fornece as estimativas dos parâmetros, tendo em conta o modelo com menor AIC.

A dissertação é composta por 7 capítulos cujos conteúdos são resumidos como se segue. O critério de selecção de modelos utilizado na dissertação é um critério baseado na teoria da informação, assim, no 2º capítulo é feita uma síntese de três abordagens estatísticas para realização de inferência: a clássica, a bayesiana e a teoria da informação, de forma a contextualizar e evidenciar as principais características de cada uma. Em cada uma daquelas abordagens existem critérios de selecção de modelos, porém será aqui dado maior destaque aos critérios baseados na teoria de informação. A inferência baseada em vários modelos (na literatura inglesa, *Multimodel inference*,

MMI) é apresentada no capítulo 3. Neste são apresentados alguns conceitos como ponderação de modelos, variância condicional e não condicional e a incerteza na escolha dos modelos. No capítulo 4 é focada a amostragem por distâncias e em particular o método de transectos lineares. Um exemplo de aplicação da ponderação de modelos, tendo por base os dados das estacas de madeiras da experiência de Laake (1978), é exposto no capítulo 5. Os resultados das simulações, que também ajudarão a dar resposta a algumas questões levantadas são apresentados no capítulo 6. Por fim, no capítulo 7 é apresentado um resumo das conclusões a que se chegaram.

A dissertação contém ainda três apêndices. No apêndice A são feitas deduções de algumas fórmulas citadas no texto e que são consideradas mais relevantes. No apêndice B são dadas algumas informações complementares, através de tabelas e gráficos referentes à análise dos dados efectuada nos capítulos 5 e 6. O apêndice C contém a descrição das etapas realizadas desde a simulação dos dados até à análise final dos mesmos, bem como algumas rotinas de programação utilizadas nessas etapas.

2. ALGUMAS ABORDAGENS À INFERÊNCIA ESTATÍSTICA E CRITÉRIOS DE SELECÇÃO DE MODELOS

2.1 Abordagens estatísticas

2.1.1 Abordagem clássica

A abordagem mais comum para realizar inferência estatística é a clássica. Esta abordagem é caracterizada por dois aspectos fundamentais: a única informação considerada relevante é a que emana da amostra recolhida e a construção e avaliação de procedimentos estatísticos tem por base o comportamento a longo prazo verificado sob condições semelhantes (Barnett, 1999).

A abordagem clássica encara a noção de probabilidade como sendo a frequência relativa de um determinado acontecimento, obtida através da realização de muitas experiências independentes, nas mesmas condições, ou em condições semelhantes.

Nesta perspectiva frequencista, os parâmetros de interesse têm um valor fixo, estimado através de amostras aleatórias repetidas infinitamente. Assim, assume-se que o parâmetro é igual ao valor médio obtido nas amostras. (Ellison, 1996).

Como referido em Hampel (1998), consideram-se duas principais escolas frequencistas: a teoria dos testes de hipóteses de Neyman-Pearson (Lehmann, 1986) e o trabalho de Fisher (1959) que contém dois ramos especiais: a abordagem da verosimilhança (Azzalini, 1996; Barnett, 1999) e as probabilidades fiduciais (Fisher, 1935a; Barnett, 1999).

A lei dos grandes números, o teorema do limite central e o conceito de distribuição de amostragem são algumas das ferramentas probabilísticas importantes utilizadas nesta abordagem.

2.1.2 Abordagem bayesiana

O Reverendo Thomas Bayes (citado em Barnett, 1999) foi o grande impulsionador da abordagem bayesiana. Esta é caracterizada pelos seguintes aspectos, que diferem substancialmente da abordagem clássica ou frequentista:

- Incorpora, nos cálculos, informação *a priori* sobre o parâmetro.
- Assume que o parâmetro de interesse não se mantém fixo, tem uma distribuição de probabilidade. Inicialmente esta é a distribuição *a priori* mas ao ser actualizada com os dados da amostra, transforma-se na distribuição *a posteriori*.
- Usa a função de verosimilhança de modo a expressar a probabilidade relativa de obter diferentes valores do parâmetro quando um conjunto de dados específico é observado (Ellison, 1996).

O Teorema de Bayes constitui a base fundamental da teoria bayesiana. Aquele afirma que a probabilidade *a posteriori*, $P(\underline{\theta}|\underline{x})$, de obter o parâmetro $\underline{\theta} \in \Theta$, dadas as observações, se determina da seguinte forma:

$$P(\underline{\theta}|\underline{x}) = \frac{P(\underline{x}|\underline{\theta}) \cdot P(\underline{\theta})}{P(\underline{x})} ,$$

onde $P(\underline{\theta})$ é a probabilidade de obter um determinado parâmetro, ou seja, a probabilidade esperada pelo investigador de observar $\underline{\theta}$ antes de realizar a experiência (*a priori*), $P(\underline{x}|\underline{\theta})$ é a função de verosimilhança de Fisher condicionada no parâmetro $\underline{\theta}$, e $P(\underline{x})$ é a distribuição marginal de \underline{x} , isto é, $P(\underline{x}) = \int P(\underline{x}|\underline{\theta}) \cdot P(\underline{\theta})d\underline{\theta}$ (no caso discreto, o integral é substituído por uma soma).

Este tipo de abordagem tem mais de dois séculos de existência, mas não tem sido muito utilizada devido à dificuldade de implementação analítica. No entanto, com o avanço da tecnologia dos computadores, ultimamente tem-se notando um aumento crescente da sua utilização. Mais detalhes sobre a metodologia bayesiana podem ser encontrados em Lee (1997) e Paulino *et al.* (2003).

2.1.3 Teoria da informação

O paradigma para realizar inferência estatística, baseado na teoria da informação surgiu na década de 70 do século passado e assenta em teorias publicadas nos anos 50 (Kullback & Leibler, 1951).

Porém o conceito de informação em estatística surgiu há já muito tempo tendo sido Fisher um dos primeiros a tentar definir o conceito de informação. Fisher (1935b) definiu a quantidade de informação contida na amostra, baseada na função de verosimilhança. A matriz de informação de Fisher, que representa essa quantidade, é definida, para qualquer $\underline{\theta} \in \Theta$, por:

$$I(\underline{\theta}) = E \left\{ - \left[\frac{\partial^2 \ln(f_2(\underline{x}|\underline{\theta}))}{\partial \theta_i \partial \theta_j} \right]_{\underline{\theta}=\underline{\theta}_0} \right\} \quad i, j = 1, \dots, k, \quad (2.1)$$

onde f_2 é o modelo de onde provêm os dados \underline{x} .

Mais tarde, surgiram outras medidas, como a de Shannon (1948), e a de Kullback e Leibler (1951), ambas relacionadas com o conceito de entropia de Boltzmann (citado em Burnham & Anderson, 2002), que foi o primeiro a estabelecer uma relação entre entropia e desordem molecular ou caos.

Kullback e Leibler (1951) quantificaram a informação perdida, quando se utiliza um modelo f_2 para aproximar a função verdadeira f_1 . Esta medida de informação perdida denomina-se informação de Kullback-Leibler (informação K-L) ou distância K-L e é definida, para funções contínuas, por

$$\mathcal{D}(f_1, f_2) = \int \ln \left(\frac{f_1(\underline{x})}{f_2(\underline{x}|\underline{\theta})} \right) f_1(\underline{x}) d\underline{x} = E \left[\ln \left(\frac{f_1(\underline{x})}{f_2(\underline{x}|\underline{\theta})} \right) \right], \quad (2.2)$$

onde $\underline{\theta} \in \Theta$ é um vector de dimensão k , \underline{x} é um vector aleatório e f_1 e f_2 são funções densidade de probabilidade em \underline{x} , n -dimensionais.

A distância K-L entre dois modelos f_1 e f_2 não é uma simples distância, é uma distância orientada pois a medida de f_1 para f_2 é diferente da medida de f_2 para f_1 .

A informação K-L é uma medida sempre positiva, excepto se f_1 e f_2 forem idênticos verificando-se $\mathcal{D}(f_1, f_2) = 0$ se e só se $f_1(\underline{x}) = f_2(\underline{x}), \forall \underline{x}$ (como se pode verificar em Burnham & Anderson, 2002:429).

A distância K-L entre dois modelos é uma quantidade fundamental em ciência e teoria da informação sendo uma base lógica para a selecção de modelos, aliada à inferência da verosimilhança (Burnham & Anderson, 2002).

Este paradigma de análise de dados baseado na informação Kullback-Leibler é considerado uma extensão da teoria da verosimilhança (Akaike, 1973).

2.2 Seleção de modelos

Três princípios fundamentais guiam a inferência nas ciências, baseada em modelos (Burnham & Anderson, 2004):

- Simplicidade e parcimónia – A inferência baseada em modelos com poucos parâmetros pode ser enviesada, mas se for baseada em modelos com muitos parâmetros pode oferecer pouca precisão. Deve, assim existir um equilíbrio entre o número de parâmetros do modelo, o enviesamento e a variância.

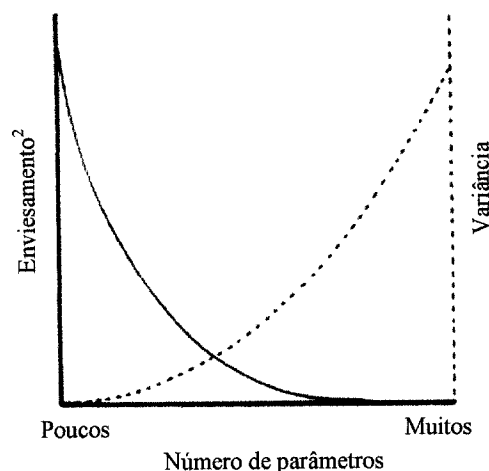


Figura 2.2.1. Princípio da parcimónia: quadrado do enviesamento (linha contínua) e variância (linha tracejada) versus número de parâmetros estimados do modelo. O melhor modelo tem dimensão próxima da intersecção das duas linhas (adaptada de Burnham & Anderson, 2001).

Este princípio é também referido como “Rasoura de Occam” devido ao seu intuito de considerar apenas o que for necessário. A finalidade é tornar os modelos o mais simples possível relativamente à sua estrutura, ao número de variáveis incluídas e ao número de parâmetros.

- Trabalhar com várias hipóteses – Segundo este princípio não existe hipótese nula podendo utilizar-se várias hipóteses plausíveis para o estudo (ou equivalentemente “modelos”) que competem entre si. Aquando da análise, um ou mais modelos podem revelar-se importantes para a realização da inferência enquanto outros poucos relevantes podem deixar de ser considerados.

- “Quantidade” de evidência – Nas ciências é fundamental existir informação quantitativa sobre o modelo que permita julgar a sua “quantidade de evidência”. Nos

testes de hipóteses tal não se verifica pois aqueles apenas fornecem uma dicotomia arbitrária: significativa ou não significativa.

Estes três princípios são aplicados quando se realiza a selecção de modelos usando a teoria da informação, ao contrário do que acontece com os testes de hipóteses.

A inferência estatística compreende três etapas: a especificação do modelo, a estimação dos parâmetros do modelo e a estimação da precisão dos estimadores.

A primeira etapa, é de extrema importância, pois sem um “bom modelo”, aquele que melhor explica o comportamento dos dados observados, não se conseguirão estimativas próximas dos valores reais dos parâmetros de interesse. Para a concretização da primeira etapa é necessário escolher um conjunto de modelos razoáveis para os dados em estudo, ajustar os modelos aos dados observados e por fim seleccionar o melhor modelo ou conjunto de melhores modelos. Seleccionar um modelo pressupõe comparar entre si os vários modelos escolhidos *a priori*. Na realização desta tarefa os investigadores usualmente utilizam uma das três técnicas estatísticas para o efeito: bondade do ajustamento, testes de hipóteses e critérios de selecção de modelos (baseados na teoria da informação ou na teoria bayesiana).

O modelo seleccionado usando a bondade do ajustamento, é aquele que originar um maior valor de R^2 ajustado. Este método favorece os modelos com mais parâmetros, ignorando o princípio da parcimónia. As estimativas dos parâmetros podem ser imprecisas.

Nos testes de hipóteses o método mais utilizado é o da razão de verosimilhanças (na literatura inglesa *likelihood ratio tests*, LRT) que consiste em comparar pares de modelos encaixados (ou seja, os modelos mais simples são casos especiais de um modelo mais complexo). É comparada a log-verosimilhança maximizada do modelo nulo (H_0) com a do alternativo (H_1), e no caso do valor p associado ser menor que um determinado valor de α (usualmente igual a 0.05), rejeita-se H_0 podendo afirmar-se que o modelo alternativo se ajusta significativamente melhor aos dados que o nulo. Com estes testes não é possível quantificar a importância relativa entre os modelos candidatos.

Através da selecção de modelos baseada em critérios de informação é possível ordenar os modelos candidatos, do melhor para o pior e indicar o peso relativo que cada um tem no conjunto de todos os modelos candidatos. Os critérios mais utilizados são o

AIC (critério de informação de Akaike) e o BIC (critério de informação bayesiano), mas existem muitos mais. O modelo seleccionado será o que originar um menor valor para o critério de informação.

Os testes de hipóteses e alguns critérios de selecção de modelos baseados na teoria da informação serão resumidos nas secções seguintes.

2.2.1 Testes de hipóteses

Uma hipótese estatística é uma afirmação acerca de um parâmetro da população ou acerca da distribuição da população. Um teste de hipóteses é um procedimento que tem por objectivo decidir, de entre duas hipóteses complementares, qual é a verdadeira. Uma das hipóteses denomina-se hipótese nula e representa-se por H_0 , enquanto a outra é denominada hipótese alternativa e representa-se por H_1 .

O teste de hipóteses é uma regra que especifica para que valores da amostra se mantém H_0 como verdadeira e para que valores da amostra H_0 é rejeitada em favor de H_1 . O subconjunto do espaço amostral para o qual H_0 é rejeitada chama-se região de rejeição ou região crítica, sendo o complementar daquela, a região de aceitação.

Para a realização do teste de hipóteses é necessário escolher uma estatística de teste cujo valor será calculado, com os dados da amostra, verificando-se em seguida se pertence à região de aceitação ou de rejeição.

Uma outra forma de tomar uma decisão relativamente à hipótese nula é comparar o valor p com o nível de significância α .

O valor p é um teste estatístico, cujo valor pertence a $[0,1]$. Pequenos valores do valor p evidenciam que H_1 é verdadeira. A regra será então rejeitar H_0 se valor $p \leq \alpha$. Neste caso, não é necessário identificar a região crítica.

A estatística do teste LRT, o mais utilizado para comparar pares de modelos encaixados tem a seguinte forma:

$$LRT = -2 \frac{\ln \left[\mathcal{L} \left(\hat{\theta}_r | f_2^*(\underline{x}) \right) \right]}{\ln \left[\mathcal{L} \left(\hat{\theta}_{r+s} | f_2(\underline{x}) \right) \right]} \sim \chi_s^2 .$$

onde $\mathcal{L} \left(\theta_r | f_2^*(\underline{x}) \right)$ e $\mathcal{L} \left(\theta_{r+s} | f_2(\underline{x}) \right)$ são verosimilhanças maximizadas dos modelos f_2^* e f_2 respectivamente. O modelo f_2^* tem menos parâmetros que o modelo global f_2 . O teste estatístico é aproximadamente um qui-quadrado cujo número de graus de liberdade

é a diferença entre os graus de liberdade dos dois modelos, assumindo que H_0 é verdadeira. Mais detalhes sobre os testes de hipóteses podem ser encontrados em Casella e Berger (2002).

2.2.2 Selecção de modelos baseada na informação K-L

Em primeiro lugar convém clarificar no que consiste a selecção de modelos baseada na teoria da informação.

Considere-se que se quer comparar um conjunto de R modelos f_{2i} , $i = 1, \dots, R$ com o modelo verdadeiro f_1 (um modelo que contém toda a informação sobre o processo ou o sistema em estudo, um modelo que representa totalmente a realidade). Pretende-se saber o quanto cada um dos modelos do conjunto se aproxima de f_1 . O objectivo é, portanto, determinar a quantidade de informação que se perde quando se utiliza um modelo para descrever a realidade, ou seja, pretende-se saber o valor de $\mathcal{D}(f_1, f_2)$, definida em (2.2), secção 2.1.3. No entanto, como a função verdadeira f_1 e os parâmetros do modelo f_2 que a aproxima não são conhecidos, a distância K-L entre f_2 e f_1 não pode ser determinada. Assim, a distância K-L tem que ser estimada.

O melhor modelo, de entre um conjunto de modelos prováveis, será aquele que minimiza a informação K-L em Θ , pois pretende-se perder o mínimo de informação possível, quando se faz a aproximação. Assim, o objectivo é encontrar um valor $\underline{\theta} \in \Theta$, representado por $\underline{\theta}_0$, que forneça o melhor modelo, isto é, que seja a solução do problema de optimização

$$\min_{\underline{\theta} \in \Theta} [\mathcal{D}(f_1, f_2)] = \int \ln \left(\frac{f_1(\underline{x})}{f_2(\underline{x}|\underline{\theta}_0)} \right) f_1(\underline{x}) d\underline{x}.$$

A informação K-L depende do parâmetro desconhecido, dada a forma do modelo. Como os dados da amostra representados por \underline{y} provêm de f_1 , pretende-se encontrar o estimador de máxima verosimilhança (na literatura inglesa *maximum likelihood estimator*, MLE) $\hat{\underline{\theta}}(\underline{y})$ e determinar uma estimativa de $\mathcal{D}(f_1, f_2(\cdot|\underline{\theta}_0))$ tal que

$$\mathcal{D}(f_1, f_2(\cdot|\hat{\underline{\theta}}(\underline{y}))) = \int \ln \left(\frac{f_1(\underline{x})}{f_2(\underline{x}|\hat{\underline{\theta}}(\underline{y}))} \right) f_1(\underline{x}) d\underline{x}. \quad (2.3)$$

Na igualdade (2.3), \underline{y} representa os dados, \underline{x} é a variável de integração, e ambas dizem respeito à função f_1 .

Por forma a simplificar a notação considera-se $\hat{\theta}(\underline{y}) = \hat{\theta}$. Usando as propriedades dos logaritmos e a definição de esperança matemática pode-se escrever (2.3) da seguinte forma:

$$\begin{aligned} \int \ln \left(\frac{f_1(\underline{x})}{f_2(\underline{x}|\hat{\theta})} \right) f_1(\underline{x}) d\underline{x} &= \int \ln(f_1(\underline{x})) f_1(\underline{x}) d\underline{x} - \int \ln(f_2(\underline{x}|\hat{\theta})) f_1(\underline{x}) d\underline{x} = \\ &= E_{\underline{x}} [\ln(f_1(\underline{x}))] - E_{\underline{x}} [\ln(f_2(\underline{x}|\hat{\theta}))]. \end{aligned}$$

A esperança matemática $E_{\underline{x}} [\ln(f_1(\underline{x}))]$ é constante pois depende apenas da função verdadeira (mas desconhecida) e será designada por C . Deste modo, para estimar $\mathcal{D}(f_1, f_2)$ é necessário apenas estimar $E_{\underline{x}} [\ln(f_2(\underline{x}|\hat{\theta}))]$.

Assim,

$$\mathcal{D}(f_1, f_2(\cdot|\hat{\theta})) = C - E_{\underline{x}} [\ln(f_2(\underline{x}|\hat{\theta}))]$$

e

$$E_{\underline{y}} [\mathcal{D}(f_1, f_2(\cdot|\hat{\theta}))] = C - E_{\underline{y}} [E_{\underline{x}} [\ln(f_2(\underline{x}|\hat{\theta}))]].$$

Note-se que ambas as esperanças (em relação a \underline{x} e a \underline{y}) são referentes à função verdadeira f_1 .

O objectivo é, portanto, encontrar um estimador para $E_{\underline{y}} E_{\underline{x}} [\ln(f_2(\underline{x}|\hat{\theta}(\underline{y})))]$.

A seguir apresenta-se o estimador daquela quantidade bem como alguns dos seus variantes.

2.2.2.1 Critério de informação de Akaike – AIC

Akaike (1973) descobriu uma relação entre a distância K-L e a máxima verosimilhança. Concluiu que era necessário estimar $E_{\underline{y}} E_{\underline{x}} [\ln(f_2(\underline{x}|\hat{\theta}(\underline{y})))]$ onde \underline{x} e \underline{y} são amostras aleatórias independentes da mesma distribuição e as esperanças matemáticas dizem ambas respeito à verdadeira função f_1 . Desta forma conseguiria um critério para a selecção de modelos.

No entanto, Akaike verificou que $\ln(\mathcal{L}(\hat{\theta}|\underline{x}))$ maximizada era enviesada positivamente, sendo o enviesamento aproximadamente igual ao número de parâmetros estimados, k . Portanto um estimador aproximadamente não enviesado de $E_{\underline{y}}E_{\underline{x}}\left[\ln\left(f_2\left(\underline{x}|\hat{\theta}(\underline{y})\right)\right)\right]$ para grandes amostras e “bons modelos”, aqueles que se ajustam bem aos dados, seria

$$\ln(\mathcal{L}(\hat{\theta}|\underline{x})) - k .$$

Akaike multiplicou ambos os termos da expressão por -2 , obtendo desta forma o chamado critério de Akaike,

$$AIC = -2\ln(\mathcal{L}(\hat{\theta}|\underline{x})) + 2k ,$$

onde k é o número de parâmetros estimados e $\mathcal{L}(\hat{\theta}|\underline{x})$ é o valor da função de verosimilhança no ponto correspondente ao MLE.

O critério de informação de Akaike pode também ser calculado recorrendo ao método dos mínimos quadrados, se todos os modelos do conjunto assumirem erros normalmente distribuídos e variância constante. Assim,

$$AIC = n\ln(\hat{\sigma}^2) + 2k$$

onde $\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n}$ e $\hat{\epsilon}_i$ são os resíduos estimados de cada modelo ajustado. Neste caso k é o número total de parâmetros da regressão estimados, incluindo a intersecção e σ^2 .

Considerando que existe um conjunto de R modelos candidatos *a priori* (escolhidos ainda antes de serem recolhidos os dados) bem definido e suportado pela ciência subjacente (que pode ser a biologia, ecologia, psicologia, economia, etc.), o AIC é calculado para cada um dos f_{2_i} modelos, $i = 1, \dots, R$. O modelo que apresentar o menor AIC é seleccionado como o melhor para os dados empíricos considerados (Burnham & Anderson, 2001).

Note-se que AIC não é um teste pois não existe nenhuma hipótese considerada nula, não existe nenhum nível α arbitrário nem é utilizada qualquer noção arbitrária de significância (Burnham & Anderson, 2001).

Em termos do critério AIC, considera-se um bom modelo aproximado aquele que se encontra mais próximo da função f_1 verdadeira, ou seja, o que tiver um menor valor para a distância K-L e conseqüentemente um menor valor de AIC. Assim ao longo da dissertação a expressão “melhor modelo” refere-se ao modelo que apresenta um menor valor de AIC.

2.2.2.2 Critério de Informação de segunda ordem – AIC_C

Sugiura (1978) propôs AIC_C , uma variante de AIC, pois este é enviesado quando o número de parâmetros considerados é grande relativamente ao tamanho da amostra. Hurvich e Tsai (1989) argumentaram que à medida que o número de parâmetros (k) do modelo aumenta relativamente ao tamanho da amostra (n), o AIC torna-se uma estimativa da informação K-L cada vez mais negativa, levando ao sob reajustamento. Assim, para corrigir o enviesamento acrescentou-se um termo ao AIC.

$$AIC_C = -2\ln(\mathcal{L}(\hat{\theta}|\underline{x})) + 2k + \frac{2k(k+1)}{n-k-1} = AIC + \frac{2k(k+1)}{n-k-1}$$

Para amostras grandes em relação ao número de parâmetros, o termo adicional tende para zero, e portanto AIC_C aproxima-se do AIC.

Burnham e Anderson (2002) aconselham a utilização de AIC_C , na análise de dados quando o tamanho da amostra é pequeno relativamente ao número de parâmetros. Este conselho estende-se também ao caso em que o número de parâmetros do modelo é muito grande. Afirmam ainda que não se devem misturar AIC e AIC_C , quando se faz a selecção de modelos.

2.2.2.3 Critério de Informação modificado para dados sobredispersos – $QAIC$ e $QAIC_C$

A sobredispersão é uma situação que acontece quando existem pequenas violações dos pressupostos como a independência e a homogeneidade dos parâmetros. Neste caso verifica-se que a variância amostral excede a variância teórica baseada no modelo assumido, (Lebreton *et al.*, 1992; Burnham & Anderson, 2002).

Apesar da sobredispersão dos dados, os estimadores dos parâmetros do modelo normalmente mantêm-se não enviesados mas a variância teórica baseada no modelo

sobrestima a precisão (McCullagh & Nelder, 1989). É portanto necessário modelar a sobredispersão e usar a teoria da quase-verosimilhança (Azzalini, 1996) como base para a análise de dados sobredispersos (McCullagh & Nelder, 1989; Lindsey, 1999).

Para contemplar o caso da sobredispersão quando se faz a selecção do modelo, Lebreton *et al.* (1992) propuseram o seguinte critério de selecção de modelos:

$$QAIC = -\frac{2\ln(\mathcal{L}(\hat{\theta}|\underline{x}))}{\hat{c}} + 2k$$

em que c é o factor de inflação da variância estimado por $\hat{c} = \frac{\chi^2}{df}$ (χ^2 é a estatística do teste qui-quadrado que mede a bondade do ajustamento do modelo global aos dados e df é o número de graus de liberdade desse modelo).

Para amostras de pequena dimensão relativamente ao número de parâmetros, tem que se acrescentar um termo de correcção, tal como acontece com o AIC. O critério a utilizar nesta situação será então

$$QAIC_c = -\frac{2\ln(\mathcal{L}(\hat{\theta}|\underline{x}))}{\hat{c}} + 2k + \frac{2k(k+1)}{n-k-1} = QAIC + \frac{2k(k+1)}{n-k-1}$$

Quando não existe sobredispersão, tem-se $c = 1$ e portanto QAIC e QAIC_c aproximam-se, respectivamente, de AIC e AIC_c.

2.2.2.4 Diferenças de Akaike, Δ_i

Os R modelos candidatos podem ser ordenados, do melhor para o pior, escalando os valores do AIC (ou AIC_c) tal que o modelo com menor critério de informação tenha valor zero, ou seja, determinando as diferenças de Akaike, para cada modelo f_{2i} , $i = 1, \dots, R$

$$\Delta_i = AIC_i - AIC_{min} .$$

Quanto maior for o valor de Δ_i menos aceitável é que o modelo f_{2i} seja o melhor modelo, dadas as observações \underline{x} .

Burnham e Anderson (2004) sugerem algumas regras empíricas, particularmente úteis para modelos encaixados, que permitem verificar o nível de suporte que o modelo tem no conjunto dos R modelos considerados: $\Delta_i \leq 2$, significa que existe um evidência

substancial de que o modelo i é o melhor modelo; $4 \leq \Delta_i \leq 7$ indica que o modelo i é menos razoável e $\Delta_i > 10$ aponta para modelos cujo suporte é essencialmente nulo.

2.2.3 Outros critérios de informação

Existem muitos outros critérios de informação, que não serão abordados nesta dissertação, por se encontrarem fora do âmbito do estudo. Entre eles, encontram-se o critério de informação bayesiano (BIC), proposto por Schwarz (1978); o critério HQ proposto por Hannan e Quinn (1979); o critério MDL (comprimento mínimo descrito) proposto por Rissanen (1989) e o critério de informação deviance (DIC) desenvolvido por Spiegelhalter *et al.* (2002).

3. INFERÊNCIA BASEADA EM VÁRIOS MODELOS - MMI

Habitualmente a selecção de modelos é vista como uma forma de seleccionar o modelo que melhor aproxima o modelo verdadeiro, de entre um conjunto de modelos candidatos. Deste modo, a inferência é condicional ao modelo seleccionado, condicional a um único modelo. Mas, se nenhum modelo do conjunto dos R modelos candidatos se destaca claramente, como sendo o modelo que melhor o aproxima, a ponderação de modelos pode ser a alternativa à selecção de um único modelo (Buckland *et al.*, 1997; Burnham & Anderson, 2002).

Em vez de se seleccionar apenas o melhor modelo pode utilizar-se todo o conjunto de modelos candidatos, ou um subconjunto deles, baseando desta forma a inferência nos vários modelos seleccionados. Assim, a inferência não é condicionada a um único modelo, mas será condicional ao conjunto de modelos utilizados para a realização da mesma. É de realçar, no entanto, que só é possível fazer inferência baseada no conjunto de modelos que tenham em comum o parâmetro de interesse ou em problemas de predição.

A escolha de um conjunto de modelos candidatos que representem a hipótese científica de interesse é muito importante para o processo de inferência, ou seja, normalmente, já se tem algum conhecimento *a priori*, do fenómeno em estudo, logo devem escolher-se modelos que sejam adequados a esse fenómeno. Por exemplo, na amostragem por distâncias a evidência empírica indica que a probabilidade de detecção dos indivíduos diminui com o aumento da distância a que se encontram do observador, portanto devem ser escolhidos modelos decrescentes. O conjunto de modelos escolhido deve ser preferencialmente pequeno e bem justificado pelo menos por estudos anteriores.

A inferência multi-modelos comporta as seguintes fases:

- ordenar os modelos ajustados, do melhor para o pior, baseando-se no valor do AIC ou Δ_i ;
- escalar os modelos de forma a obter um peso relativo, para cada modelo f_{2i} $i = 1, \dots, R$, através dos pesos de Akaike (ω_i), definidos na secção 3.1;

- estimar os parâmetros de interesse com base na ponderação dos modelos considerados, cujo estimador é apresentado na secção 3.2.

3.1 Pesos de Akaike

A ordenação dos modelos do conjunto de modelos candidatos a melhor modelo pode ser feita usando as diferenças de Akaike (Δ_i). Para poder quantificar a probabilidade do modelo f_{2i} , $i = 1, \dots, R$ ser o melhor modelo surgiu o conceito de verosimilhança do modelo, dadas as observações, $\mathcal{L}(f_{2i}|\underline{x})$. Esta verosimilhança, proporcional a $\exp\left(-\frac{1}{2}\Delta_i\right)$, representa uma medida relativa da evidência de cada modelo do conjunto. Assim,

$$\mathcal{L}(f_{2i}|\underline{x}) \propto \exp\left(-\frac{1}{2}\Delta_i\right)$$

(Burnham & Anderson, 2002:74).

Para melhor se poder interpretar a verosimilhança relativa de cada modelo, no conjunto dos R modelos normaliza-se $\mathcal{L}(f_{2i}|\underline{x})$. Esta medida, denominada pesos de Akaike, é dada por:

$$\omega_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^R \exp\left(-\frac{1}{2}\Delta_r\right)}$$

onde Δ_i são as diferenças de Akaike relativas ao modelo i .

Em Buckland *et al.* (1997) Δ (diferenças de Akaike) foi substituído por AIC na referida fórmula.

3.2 Estimação de parâmetros usando a ponderação de modelos

Se no conjunto dos R modelos considerados *a priori* não existir nenhum que se evidencie como o melhor, ou seja, que $\omega_i \geq 0.9$ (segundo Burnham & Anderson, 2002), pode fazer-se a inferência considerando todos os modelos do conjunto, em vez

de apenas um melhor modelo, com o mínimo AIC. Neste caso, a estimação dos parâmetros é feita com recurso aos pesos de Akaike.

O estimador do parâmetro de interesse é

$$\hat{\theta} = \sum_{i=1}^R \omega_i \hat{\theta}_i . \quad (3.1)$$

onde $\hat{\theta}_i$ é a estimativa do parâmetro θ considerando o modelo i e ω_i é o peso de Akaike que o modelo i tem no conjunto dos R modelos usados na estimação do parâmetro de interesse (Buckland *et al.*, 1997).

3.3 Incerteza na selecção de modelos

Ao estimar o parâmetro baseando o estudo num único modelo, a variabilidade observada diz apenas respeito a esse modelo, a variância é condicional ao modelo seleccionado como melhor modelo. Por outro lado se a estimação do parâmetro for feita com base nos R modelos candidatos, uma outra fonte de variabilidade terá que ser considerada, a que corresponde à incerteza na selecção de modelos. Assim, obtém-se uma variância, não condicional a um único modelo, mas sim ao conjunto de modelos considerados.

O estimador da variância de $\hat{\theta}$, que incorpora um termo para a variância amostral condicional ao modelo f_{2i} e um termo para a incerteza relativa à selecção do modelo é dado por

$$\widehat{var}(\hat{\theta}) = \left[\sum_{i=1}^R \omega_i \sqrt{\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2} \right]^2 . \quad (3.2)$$

onde $\hat{\theta}_i$ representa a estimativa do parâmetro obtida para θ a partir do modelo f_{2i} e $\hat{\theta}$ é a estimativa obtida para θ usando o estimador (3.1) (Buckland *et al.*, 1997).

A dedução da fórmula (3.2) é apresentada no apêndice A, secção A.1.

3.4 “Conjunto” de confiança para o melhor modelo

Burnham e Anderson (2002) propõem três formas de construir um conjunto de confiança para o melhor modelo.

A primeira proposta consiste em somar os pesos de Akaike dos vários modelos, do maior para o menor, até se atingir um valor maior ou igual a 0.95. Estes modelos constituirão o conjunto de confiança para o melhor modelo.

Uma segunda proposta para a construção do conjunto de confiança baseia-se na ideia de que Δ_i é uma variável aleatória com uma distribuição de amostragem.

Considere-se que o melhor modelo esperado está no conjunto de modelos candidatos e se denomina $f_{2_{best}}$. Então Δ_i de conceptual interesse é

$$\Delta_p = AIC_{best} - AIC_{min} .$$

O “ p ” em Δ_p significa que Δ é um valor pivotal conceptual em vez do actual Δ_i que se calcula a partir dos dados reais. A variável aleatória não observada Δ_p é análoga a $\theta - \hat{\theta}$ que pode ser usada (depois de normalizada por $\widehat{se}(\hat{\theta})$) como valor pivotal para a construção do intervalo de confiança de θ . Várias simulações foram realizadas por Burnham e Anderson (2002) tendo-se concluído que a distribuição de amostragem de Δ_p tem estabilidade e que normalmente o percentil 95 de Δ_p é obtido para valores de Δ muito menores que 10 e até geralmente menores que 7 (frequentemente perto de 4 nas situações mais simples). Portanto, a construção do conjunto de confiança de 95% pode basear-se na regra de que deve conter todos os modelos f_{2_i} , $i = 1, \dots, R$ com Δ_i menor ou igual a um determinado valor que se situa entre 4 e 7 (Burnham & Anderson, 2002).

Por fim, uma última proposta, considerada muito útil, assenta na inferência baseada na verosimilhança. O conjunto de confiança inclui todos os modelos para os quais se verifica a seguinte relação

$$\frac{\mathcal{L}(f_{2_i} | \underline{x})}{\mathcal{L}(f_{2_{min}} | \underline{x})} > \text{“ponto de corte”}$$

onde $\mathcal{L}(f_{2_i} | \underline{x})$ é a verosimilhança do modelo f_{2_i} e o “ponto de corte” pode ser aproximadamente $\frac{1}{8}$.

A escolha de $\frac{1}{8}$ para ponto de corte baseia-se no seguinte:

$$\begin{aligned} \frac{\mathcal{L}(f_{2_i} | \underline{x})}{\mathcal{L}(f_{2_{min}} | \underline{x})} &= \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\exp\left(-\frac{1}{2}\Delta_{min}\right)} = \exp\left(-\frac{1}{2}\Delta_i + \frac{1}{2}\Delta_{min}\right) = \exp\left[-\frac{1}{2}(\Delta_i - \Delta_{min})\right] = \\ &= \exp\left(-\frac{1}{2}\Delta_i\right) \end{aligned}$$

Note-se que $\Delta_{min} = 0$ porque representa a diferença de Akaike do modelo que fornece um menor AIC com ele próprio.

As regras empíricas apresentadas em Burnham e Anderson (2002) afirmam que existe ainda alguma evidência (com uma confiança de cerca de 95% segundo as simulações por eles realizadas) de que o modelo i é o melhor modelo, se $4 \leq \Delta_i \leq 7$. Assim,

$$\begin{aligned} 4 \leq \Delta_i \leq 7 &\Leftrightarrow \frac{4}{2} \leq \frac{1}{2}\Delta_i \leq \frac{7}{2} \Leftrightarrow -2 \geq -\frac{1}{2}\Delta_i \geq -3.5 \Leftrightarrow \\ &\Leftrightarrow \exp(-2) \geq \exp\left(-\frac{1}{2}\Delta_i\right) \geq \exp(-3.5) \Leftrightarrow 0.135 \geq \exp\left(-\frac{1}{2}\Delta_i\right) \geq 0.03 \end{aligned}$$

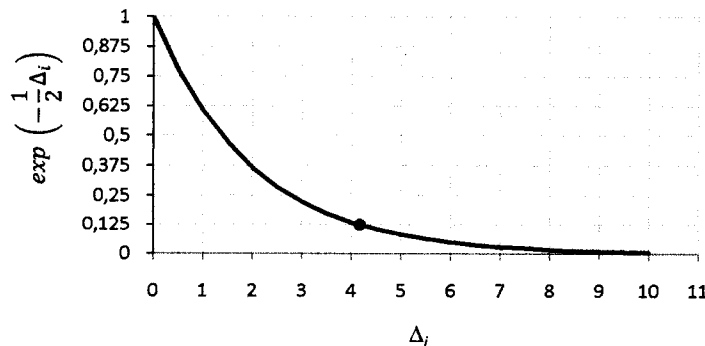


Figura 3.4.1. Relação entre as diferenças de Akaike e a verosimilhança do modelo (evidência relativa do modelo).

Como quanto mais afastado se encontrar Δ_i de 4 menor será a evidência do modelo, estabeleceu-se o ponto de corte igual a $\frac{1}{8}$ o qual corresponde a um $\Delta_i \approx 4.159$.

Esta última abordagem tem a vantagem de que a adição ou a eliminação de modelos não altera o valor do “ponto de corte”, e assim o conjunto de confiança é invariante. O mesmo não se verifica quando se utilizam os pesos de Akaike, onde uma alteração no conjunto dos R modelos candidatos pode alterar o conjunto de confiança quando se somam os pesos de Akaike, pois obriga a calcular novamente os pesos sempre que se adiciona ou elimina um modelo.

4. AMOSTRAGEM POR DISTÂNCIAS

Vários métodos têm sido utilizados para estimar a abundância ou densidade de populações animais. O mais recente é o método de amostragem por distâncias, que consiste na recolha de distâncias a que se encontra um objecto (animal ou planta) de uma linha (transecto linear) ou de um ponto (transecto pontual). Nem todos os objectos que se encontram na área examinada são detectados, mas um pressuposto fundamental da amostragem por distâncias refere que aqueles que estão sobre a linha central do transecto ou ponto são detectados com probabilidade 1. Intuitivamente, a detecção dos objectos diminui à medida que a distância a que se encontram do transecto ou do ponto, aumenta. Assim, o objectivo principal da amostragem por distância é encontrar uma função de detecção que se ajuste bem à distribuição das frequências das distâncias recolhidas pelo observador, a qual é depois usada para estimar a densidade populacional e/ou a abundância.

Portanto, o conceito de função de detecção é o alicerce da amostragem por distâncias. Um “bom” modelo para a função de detecção é de primordial importância, para a obtenção das estimativas.

No caso dos transectos lineares, a função de detecção é definida pela probabilidade de detectar um objecto, dado que se encontra a uma distância x do transecto, ou seja,

$$g(x) = \text{prob}\{\text{detectar}|\text{distância } x\}.$$

A função de detectabilidade não é em geral uma função densidade de probabilidade. É a partir daquela função que se estima a função densidade de probabilidade para estimar os parâmetros de interesse da população.

4.1 Função de detectabilidade

A função de detecção é desconhecida (sendo estimada a partir dos dados da amostra) e pode assumir formas variadas, mas o modelo para esta função deve possuir três características desejáveis: robustez, critério de forma e eficiência. A robustez, característica mais importante do modelo, significa que o modelo deve ser flexível,

podendo manter a sua forma genérica quando ajustada a vários conjuntos de dados. A segunda característica, indica que o modelo deve possuir um “ombro” no início, ou seja, a probabilidade de detectar um objecto perto da linha do transecto deve ser constante. Matematicamente, isto traduz-se em $g'(0) = 0$. A última característica está relacionada com a precisão dos estimadores derivados do modelo. É desejável que a variância desses estimadores seja pequena.

Buckland, *et al.* (2001) sugerem que a modelação da função de detecção passe por duas fases. A primeira, encarada como ponto de partida, consiste no ajustamento aos dados de uma função designada por função chave. A segunda fase, consiste em acrescentar à função chave uma expansão em série que melhorará o ajustamento, se não for suficiente utilizar apenas a função chave. Assim a função de detecção será da forma:

$$g(x) = \text{função chave}(x)[1 + \text{expansão em série}(x)].$$

No programa DISTANCE, software comumente utilizado para a análise de dados obtidos em amostragem por distâncias e utilizado na análise das simulações desta dissertação, estão disponíveis 4 funções chave (exponencial negativa, *half-normal*, *hazard-rate* e uniforme) e três expansões em série (co-senos, polinómios simples e polinómios Hermite, como é indicado no quadro 4.1.1. Cada função chave pode ser combinada com uma expansão em série.

Quadro 4.1.1. *Funções chave e expansões em série disponíveis no DISTANCE.*

Função chave	Expansão em série ^a
Uniforme: $\frac{1}{w}$	Co-seno: $\sum_{j=1}^m a_j \cos\left(\frac{j\pi x}{w}\right)$
<i>Half-normal</i> : $e^{-\frac{x^2}{2\sigma^2}}$	Polinómios simples: $\sum_{j=1}^m a_j \left(\frac{x}{w}\right)^{2j}$
<i>Hazard-rate</i> : $1 - e^{-\left(\frac{x}{\sigma}\right)^{-b}}$	Polinómios Hermite: $\sum_{j=1}^m a_j H_{2j}\left(\frac{x}{\sigma}\right)$
Exponencial Negativa: $e^{-\frac{x}{\lambda}}$	

^a As expansões em série podem ter que ser ligeiramente modificadas segundo a função chave associada.

Apesar de disponível no software DISTANCE, Buckland *et al.* (2001) não aconselha o uso do modelo exponencial negativo, a não ser em casos extremos em que houve uma recolha insuficiente de dados e exista uma forte suspeição de que os dados das distâncias são muito “aguçados”. Em detrimento desta sugere a utilização do modelo *hazard-rate*, cuja forma oferece uma maior flexibilidade de ajustamento e garante o critério de forma.

4.2 Transectos lineares

No método de amostragem por transectos lineares, o observador percorre uma linha e mede as distâncias perpendiculares a ela, a que o objecto é detectado. A localização da linha de comprimento conhecido L , na área em estudo A , assume-se aleatória. Se existirem t linhas, o comprimento total será $\sum_{i=1}^t l_i = L$, sendo l_i o comprimento de cada uma das linhas. Estas linhas podem ter todas a mesma direcção, como mostra a figura 4.2.1 ou direcções diferentes desde que não se interceptem (este delineamento não é muito usado por ser difícil de implementar) podendo também ter comprimentos diferentes (Burnham *et al.*, 1980).

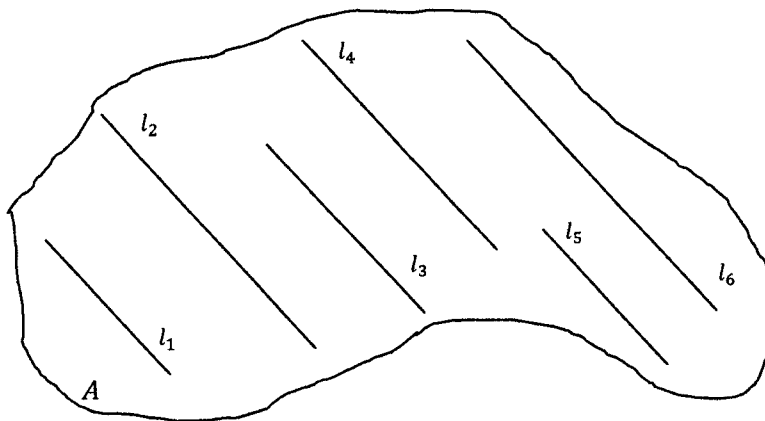


Figura 4.2.1. Transectos lineares com a mesma direcção e localização aleatória na área em estudo. A soma dos comprimentos das linhas é igual a L . (adaptada de Burnham *et al.*, 1980).

Como a partir de certa distância do transecto não é possível detectar objectos, pode definir-se que só se consideram os objectos detectados até uma distância máxima

w para ambos os lados do transecto, como é apresentado na figura 4.2.2. Tendo a área examinada a forma rectangular a superfície é igual a $a = 2wL$.

As estimativas dos parâmetros, como a densidade populacional e a abundância são determinadas com base nas observações recolhidas, x_i , as distâncias perpendiculares.

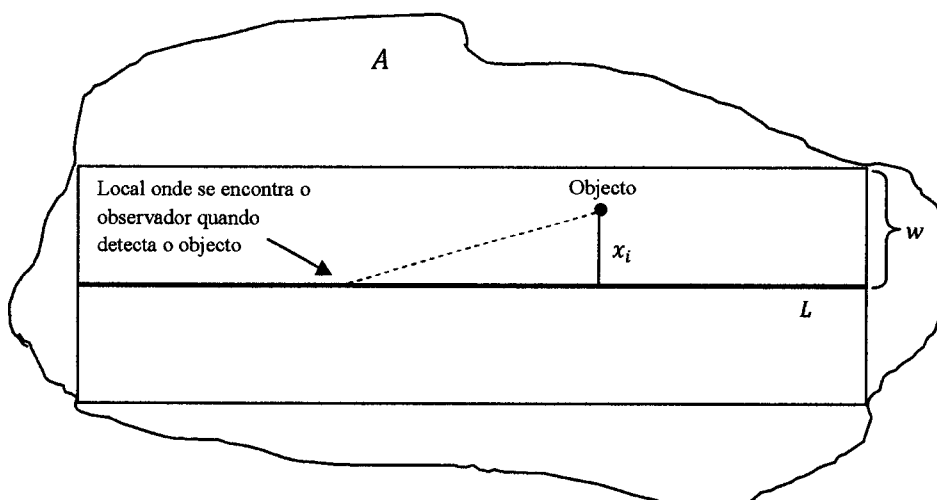


Figura 4.2.2. Representação esquemática de como são medidas as distâncias. Neste esquema utiliza-se apenas um transecto e define-se uma distância máxima de observação igual a w (adaptada de Buckland *et al.*, 2001).

Em Buckland *et al.* (2001) são enumerados os pressupostos do método de amostragem dos transectos lineares, que devem ser garantidos para obter estimativas, com confiança, dos parâmetros. Os pressupostos são aqui apresentados, de forma resumida, e por ordem de importância.

1 – A detecção na linha é certa, ou seja, $g(0) = 1$.

Todos os objectos que se encontram sobre a linha ou muito próximos dela deveriam ser detectados pelo observador. Se isso não acontecer o estimador da densidade populacional será negativamente enviesado, isto é, $E(\hat{D}) < D$.

2 – Os objectos não se movem até serem detectados.

Se os animais se moverem, o registo das distâncias é feito relativamente ao local onde se encontram depois do movimento em vez de serem feitas relativamente ao local de detecção original. O enviesamento da densidade populacional poderá ser grande caso o observador registe várias vezes o mesmo animal que se movimentou sem que o observador se aperceba. No caso do deslocamento se dever à presença do observador, a



densidade populacional poderá ser negativamente enviesada, se o animal se afastar do observador e positivamente enviesada caso contrário.

3 – Não existem erros de medição das distâncias.

É importante que as medições sejam realizadas com rigor, especialmente perto ou sobre a linha. Se os erros de medição forem pequenos e aleatórios não se verificam problemas na estimação da densidade, principalmente se a amostra for grande, mas erros sistemáticos de medição poderão provocar sérios problemas na estimação.

4 – Os objectos são independentes e estão aleatoriamente distribuídos.

A teoria da máxima verosimilhança utilizada na obtenção dos estimadores dos parâmetros e da variância amostral teórica assume independência entre detecções. Assim, se este pressuposto falhar, apesar de provocar pouco efeito nos estimadores dos parâmetros provoca enviesamento no estimador da variância teórica.

4.3 Estimadores de D , N e $var(\hat{D})$

A densidade populacional por unidade de área, D , é dada por $D = \frac{N}{A}$, onde N representa o número total de indivíduos que existem na área de dimensão A . Como ao percorrer a linha recta apenas n indivíduos são detectados na área coberta ($a = 2wL$), um estimador da densidade populacional é dado por

$$\hat{D} = \frac{n}{2wL\hat{P}_a} \quad , \quad (4.1)$$

onde P_a é a probabilidade de um objecto aleatoriamente escolhido ser detectado na área coberta e é estimada através de uma função de detecção, $g(x)$ correspondendo à probabilidade condicionada de detectar um objecto sabendo que está a uma certa distância x da linha, com $0 \leq x \leq w$ e assumindo $g(0) = 1$.

Seja $\mu = \int_0^w g(x)dx$ então $P_a = \frac{\mu}{w}$. Como a função g também não é conhecida, é necessário estimar μ e assim

$$\hat{P}_a = \frac{\hat{\mu}}{w} \quad . \quad (4.2)$$

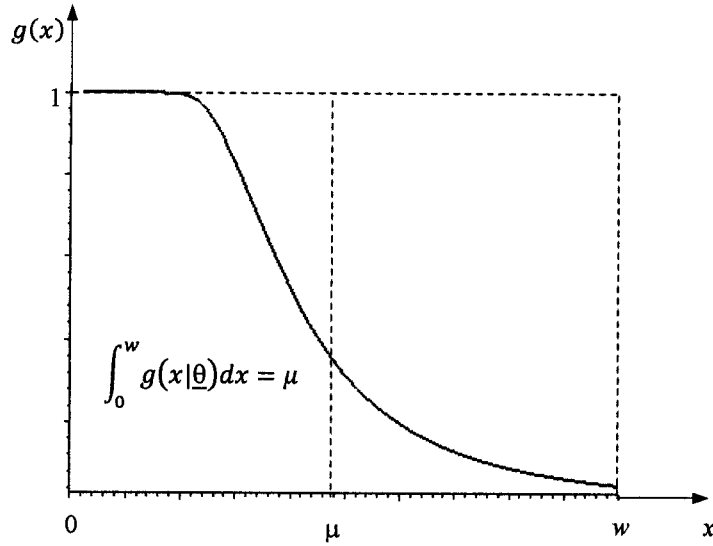


Figura 4.3.1. Probabilidade de detectar um objecto dentro da área examinada, a qual corresponde à área μ abaixo da função de detecção $g(x)$ quando expressa como a proporção da área w do rectângulo (adaptado de Buckland *et al.* 2001).

O integral da função g é definido como

$$\int g(x|\underline{\theta}) dx = \begin{cases} \mu, & \text{se } 0 \leq x \leq w \\ 0, & \text{caso contrário} \end{cases}.$$

A função de detecção $g(x)$ não é uma função densidade de probabilidade, mas para o ser basta apenas dividi-la por μ . Assim,

$$f(x|\underline{\theta}) = \frac{g(x|\underline{\theta})}{\int_0^w g(x|\underline{\theta}) dx} = \frac{g(x|\underline{\theta})}{\mu}.$$

A função $f(x|\underline{\theta})$ é uma função densidade de probabilidade (fdp), pois é sempre positiva ($g(x|\underline{\theta}) \geq 0 \forall x$) e o seu integral é igual a 1, como se pode verificar abaixo:

$$\int_0^\infty f(x|\underline{\theta}) dx = \int_0^\infty \frac{g(x|\underline{\theta})}{\int_0^w g(x|\underline{\theta}) dx} dx = \int_0^\infty \frac{g(x|\underline{\theta})}{\mu} dx = \frac{1}{\mu} \int_0^\infty g(x|\underline{\theta}) dx = \frac{1}{\mu} \cdot \mu = 1.$$

Como um dos pressupostos da metodologia é $g(0) = 1$ tem-se $f(0) = \frac{g(0)}{\mu}$. Então

$$\mu = \frac{1}{f(0)}.$$

Portanto, um estimador de $f(0)$ é dado por:

$$\hat{f}(0) = \frac{1}{\hat{\mu}}.$$

O inverso do valor da função densidade de probabilidade no ponto de abscissa zero é o estimador de μ , ou seja,

$$\hat{\mu} = \frac{1}{\hat{f}(0)}. \quad (4.3)$$

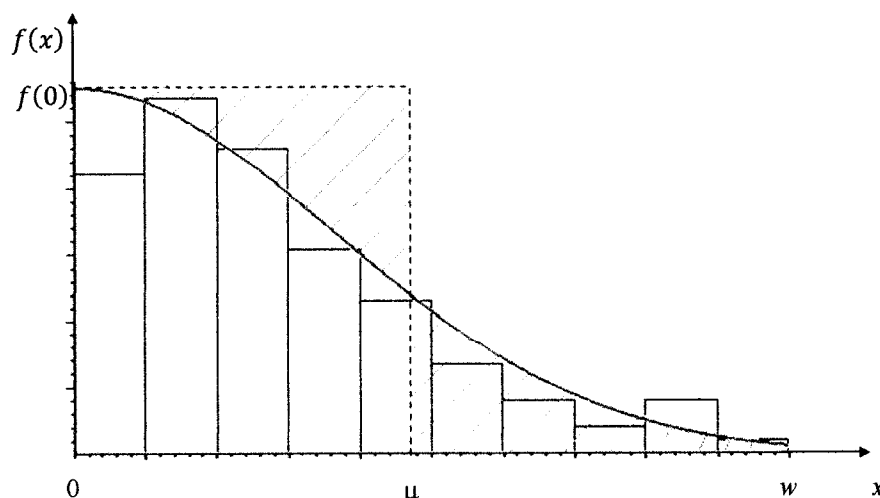


Figura 4.3.2. Função densidade de probabilidade, $f(x)$, sobre o histograma das frequências das distâncias perpendiculares. Como $f(x)$ é uma fdp a área abaixo da curva é unitária. As áreas tracejadas são iguais logo a área do retângulo também é igual a um. Portanto $\mu = \frac{1}{f(0)}$ (adaptado de Buckland *et al.*, 2001).

Substituindo (4.2) e (4.3) em (4.1), obtém o estimador da densidade populacional

$$\hat{D} = \frac{n}{2wL \frac{\hat{\mu}}{w}} = \frac{n}{2L \frac{1}{\hat{f}(0)}} = \frac{n\hat{f}(0)}{2L}. \quad (4.4)$$

A abundância é igual ao produto entre a densidade populacional e a área da região em que se encontra a população em estudo. Portanto, assumindo que a área em estudo A corresponde na totalidade à área coberta a , um estimador da abundância é $\hat{N} = \hat{D} \cdot a$, que pode ser escrito como

$$\hat{N} = \frac{n\hat{f}(0)}{2L} \cdot 2wL = nw\hat{f}(0)$$

A variância do estimador da densidade populacional definida em (4.4) provém de duas fontes de variabilidade: o tamanho da amostra e o valor da estimativa da função de detecção na distância zero.

Supondo que n e $\hat{f}(0)$ são independentes e que $\hat{f}(0)$ é não enviesado, um estimador da variância de \hat{D} é dado por

$$\widehat{var}(\widehat{D}) = \widehat{D}^2 \left\{ \frac{\widehat{var}(n)}{n^2} + \frac{\widehat{var}[\hat{f}(0)]}{[\hat{f}(0)]^2} \right\} \quad (4.5)$$

No apêndice A, secção A.2 mostra-se uma dedução da expressão (4.5).

Segundo Buckland *et al.* (2001), \widehat{D} é positivamente enviesada e obtém-se um intervalo de confiança com melhor cobertura assumindo que tem distribuição log-normal. Assim um intervalo de confiança de $100(1 - 2\alpha)\%$ para \widehat{D} , é dado por:

$$\left(\frac{\widehat{D}}{C}, \widehat{D} \cdot C \right)$$

onde $C = \exp \left[z_{\alpha} \sqrt{\widehat{var}(\ln \widehat{D})} \right]$ e $\widehat{var}(\ln \widehat{D}) = \ln \left[1 + \frac{\widehat{var}(\widehat{D})}{\widehat{D}^2} \right]$.

5. APLICAÇÃO DA PONDERAÇÃO DE MODELOS EM TRANSECTOS LINEARES

Neste capítulo pretende-se fazer uma aplicação da ponderação de modelos, usando alguns dados da experiência realizada por Laake (1978), com estacas de madeira.

Laake realizou um estudo controlado com estacas de madeira para avaliar o desempenho do método dos transectos lineares. Numa determinada área distribuíram-se aleatoriamente 150 estacas, obtendo-se uma densidade de 37.5 estacas/ha. O transecto linear, com 1000 metros de comprimento e $w = 20$, foi percorrido, em momentos diferentes, por 11 observadores. Cada observador registou as distâncias, perpendiculares ao transecto, a que detectou as estacas. Sendo este um estudo em que a densidade populacional é conhecida permite-nos fazer comparações e chegar a determinadas conclusões que de outra forma não se conseguiriam obter.

Na aplicação da ponderação de modelos serão utilizadas apenas as distâncias reais, correspondentes às registadas pelo observador 4. Os resultados obtidos através da ponderação de modelos serão comparados com os alcançados com a selecção de modelos tradicional em que se utiliza o AIC como critério para seleccionar o modelo que melhor se ajusta aos dados. Na realização da análise será utilizado o programa DISTANCE. Serão ajustados aos dados da amostra todos os modelos que combinam a função chave com a expansão em série e fixado em três o número máximo de parâmetros de cada modelo. Como é desaconselhado o uso do modelo exponencial negativo, de forma a verificar o efeito causado por este modelo na ponderação de modelos, estudaram-se dois cenários: um em que o conjunto de modelos é constituído pelas quatro funções chave combinadas com as expansões em série, e outro em que o modelo exponencial negativo não é incluído no conjunto de modelos considerados.

Ao longo deste capítulo, a expressão “melhor modelo” refere-se, como já foi dito no capítulo 2, ao modelo que apresenta o menor valor da informação K-L estimado, ou seja, o modelo com menor AIC.

5.1 Análise dos resultados

A primeira análise efectuada reporta-se ao cenário em que se ajustaram aos dados das estacas todas as funções chave combinadas com as expansões em série disponíveis no software DISTANCE. Para cada função chave seleccionou-se o modelo (função chave + expansão em série) que apresentava menor AIC. O modelo exponencial negativo, de entre os quatro considerados, foi o seleccionado, como se mostra no quadro 5.1.1. No entanto, é possível verificar, no quadro 5.1.2, que o modelo exponencial negativo sobrestima o valor dos parâmetros D e N . Os restantes modelos subestimam os parâmetros, sendo o modelo *half-normal* com 1 co-seno o que subestima ligeiramente D e N . A figura B.1 (apêndice B) permite verificar visualmente que qualquer um dos modelos apresenta um ajustamento bastante razoável aos dados, o qual pode ser comprovado pelos valores do valor p associado a cada modelo, que se encontram no quadro 5.1.1

Quadro 5.1.1. Modelo seleccionado para cada função chave.

Função chave	Expansão em série	Nº total de parâmetros	AIC	Δ_i ^b	valor p ^c
Exponencial negativa ^a	---	1	330.29	0.00	0.94
<i>Half-normal</i>	Co-seno	2	331.04	0.76	0.93
<i>Hazard-rate</i> ^a	---	2	331.10	0.81	0.95
Uniforme	Co-seno	2	331.42	1.13	0.96

^a Obteve-se um bom ajustamento sem ser necessário recorrer às expansões em série.

^b Diferenças de Akaike $\Delta_i = AIC_i - AIC_{min}$, $i = 1, \dots, 4$.

^c Teste para a bondade do ajustamento.

Quadro 5.1.2. Estimativas de D e N para cada modelo.

Modelo	\hat{D}	% Enviesamento (\hat{D})	\hat{N}	% Enviesamento (\hat{N})	% CV ^a
Exponencial negativo	43.63	16.35	175	16.67	20.79
<i>Half-normal</i> + 1 co-seno	37.00	-1.31	148	-1.33	19.90
<i>Hazard-rate</i>	33.37	-11.00	133	-11.33	20.99
Uniforme + 2 co-senos	33.54	-10.56	134	-10.67	17.44

^a Coeficiente de variação

Se não se considerar o modelo exponencial negativo, como é aconselhado nos dados com este tipo de histograma, o modelo com menor valor de AIC é *half-normal* com 1 co-seno. Porém este modelo é o que oferece um valor p mais pequeno e assim os outros dois modelos considerados ajustam-se ligeiramente melhor aos dados.

Os quatro modelos considerados apresentam valores para o AIC muito próximos uns dos outros. Esta proximidade dos valores do AIC poderá tornar difícil a escolha de um modelo em detrimento de outro, pois nenhum se destaca claramente como sendo o melhor. Curiosamente o modelo que apresenta o maior valor de AIC, é o que apresenta um melhor ajustamento aos dados.

Como as diferenças de Akaike são todas inferiores a 2, segundo Burnham e Anderson (2004), existe uma forte evidência de que os modelos, para os quais isso se verifica, são bons candidatos a melhor modelo. Observando os pesos de Akaike, indicados no quadro 5.1.3 verifica-se que em nenhum caso se tem $\omega_i \geq 0.9$, $i = 1, \dots, 4$ (que representa uma forte evidência de que o modelo é o melhor modelo do conjunto) e portanto nenhum modelo se destaca claramente relativamente aos outros. Não havendo nenhum modelo considerado o único melhor modelo, a inferência pode ser realizada de forma não condicional a um único modelo mas sim ao conjunto de modelos candidatos.

Assim, aplicou-se a ponderação de modelos ao mesmo conjunto de dados das estacas, primeiro usando os quatro modelos e depois excluindo o modelo exponencial negativo do conjunto de modelos possíveis.

Quadro 5.1.3. Modelos analisados e respectivos AIC, Δ_i e ω_i (inclui o modelo EN).

Modelo	AIC	Δ_i	ω_i^a
Exponencial negativa	330.29	0.00	0.34
<i>Half-normal</i> + 1 co-seno	331.04	0.76	0.24
<i>Hazard-rate</i>	331.10	0.81	0.23
Uniforme + 2 co-senos	331.42	1.13	0.20

$$^a \text{Pesos de Akaike } \omega_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^R \exp\left(-\frac{1}{2}\Delta_r\right)}, \quad R = 4.$$

Ao excluir o modelo exponencial negativo do conjunto de modelos candidatos, é necessário recalcular os pesos de Akaike, uma vez que estes devem somar 1. O quadro 5.1.4 mostra os novos pesos de cada modelo, notando-se também nesta situação a

inexistência de um modelo que se destaque nitidamente dos restantes, ou seja, $\omega_i \geq 0.9$ para algum $i = 1, \dots, 3$.

Quadro 5.1.4. Modelos analisados e respectivos AIC, Δ_i e ω_i (não inclui o modelo EN).

Modelo	AIC	Δ_i	ω_i ^a
<i>Half-normal</i> + 1 co-seno	331.04	0.00	0.36
<i>Hazard-rate</i>	331.10	0.06	0.35
Uniforme + 2 co-senos	331.42	0.38	0.30

^a Pesos de Akaike $\omega_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_{r=1}^R \exp(-\frac{1}{2}\Delta_r)}$, $R = 3$.

As estimativas dos parâmetros usando a ponderação de modelos, comportam-se de forma diferente, quando o modelo exponencial negativo está contido no conjunto de modelos a ponderar e quando não está (quadro 5.1.5). Se o modelo exponencial negativo fizer parte do conjunto de modelos candidatos, obtêm-se estimativas mais próximas dos valores reais dos parâmetros com a ponderação de modelos do que com a selecção, por meio do AIC, de um único modelo. No entanto, observa-se uma situação contrária, quando o modelo exponencial negativo não está contido no conjunto dos candidatos a melhor modelo.

Quanto ao enviesamento verifica-se que este é quase negligenciável no cenário A da ponderação de modelos e no cenário B da selecção de modelos tradicional. Observa-se ainda que a ponderação de modelos realizado com o conjunto de modelos onde está incluído o modelo exponencial negativo fornece estimativas menos enviesadas do que qualquer um dos outros cenários apresentados no quadro 5.1.5.

Outro facto que se observa, quer na selecção de modelos tradicional, quer na ponderação de modelos, é que em ambas as metodologias o coeficiente de variação diminui, quando se muda do cenário A para o cenário B. Esta situação não é de estranhar uma vez que o modelo exponencial negativo normalmente produz estimativas menos precisas. Também se verifica que o coeficiente de variação é maior na ponderação de modelos do que na selecção de modelos tradicional.

Parece que o modelo exponencial negativo, apesar de não ser muito alvitado o seu uso na selecção de modelos tradicional, poderá ter um papel relevante quando se utiliza a ponderação de modelos. Uma razão para tal situação poderá ser o facto de ao

sobrestimar o valor dos parâmetros, na ponderação de modelos o modelo exponencial negativo atenua o efeito da subestimação dos parâmetros provocada pelos restantes modelos conseguindo-se desta forma melhores estimativas.

Quadro 5.1.5. *Estimativas dos parâmetros para a selecção de um único modelo e para a ponderação de modelos, considerando dois cenários.*

	Seleccção de um único modelo		Ponderação de modelos	
	Cenário A ^a (EN)	Cenário B ^b (HN)	Cenário A ^a (com EN)	Cenário B ^b (sem EN)
\hat{D}	43.63	37.01	37.77	34.72
% Enviesamento (\hat{D})	16.36	-1.31	0.73	-7.41
% CV (\hat{D})	20.79	19.90	23.11	20.15
\hat{N}	175	148	151	139
% Enviesamento (\hat{N})	16.67	-1.33	0.73	-7.57
% CV (\hat{N})	20.79	19.90	23.22	20.18

^a O cenário A corresponde à análise usando os quatro modelos.

^b O cenário B corresponde à análise usando os três modelos (exclui-se o modelo exponencial negativo da análise).

Nota: Na selecção de um único modelo, no cenário A foi seleccionado o modelo exponencial negativo e no cenário B foi seleccionado o modelo half-normal com 1 co-seno.

Este exemplo das estacas indicia que a inclusão do modelo exponencial negativo no conjunto de modelos candidatos a melhor modelo pode ser importante para a obtenção de estimativas menos enviesadas aquando da realização da ponderação de modelos. Será que as estimativas dadas pela ponderação de modelos são sempre melhoradas (em termos de diminuir o enviesamento) quando se inclui o modelo exponencial negativo no conjunto de modelos candidatos? No próximo capítulo tentar-se-á dar resposta a esta questão.

6. RESULTADOS DAS SIMULAÇÕES

Estudos já realizados por Stefanski *et al.* (manuscrito em revisão) apontam no sentido de um enviesamento provocado pela selecção de modelos, usando o critério AIC, realizada no programa DISTANCE. Pretende-se neste capítulo aprofundar a conclusão a que chegaram aqueles investigadores, verificando se para além do enviesamento do modelo também se verifica um enviesamento das estimativas obtidas dos parâmetros de interesse.

Um outro objectivo deste capítulo é confirmar se a ponderação de modelos fornece estimativas dos parâmetros de interesse menos enviesadas do que as obtidas através de um único modelo seleccionado. Em caso afirmativo, pretende-se identificar o tipo de cenário em que ocorre o melhor desempenho da ponderação de modelos.

De forma a atingir os objectivos acima referidos realizam-se simulações Monte Carlo, no software R versão 2.7.0, de amostras de duas funções densidade de probabilidade: *half-normal* e *hazard-rate*. A escolha da primeira prende-se com o facto de no capítulo anterior se ter verificado que o modelo *half-normal* era um modelo que fornecia estimativas dos parâmetros em estudo muito próximas dos valores reais. A segunda função a simular foi escolhida por ser uma função bastante flexível e, portanto fácil de se ajustar a vários tipos de dados.

Para ambas as funções densidade de probabilidade foram simuladas 1000 amostras. A execução da análise automatizada destas amostras foi realizada com recurso ao MS-DOS que permitiu correr o executável MCDS.exe do DISTANCE versão 5.0 release 2. Algumas amostras das duas simulações foram excluídas por apresentarem erros de compilação ou em que o DISTANCE realizava a análise mas os resultados não eram aceitáveis. É de salientar que as amostras utilizadas em todos os cenários de cada uma das simulações foram as mesmas, com excepção daquelas que foi necessário retirar, quando se truncaram os dados, devido aos erros (no caso da simulação da função *half-normal* foram retiradas 9 amostras e no caso da simulação da função *hazard-rate* foram retiradas 11 amostras). Para mais detalhes sobre a implementação das simulações veja-se o apêndice C.

Na estimação do parâmetro de escala das funções densidade de probabilidade *half-normal* e *hazard-rate*, consideraram-se os valores reais $N = 150$, $L = 1000 m$, $w = 20 m$ e $D = 37.5$ estacas/ha, da experiência de Laake (1978), já mencionada no capítulo cinco. No caso da função densidade de probabilidade *hazard-rate* o parâmetro de forma foi fixado em 3. As amostras simuladas são de dimensão $n = 70$, um valor compreendido entre 60-80, aconselhado por Buckland *et al.* (2001).

A cada amostra ajustaram-se as quatro funções chave combinadas com as três expansões em série disponíveis no software. O modelo ajustado aos dados poderá ter sido formado apenas pela função chave, ou ser uma combinação da função chave com uma expansão em série, no caso da função chave não ter sido suficiente para obter um bom ajustamento. Estabeleceu-se que o número máximo de parâmetros (número de parâmetros da função chave + número de ajustamentos) de cada modelo seria igual a 3, como é recomendado em Buckland *et al.* (2001).

A selecção do melhor modelo para cada amostra processou-se da seguinte forma:

- para cada função chave (combinada ou não com uma expansão em série) escolheu-se o modelo que apresentou o menor AIC;
- posteriormente, de entre os quatro modelos, referentes a cada uma das funções chave (combinadas ou não com uma expansão em série), seleccionou-se o que exibiu um menor valor de AIC.

Nas secções seguintes serão apresentados os resultados da análise das amostras simuladas, considerando quatro cenários distintos:

- ajustamento das quatro funções chaves combinadas com as expansões em série e considerando os dados não truncados;
- ajustamento excluindo o modelo exponencial negativo da análise e considerando os dados não truncados;
- ajustamento das quatro funções chaves combinadas com as expansões em série e considerando os dados truncados no percentil 95;
- ajustamento excluindo o modelo exponencial negativo da análise e considerando os dados truncados no percentil 95;

Duas razões levaram à decisão de incluir o modelo exponencial negativo neste estudo, a saber, o facto de na literatura este modelo ser por vezes seleccionado (quando

é considerado) e a necessidade de verificar qual a influência da sua utilização na estimação do parâmetro $f(0)$.

Sendo $f(0)$ o parâmetro de interesse para estimar para D e N serão apenas apresentados resultados para este parâmetro.

Para os quatro cenários atrás mencionados serão comparados os resultados obtidos por meio da selecção de um único modelo baseado no mínimo valor do AIC, com os resultados alcançados com a ponderação de modelos.

6.1 Resultados obtidos na simulação da *half-normal*

6.1.1 Enviesamento do AIC e das estimativas

O primeiro aspecto verificado prende-se com o facto do programa DISTANCE nem sempre seleccionar o modelo verdadeiro, tal como tinha sido observado em Stefanski *et al.* (manuscrito em revisão).

Os resultados das simulações mostram que a selecção do modelo *half-normal* fica muito aquém do esperado, em qualquer dos cenários estudados, como se pode confirmar no quadro 6.1.1.1. O melhor cenário, aquele em que o modelo em questão foi mais vezes seleccionado, verifica-se para dados não truncados e quando se exclui o modelo exponencial negativo da análise. Neste cenário, de entre os três modelos possíveis, o *half-normal* foi seleccionado 44.2% das vezes, nas 1000 amostras estudadas.

Verifica-se também que o facto de truncar os dados piorou significativamente os resultados no que diz respeito à selecção do modelo verdadeiro. Este resultado talvez seja de esperar pois os dados foram gerados de uma distribuição não truncada, mas em que a probabilidade de se obter um valor superior a 20 é quase zero.

Quadro 6.1.1.1 *Seleccção do modelo half-normal nos quatro cenários estudados.*

	Com exponencial	Sem exponencial
Sem truncar	35.5%	44.2%
Truncar	11.1%	15.4%

Comparando os resultados das simulações para os quatro modelos seleccionados nota-se que o modelo uniforme é o mais seleccionado em três dos cenários investigados, como se pode observar na figura 6.1.1.1, a qual representa a distribuição da selecção dos vários modelos para os quatro cenários analisados. A excepção é observada no caso dos dados não truncados e quando o modelo exponencial negativo é excluído da análise.

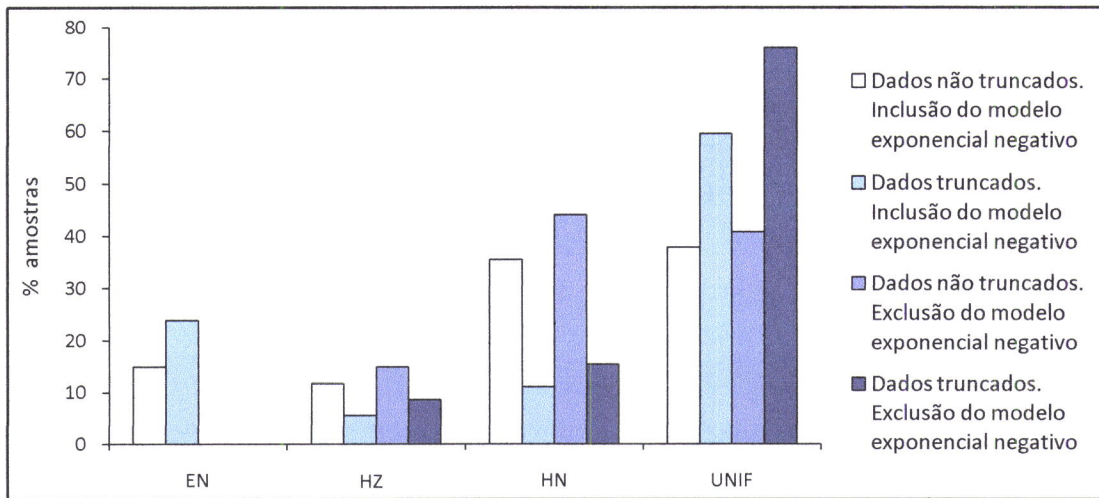


Figura 6.1.1.1 Percentagem de amostras em que cada modelo foi seleccionado para cada um dos quatro cenários (função simulada: *half-normal*).

Buckland *et al.* (2001) afirmam que os dados devem ser truncados de forma a eliminar possíveis *outliers* que tornam difícil o ajustamento do modelo aos dados e não aconselham a inclusão do modelo exponencial negativo na análise a não ser pelas razões já enunciadas no capítulo 5 desta dissertação. Portanto, é bastante curioso, que perante aquele cenário, em mais de 75% das amostras, o modelo mais seleccionado não seja o *half-normal* mas sim o uniforme. Uma possível explicação para este aspecto é que ao truncar elimina-se parte da cauda favorecendo-se o aparecimento do modelo uniforme com um pequeno decréscimo modelado pelo co-seno.

Ao analisar mais detalhadamente as combinações função chave + expansão em série nota-se efectivamente uma tendência para escolher modelos com menos parâmetros em vez daqueles que têm mais parâmetros. O quadro 6.1.1.2 mostra um resumo dos modelos mais seleccionados notando-se uma superioridade na escolha do modelo uniforme com um co-seno. No apêndice B, quadro B.1, é apresentada uma versão completa do quadro 6.1.1.2.

Quadro 6.1.1.2. *Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série (função simulada: half-normal).*

Função chave	Expansão em série ^a	CENÁRIOS			
		Sem truncar e com EN	Sem truncar e sem EN	Truncar e com EN	Truncar e sem EN
<i>Half-normal</i>	---	34.90%	41.80%	10.09%	13.02%
Uniforme	Polinómio Simples (1)	14.20%	14.40%	4.74%	4.84%
	Polinómio Hermite (1)	0.00%	0.00%	3.94%	4.14%
	Co-seno (1)	20.90%	22.20%	45.41%	60.14%
<i>Hazard-rate</i>	---	10.80%	12.30%	5.55%	8.48%
Exponencial negativa	---	12.10%	---	23.82%	---

^a Os números que se encontram entre parêntesis em frente das expansões em série referem-se ao número de ajustamentos efectuados.

Da análise dos quadros 6.1.1.1 e 6.1.1.2 e figura 6.1.1.1 conclui-se que existe um enviesamento na selecção de modelos feita usando o critério AIC, realizada no software DISTANCE. O enviesamento observado é maior quando se utilizam os dados truncados, notando-se que entre 84.6% e 88.9% das amostras o modelo seleccionado não é o verdadeiro, aquele que gerou os dados. No entanto, quando os dados não estão truncados, a percentagem de vezes que outro modelo é seleccionado é menor face ao caso dos dados truncados mas mesmo assim é uma percentagem elevada (entre 55.8% e 64.5%). Há uma tendência para escolher o modelo uniforme combinado com um co-seno, em vez do modelo *half-normal*, possivelmente por serem modelos com formas muito semelhantes e oferecerem ambos um bom ajustamento aos dados.

Além do enviesamento na escolha do modelo nota-se também um enviesamento nas estimativas de $f(0)$, ou seja, por vezes, o modelo que apresenta menor valor de AIC, e que é o seleccionado, produz estimativas ligeiramente mais afastadas do valor verdadeiro do que aqueles que não são seleccionados. Esta situação verifica-se em mais de metade das amostras e para todos os cenários analisados, como é possível verificar no quadro 6.1.1.3.

Quadro 6.1.1.3. *Percentagem de amostras em que o modelo seleccionado fornece estimativas de $f(0)$ mais enviesadas face a outros modelos do conjunto (função simulada: half-normal).*

Cenários		Estimativa mais enviesada ^a
Dados não truncados	Inclusão do modelo exponencial negativo	79.0%
	Exclusão do modelo exponencial negativo	66.9%
Dados truncados	Inclusão do modelo exponencial negativo	81.3%
	Exclusão do modelo exponencial negativo	72.1%

^a A estimativa de $f(0)$ fornecida pelo modelo seleccionado tem um enviesamento relativo maior do que as fornecidas pelos modelos não seleccionados.

É importante saber qual é a percentagem de enviesamento das estimativas fornecidas pelos modelos seleccionados, uma vez que o modelo seleccionado nem sempre apresenta a estimativa de $f(0)$ menos enviesada. Os histogramas apresentados nas figuras 6.1.1.2 a – d) mostram a distribuição, em percentagem, dos enviesamentos de $\hat{f}(0)$ correspondentes às amostras consideradas no quadro 6.1.1.3. Analisando aquelas, nota-se que entre 65% e 80% das amostras, os modelos seleccionados apresentam estimativas de $f(0)$ com um enviesamento entre -20% e 20%, nos cenários estudados. Observa-se também que nos cenários com os dados não truncados, entre 57% e 61% das amostras, o modelo seleccionado fornece estimativas com enviesamento negativo, ou seja, as estimativas são geralmente subestimadas. Quando os dados são truncados verifica-se uma situação contrária, as estimativas fornecidas pelos modelos são sobrestimadas, entre 71% e 75% das amostras. Uma justificação provável para esta situação é o facto de ao truncar o valor de $\hat{\mu}$ ser menor do que é efectivamente tornando $\hat{f}(0)$ maior.

Convém salientar que um número considerável de amostras (entre 7% e 20%) apresenta uma percentagem de enviesamento para $\hat{f}(0)$ superior a 30% podendo-se considerar um enviesamento demasiado severo mesmo que existam já alguns valores de referência, para eventual calibração

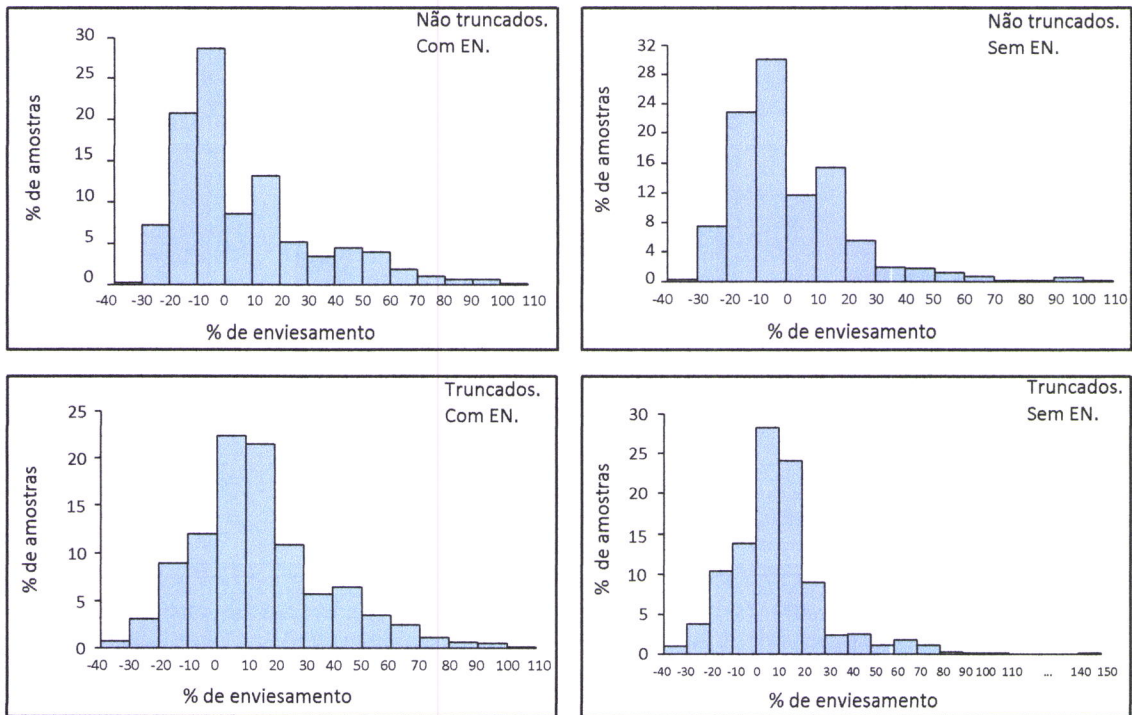


Figura 6.1.1.2 Distribuição do enviesamento de $\hat{f}(0)$ para as amostras em que o modelo seleccionado produz estimativas mais enviesadas que os restantes (função simulada: *half-normal*).

6.1.2 Ponderação de modelos versus selecção de modelos tradicional

Sendo um dos objectivos do capítulo inferir acerca das estimativas de $f(0)$ fornecidas pela ponderação de modelos determinou-se o enviesamento relativo de $f(0)$ para esta abordagem e para a selecção de modelos tradicional. Compararam-se os resultados obtidos com as duas abordagens e calculou-se a frequência relativa, em percentagem, do número de amostras em que a ponderação de modelos fornece estimativas de $f(0)$ com enviesamento relativo menor do que a selecção de modelos tradicional. A figura 6.1.2.1 evidência uma melhoria das estimativas de $f(0)$ fornecidas pela ponderação de modelos, por apresentarem menor enviesamento relativo, em qualquer um dos cenários. No entanto, essa melhoria é mais destacada no cenário em que não se truncam os dados e se inclui o modelo exponencial negativo no conjunto de modelos, onde em cerca de 62% das amostras o $\hat{f}(0)$ da ponderação de modelos é menos enviesado do que o da selecção de modelos tradicional. Nos outros cenários aquela percentagem situa-se ligeiramente abaixo dos 60%.

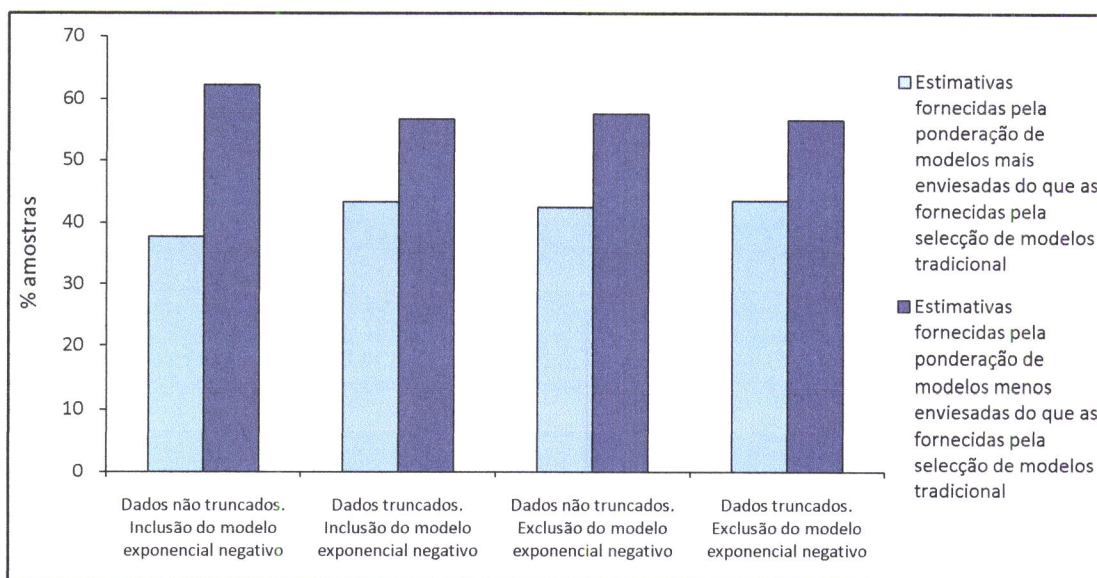


Figura 6.1.2.1 Comparação das estimativas de $f(0)$ dadas pela ponderação de modelos e pela selecção de modelos tradicional (função simulada: *half-normal*).

Ao analisar as distribuições dos enviesamentos das várias amostras, para os cenários estudados, nota-se que em média as estimativas de $f(0)$ obtidas quando não se trunca e o modelo exponencial negativo não está incluído no conjunto de modelos considerados são menos enviesadas, que nos restantes casos. Naquele cenário observa-se também uma menor variabilidade do enviesamento relativamente aos outros. Este aspecto é verificado quer para a ponderação de modelos, quer para a selecção de modelos tradicional. Observa-se também que o enviesamento médio é ligeiramente negativo, apenas para o cenário mencionado, e só no caso da ponderação de modelos obtendo-se naquela situação o enviesamento médio mais pequeno de todos os cenários examinados, como mostra o quadro 6.1.2.1.

Os histogramas referentes às distribuições das percentagens de enviesamento de $\hat{f}(0)$ podem ser consultados no apêndice B, figura B.2.

Quadro 6.1.2.1. Média e desvio-padrão da percentagem de enviesamento de $\hat{f}(0)$ obtida nos quatro cenários, para a selecção de modelos tradicional e para a ponderação de modelos (função simulada: half-normal).

Cenários	Média		Desvio-padrão	
	Seleção tradicional	Ponderação de modelos	Seleção tradicional	Ponderação de modelos
Dados não truncados. Inclusão do modelo exponencial negativo.	4.66	4.03	22.50	17.57
Dados não truncados. Exclusão do modelo exponencial negativo.	0.19	- 0.32	16.97	14.35
Dados truncados. Inclusão do modelo exponencial negativo.	12.44	11.45	21.74	17.48
Dados truncados. Exclusão do modelo exponencial negativo.	7.52	6.99	18.61	16.20

6.2 Resultados obtidos na simulação da *hazard-rate*

6.2.1 Enviesamento do AIC e das estimativas

Como se pode observar no quadro 6.2.1.1, apesar dos dados serem oriundos de uma função densidade de probabilidade *hazard-rate*, em mais de 60% das amostras, o modelo *hazard-rate* não foi o mais seleccionado, para os cenários dos dados não truncados. Nos cenários de dados truncados, a percentagem de amostras em que o modelo *hazard-rate* é seleccionado ainda é menor relativamente aos dados não truncados, atingindo percentagens inferiores a 16%.

Quadro 6.2.1.1. Selecção do modelo *hazard-rate* nos quatro cenários estudados

	Com exponencial	Sem exponencial
Sem truncar	33.70%	36.50%
Truncar	14.16%	15.37%

A comparação dos vários modelos seleccionados nos quatro cenários, permite concluir que o modelo uniforme é bastante mais seleccionado que os outros modelos no

caso dos dados truncados, mas perde importância relevante quando não se truncam os dados. Os modelos *hazard-rate* e *half-normal* são os mais escolhidos quando os dados não são truncados. Na figura 6.2.1.1 é possível ver a percentagem de selecção dos vários modelos nos cenários estudados.

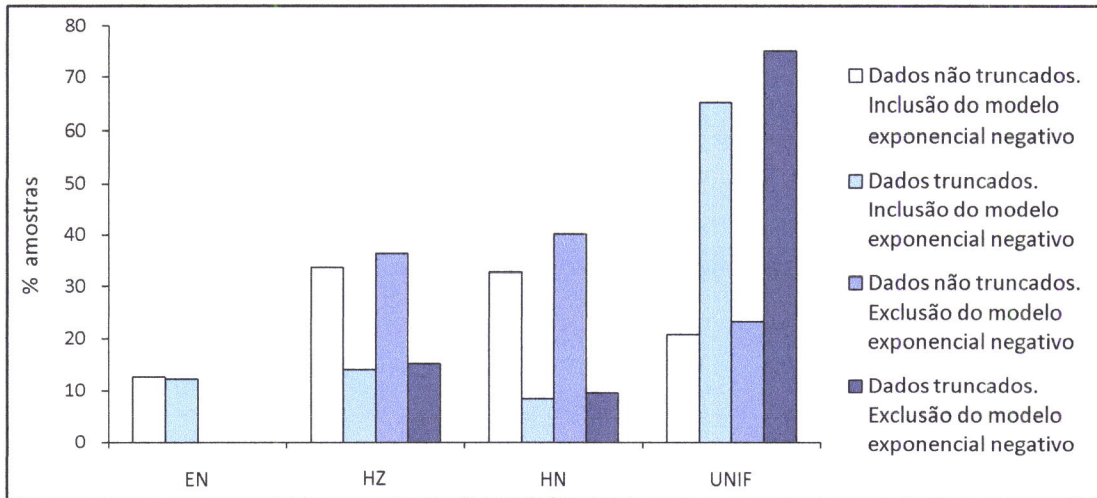


Figura 6.2.1.1 Percentagem de amostras em que cada modelo foi seleccionado para cada um dos quatro cenários (função simulada: *hazard-rate*).

Observando as combinações função chave + expansão em série, verifica-se, tal como se pode ver no quadro 6.2.1.2, que:

- no caso dos dados truncados, o modelo uniforme com um co-seno é escolhido em mais de 50% das amostras, quando se usa o AIC, como critério de selecção de modelos;

- no caso dos dados não truncados o modelo *hazard-rate* sem termos de ajustamento é o mais seleccionado, não chegando, no entanto, a percentagem de amostras aos 40%;

- o modelo que mais “beneficia” com a exclusão do modelo exponencial negativo é o *half-normal* sem termos de ajustamento, no caso dos dados não truncados (cuja selecção aumenta cerca de 5%) e o uniforme com 1 co-seno, no caso dos dados truncados (cuja selecção aumenta cerca de 8%);

A informação completa sobre a percentagem de amostras para as quais são seleccionados os vários modelos pode ser consultada no quadro B.2 (apêndice B).

É de salientar que nos cenários com dados não truncados o modelo de onde provêm os dados é o mais seleccionado mas a percentagem de amostras para as quais isso se verifica não é muito diferente da percentagem de amostras que selecciona o

modelo *half-normal* sem termos de ajustamento. Esta situação pode dever-se ao facto das duas funções oferecerem bons ajustamentos aos dados e o modelo *half-normal* ter menos parâmetros para estimar do que o modelo *hazard-rate* sendo portanto seleccionado quase tantas vezes como o *hazard-rate*. Nos cenários com dados truncados existe uma tendência clara para seleccionar o modelo uniforme com um co-seno. Uma possível razão para esta tendência pode ter a ver com o número de parâmetros dos dois modelos, uniforme com 1 co-seno (1 parâmetros) e *hazard-rate* (2 parâmetros). Se ambos tiverem valores correspondentes à verosimilhança semelhantes, o modelo *hazard-rate* por ter dois parâmetros será mais penalizado do que o modelo uniforme com 1 co-seno, que só tem um parâmetro e assim terá um valor de AIC menor o que determina a sua selecção (recorde-se que o AIC é calculado através da soma de duas parcelas: uma tem em conta o valor da verosimilhança e a outra o número de parâmetros estimados do modelo). Outra razão plausível é o facto de ao truncar se eliminar a cauda da distribuição obtendo-se um “ombro” maior relativamente ao novo w usado e assim assemelhar-se mais ao modelo uniforme modelado por um co-seno.

Quadro 6.2.1.2. *Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série (função simulada: hazard-rate).*

Função chave	Expansão em série ^a	CENÁRIOS			
		Sem truncar e com EN	Sem truncar e sem EN	Truncar e com EN	Truncar e sem EN
<i>Half-normal</i>	---	30.20%	35.10%	8.29%	9.20%
Uniforme	Polinómio Simples (1)	3.80%	3.80%	4.25%	4.55%
	Polinómio Hermite (1)	0.00%	0.00%	3.34%	3.44%
	Co-seno (1)	6.90%	7.20%	51.16%	59.45%
<i>Hazard-rate</i>	---	33.60%	36.00%	14.05%	15.17%
Exponencial negativa	---	11.60%	---	12.04%	---

^a Os números que se encontram entre parêntesis em frente das expansões em série referem-se ao número de ajustamentos efectuados.

Depois de se confirmar que a selecção de um único modelo através do AIC é frequentemente enviesada, observa-se também que o modelo escolhido em mais de 60%

das amostras apresenta estimativas mais enviesadas dos que os modelos não seleccionados, como é possível ver no quadro 6.2.1.3.

Quadro 6.2.1.3. *Percentagem de amostras em que o modelo seleccionado fornece estimativas mais enviesadas face a outros modelos do conjunto (função simulada: hazard-rate).*

Cenários		Estimativa mais enviesada ^a
Dados não truncados	Inclusão do modelo exponencial negativo	69.8%
	Exclusão do modelo exponencial negativo	64.4%
Dados truncados	Inclusão do modelo exponencial negativo	82.0%
	Exclusão do modelo exponencial negativo	76.7%

^a A estimativa de $f(0)$ fornecida pelo modelo seleccionado tem um enviesamento relativo maior do que as fornecidas pelos modelos não seleccionados.

As figuras 6.2.1.2 a) – d) mostram que o enviesamento de $\hat{f}(0)$ relativo aos modelos a que se referem o quadro 6.2.1.3 está compreendido entre -20% e 20%, numa percentagem de amostras que vai desde 54% a 78%. Contudo, observa-se que em qualquer dos cenários, o enviesamento é sobretudo positivo. De facto, cerca de 90% das amostras apresentam enviesamento positivo de $\hat{f}(0)$, no caso dos dados truncados, enquanto no caso dos dados não truncados aquela percentagem se situa entre 66% e 68%.

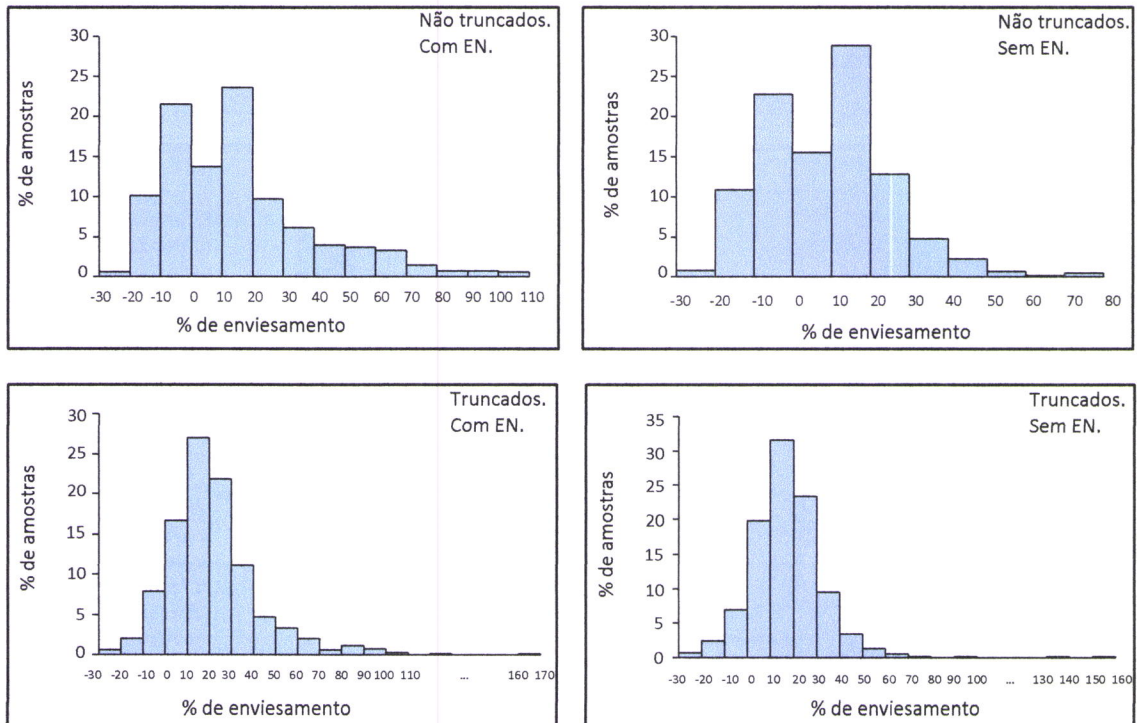


Figura 6.2.1.2 Distribuição do enviesamento de $\hat{f}(0)$ para as amostras em que o modelo seleccionado produz estimativas mais enviesadas que outros modelos do conjunto (função simulada:

6.2.2 Ponderação de modelos versus selecção de modelos tradicional

À semelhança do que foi feito com a simulação da função *half-normal*, também para a função *hazard-rate* se compararam os enviesamentos de $\hat{f}(0)$, produzidos quer pela ponderação de modelos, quer pela selecção de um único modelo. Os resultados são apresentados na figura 6.2.2.1, onde é apresentada a percentagem de amostras em que a ponderação de modelos apresenta estimativas de $f(0)$ menos afastadas do valor verdadeiro do que a selecção de modelos tradicional, bem como a percentagem de amostras em que se verifica o contrário.

Ao truncar os dados ocorrem duas situações: se o modelo exponencial negativo não estiver no conjunto de modelos considerados, a ponderação de modelos apresenta estimativas menos enviesadas que a selecção de modelos tradicional, em cerca de 60% das amostras, mas se o modelo exponencial for incluído no conjunto de modelos considerados, nota-se um equilíbrio entre a ponderação de modelos e a selecção de modelos tradicional. No caso dos dados não truncados, entre 53% e 57% das amostras,

as estimativas de $f(0)$ fornecidas através da ponderação de modelos são menos enviesadas do que as fornecidas por um único modelo seleccionado.

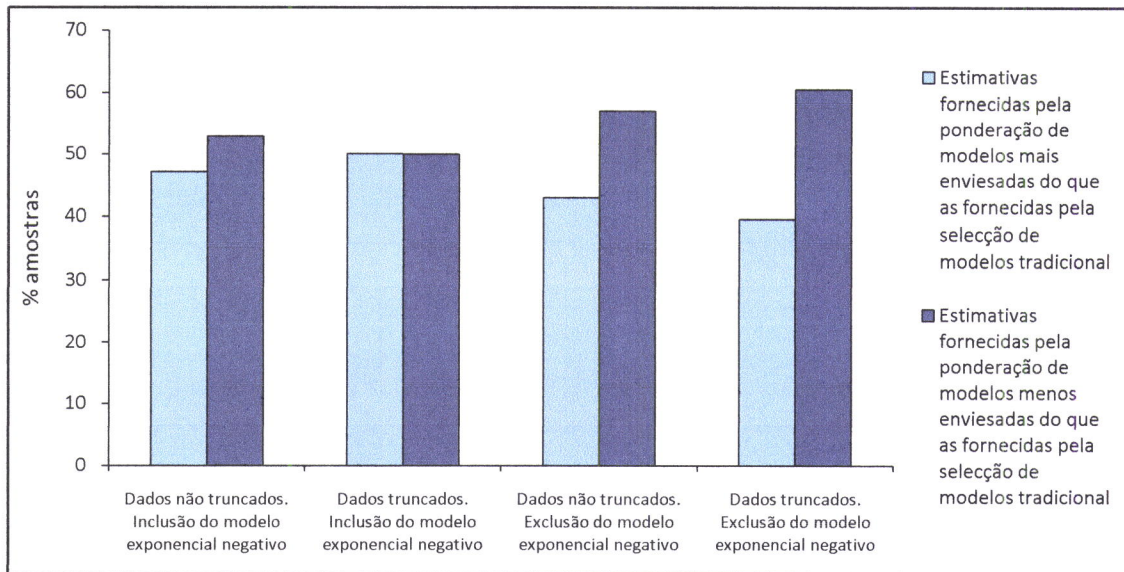


Figura 6.2.2.1 Comparação das estimativas de $f(0)$ dadas pela ponderação de modelos e pela selecção de modelos tradicional (função simulada: *hazard-rate*).

Depois de verificar que a estimativa de $f(0)$ é enviesada, quer se utilize a selecção de modelos tradicional, quer se utilize a ponderação de modelos, importa averiguar o que acontece com a distribuição das percentagens de enviesamento das estimativas de $f(0)$ das várias amostras simuladas. Idealmente o enviesamento médio de $\hat{f}(0)$ deveria ser aproximadamente zero, no entanto isso não se verifica, em nenhuma das abordagens, como se pode observar no quadro 6.2.2.1. Verifica-se ainda que as médias obtidas nas duas abordagens têm valores muito semelhantes. O cenário em que o enviesamento médio é menor, apesar de se situar ligeiramente acima de 7, é o correspondente aos dados não truncados e sem considerar o modelo exponencial negativo. É também para aquele cenário que se obtém um menor desvio-padrão do enviesamento de $\hat{f}(0)$, para a ponderação de modelos, pois no caso da selecção de modelos tradicional a menor variabilidade corresponde ao cenário de dados truncados e sem incluir o modelo exponencial negativo. Nota-se também que em qualquer cenário da ponderação de modelos se obtém um desvio-padrão do enviesamento de $\hat{f}(0)$ menor do que o conseguido com a selecção de modelos tradicional.

Os histogramas referentes às distribuições das percentagens de enviesamento de $\hat{f}(0)$ podem ser consultados no apêndice B, figura B.3.

Quadro 6.2.2.1. Média e desvio-padrão da percentagem de enviesamento de $\hat{f}(0)$ obtido nos quatro cenários, para a selecção de modelos tradicional e para a ponderação de modelos (função simulada: hazard-rate).

Cenários	Média		Desvio-padrão	
	Seleccção tradicional	Ponderação de modelos	Seleccção tradicional	Ponderação de modelos
Dados não truncados. Inclusão da exponencial negativa	11.76	11.75	21.35	16.03
Dados não truncados. Exclusão da exponencial negativa	7.05	7.19	18.87	12.78
Dados truncados. Inclusão da exponencial negativa	19.11	19.02	19.94	16.11
Dados truncados. Exclusão da exponencial negativa.	15.58	15.19	15.62	14.09

6.3 Comparação entre os resultados dos modelos *half-normal* e *hazard-rate*

Os resultados obtidos com as simulações de amostras oriundas das funções densidades de probabilidade *half-normal* e *hazard-rate* mostram que o AIC tende claramente (em mais de 75% das amostras) a seleccionar o modelo uniforme, quando se truncam os dados e não se inclui o modelo exponencial negativo no conjunto de modelos considerados. Em ambas as simulações, em cerca de 60% das amostras o modelo uniforme é combinado com um co-seno.

No que respeita à selecção do modelo verdadeiro, conclui-se que o cenário “mais favorável” a essa selecção, nas duas simulações, é aquele em que não se truncam os dados nem se incluiu o modelo exponencial negativo no conjunto de modelos. No entanto, apesar de ser o melhor cenário, o modelo *half-normal* só foi seleccionado 44% das vezes e o *hazard-rate* nem sequer chegou aos 37%, nas respectivas simulações.

Se o modelo exponencial negativo não estiver incluído no conjunto de modelos considerados, a percentagem que lhe correspondia quando era seleccionado é distribuída pelos outros modelos, verificando-se situações distintas quando se truncam e quando não se truncam os dados:

- nos cenários em que se efectua a truncatura, o modelo uniforme é o mais beneficiado com a ausência do modelo exponencial negativo. A selecção do modelo uniforme aumenta 15% no caso da simulação da *half-normal*, e 10% no caso da simulação da *hazard-rate*, face ao cenário que considera também o modelo exponencial negativo.

- nos cenários em que não é efectuada a truncatura, nota-se um aumento da selecção do modelo *half-normal*, em ambas as simulações, mas que não chega aos 9% comparado com o cenário onde se inclui o modelo exponencial negativo.

6.4 Comparação entre os resultados obtidos com o AIC e com o AIC_C

Já foi referido no capítulo 3, página 14, desta dissertação que o critério AIC_C deve ser preferido em relação ao critério AIC, no caso de amostra pequenas relativamente ao número de parâmetros estimados do modelo, ou quando o número de parâmetros é muito grande. Neste exemplo concreto, em que o tamanho da amostra é 66 ou 70 e no máximo o modelo tem três parâmetros é feita uma comparação entre os resultados obtidos pelos dois critérios de selecção de modelos.

Os critérios AIC e AIC_C diferem um do outro apenas no que diz respeito ao termo de penalização pois o termo relativo à verosimilhança é igual em ambos. Se o termo correspondente à penalização não for muito diferente para os dois critérios pouca influência tem na selecção do modelo.

A partir do quadro 6.4.1 é possível notar que para o tamanho de amostra e número de parâmetros considerados existe uma diferença mínima nos termos de penalização dos dois critérios e portanto a selecção de modelos pouco será afectada por essa diferença.

Quadro 6.4.1. Comparação entre os termos de penalização dos critérios AIC e AIC_C.

Número de parâmetros do modelo (k)	Termo de penalização no AIC ($2k$)	Termo de penalização no AIC _C $\left(\frac{2k(k+1)}{n-k-1}\right)$	
	$n = 66 ; n = 70$	$n = 66$	$n = 70$
0	0	0.00	0.00
1	2	2.06	2.06
2	4	4.19	4.18
3	6	6.40	6.36

No caso das simulações da função *half-normal*, quer se utilize o AIC ou o AIC_C, o modelo mais seleccionado não se altera em nenhum dos cenários, como se pode ver no quadro 6.4.2. Nos cenários onde não se truncam os dados, a diferença da percentagem de amostras seleccionadas para cada modelo é no máximo de 1.6% o que representa, em termos absolutos 16 amostras. Já no caso dos dados truncados essa diferença não chega a 1%, representando menos de 10 amostras. De facto, seria de esperar que fosse praticamente irrelevante usar AIC ou AIC_C nesta situação pois a função *half-normal* simulada só tem um parâmetro e em mais de 78% das amostras o modelo seleccionado, mesmo não sendo sempre o verdadeiro tem 0 ou 1 parâmetro (ver quadro B.1 do apêndice B), o que comparativamente com o tamanho das amostras é pequeno e assim não se justifica o uso do AIC_C.

Quadro 6.4.2. Percentagem de amostras para as quais cada modelo é seleccionado, usando o AIC e o AIC_C, em cada cenário analisado (função simulada: *half-normal*).

Cenários	Não truncar Com EN		Não truncar Sem EN		Truncar Com EN		Truncar Sem EN	
	AIC	AIC _C	AIC	AIC _C	AIC	AIC _C	AIC	AIC _C
EN	14.9	14.8	----	----	23.9	24.1	----	----
HZ	11.7	10.5	14.9	13.7	5.5	4.9	8.6	8.0
HN	35.5	36.8	44.2	45.6	11.1	11.1	15.4	15.4
UNIF	37.9	37.9	40.9	40.7	59.4	59.8	76.0	76.6

Relativamente à percentagem de enviesamento das estimativas de $f(0)$, quer se utilize a selecção de modelos tradicional, quer se utilize a ponderação de modelos, notam-se apenas pequenas diferenças ao usar um ou outro critério, como é possível ver no quadro 6.4.3. Observa-se também uma certa consistência na distribuição da

percentagem de enviesamento, ou seja, os intervalos $[-10, 0[$ e $[0, 10[$ são os que registam a maior percentagem de amostras, nos cenários com dados não truncados e com dados truncados, respectivamente, independentemente do critério ou metodologia utilizados.

Quadro 6.4.3 *Percentagem de amostras com enviesamento de $\hat{f}(0)$ entre -20% e 20%, para cada um dos cenários analisados, considerando a selecção de modelos tradicional e a ponderação de modelos (função simulada: half-normal).*

Cenário	Método	Critério	% de enviesamento				Total
			$[-20, -10[$	$[-10, 0[$	$[0, 10[$	$[10, 20[$	
Não truncar. Com EN.	Seleccção tradicional	AIC	16.6	32.3	16.5	11.4	76.8
		AIC _C	16.7	32.3	17.0	11.0	77.0
	Ponderação de modelos	AIC	16.4	31.8	20.1	12.9	81.2
		AIC _C	16.4	32.2	20.1	12.6	81.3
Não truncar. Sem EN.	Seleccção tradicional	AIC	17.6	35.8	19.7	12.0	85.1
		AIC _C	17.6	35.7	20.5	11.6	85.4
	Ponderação de modelos	AIC	20.3	37.3	20.3	9.9	87.8
		AIC _C	20.2	38.5	19.9	9.7	88.3
Truncar. Com EN.	Seleccção tradicional	AIC	7.4	17.4	26.2	19.4	70.4
		AIC _C	7.5	17.1	26.2	19.5	70.3
	Ponderação de modelos	AIC	7.1	17.4	27.1	20.8	72.4
		AIC _C	7.2	17.2	27.4	20.6	72.4
Truncar. Sem EN.	Seleccção tradicional	AIC	8.3	21.5	30.4	20.9	81.1
		AIC _C	8.5	21.3	30.5	21.1	81.4
	Ponderação de modelos	AIC	10.3	21.6	31.0	19.5	82.4
		AIC _C	10.0	21.8	31.4	19.1	82.3

No caso das simulações da função *hazard-rate* verificam-se algumas diferenças na selecção de modelos para o cenário de dados não truncados e inclusão do modelo

exponencial negativo no conjunto de modelos. Observa-se que ao usar o critério AIC a função mais vezes seleccionada é a *hazard-rate*, enquanto a função *half-normal* é a mais escolhida através do critério AIC_C . Esta situação é devida aos casos em que as duas funções apresentam valores de AIC muito próximos e ao alterar o critério para AIC_C , a penalização da função *hazard-rate* ser maior (pois tem 2 parâmetros) do que a da *half-normal*. Assim obtém um valor de AIC_C menor para a função *half-normal* do que para a função *hazard-rate*, tornando-a mais seleccionada. Nos restantes cenários, a função mais escolhida é a mesma, quer se use um ou outro critério, como se observa no quadro 6.4.4. Verifica-se ainda que o modelo verdadeiro (*hazard-rate*) é, em média, cerca de 2% menos escolhido usando o critério AIC_C do que usando o critério AIC.

Quadro 6.4.4. *Percentagem de amostras para as quais cada modelo é seleccionado, usando o AIC e o AIC_C , em cada cenário analisado (função simulada: hazard-rate).*

Cenários	Não truncar Com EN		Não truncar Sem EN		Truncar Com EN		Truncar Sem EN	
	AIC	AIC_C	AIC	AIC_C	AIC	AIC_C	AIC	AIC_C
EN	12.7	13.4	----	----	12.3	12.5	----	----
HZ	33.7	31.2	36.5	34.1	14.2	12.8	15.4	13.9
HN	32.8	36.1	40.2	43.8	8.4	8.7	9.6	9.9
UNIF	20.8	19.3	23.3	22.1	65.1	66.0	75.0	76.2

Quanto à percentagem de enviesamento de $\hat{f}(0)$, os resultados obtidos através da selecção de modelos feita com o AIC são semelhantes aos obtidos usando o critério AIC_C , mesmo no cenário em que o modelo mais seleccionado usando um ou outro critério difere. Esta semelhança nos resultados é obtida quer na selecção de modelos tradicional, quer na ponderação de modelos, como mostra o quadro 6.4.5. Através daquele quadro verifica-se também, como no caso da simulação da *half-normal*, que existe uma certa consistência na distribuição dos enviesamentos produzidos pelos dois métodos e para os dois critérios de selecção. Assim, observa-se que entre 25% e 33% de amostras, a percentagem de enviesamento de $\hat{f}(0)$ se situa no intervalo $[0, 10[$, quando os dados não são truncados e no intervalo $[10, 20[$, quando os dados são truncados.

Comparativamente com a simulação da *half-normal* nota-se para a simulação da *hazard-rate* um deslocamento positivo de cerca de 10% na distribuição dos enviesamentos de $\hat{f}(0)$, ou seja, há uma tendência para que as estimativas de $f(0)$ da

função *hazard-rate* sejam mais enviesadas positivamente do que as da função *half-normal*.

Quadro 6.4.5. *Percentagem de amostras com enviesamento de $\hat{f}(0)$ entre -20% e 20%, para cada um dos cenários analisados, considerando a selecção de modelos tradicional e a ponderação de modelos (função simulada: hazard-rate).*

Cenário	Método	Critério	% de enviesamento				Total
			[-20, -10[[-10, 0[[0, 10[[10, 20[
Não truncar. Com EN.	Seleccção tradicional	AIC	7.3	24.1	24.8	20.2	76.4
		AIC _C	7.4	23.6	24.7	19.8	75.5
	Ponderação de modelos	AIC	4.8	19.4	27.8	22.8	74.8
		AIC _C	4.8	19.6	27.1	23.1	74.6
Não truncar. Sem EN.	Seleccção tradicional	AIC	7.5	26.0	27.4	22.6	83.5
		AIC _C	7.6	25.5	27.8	22.5	83.4
	Ponderação de modelos	AIC	5.7	25.4	32.5	22.0	85.6
		AIC _C	6.0	25.4	32.5	22.0	85.9
Truncar. Com EN.	Seleccção tradicional	AIC	1.6	10.3	21.1	25.4	58.4
		AIC _C	1.3	10.3	21.2	25.4	58.2
	Ponderação de modelos	AIC	1.0	7.2	21.4	28.9	58.5
		AIC _C	0.9	6.9	21.3	28.9	58.0
Truncar. Sem EN.	Seleccção tradicional	AIC	1.8	10.5	24.0	28.5	64.8
		AIC _C	1.5	10.5	24.1	28.8	64.9
	Ponderação de modelos	AIC	1.7	9.1	27.0	32.0	69.8
		AIC _C	1.6	9.1	26.7	32.4	69.8

7. CONCLUSÃO

A selecção de modelos baseada na informação Kullback-Leibler (K-L) é uma metodologia recente comparada com os testes de hipóteses e que tem vindo a ser cada vez mais utilizada em várias áreas da ciência. Escolher o melhor modelo de um conjunto de vários modelos possíveis é uma tarefa de extrema importância para a posterior estimação de parâmetros baseada no modelo seleccionado.

A informação K-L, que representa a distância do modelo f_2 (que aproxima a função verdadeira) para a função “verdadeira” f_1 não pode ser utilizada como critério para seleccionar o melhor modelo pois nem a função f_1 nem os parâmetros da função f_2 são conhecidos. É, portanto, necessário utilizar um critério para a selecção de modelos que seja um estimador da informação K-L. O critério utilizado nesta dissertação foi o critério de informação de Akaike (AIC). O modelo seleccionado, de entre um conjunto de modelos escolhido *a priori*, é o que apresentar o menor valor para o AIC. A realização da inferência é, portanto, baseada num único modelo – selecção de modelos tradicional. Mas, se os valores dos AIC’s dos vários modelos do conjunto considerado estiverem muito próximos uns dos outros, a inferência pode ser realizada com base no conjunto de modelos escolhidos *a priori*, em vez de assentar apenas num único modelo. Este método chamado ponderação de modelos, em que são ponderados os modelos, considera também a incerteza na escolha do modelo, o que não acontece com a selecção de modelos tradicional.

Nesta dissertação foi avaliada a performance do AIC, no contexto da amostragem por distâncias, mais precisamente nos transectos lineares. Foi também comparada a selecção de modelos tradicional com a ponderação de modelos, no que diz respeito ao enviesamento das estimativas fornecidas por ambos os métodos. Para a concretização destes objectivos efectuaram-se simulações Monte Carlo, para gerar amostras, provenientes de funções densidade de probabilidade. As amostras foram posteriormente analisadas com o software DISTANCE, provavelmente o mais utilizado em amostragem por distâncias, para a estimação da densidade populacional de animais selvagens. Na análise executada consideraram-se quatro cenários distintos, tendo-se chegado a algumas conclusões que a seguir se apresentam.

A primeira conclusão a que se chegou diz respeito ao AIC. Este critério de selecção de modelos baseado na teoria da informação, é enviesado na escolha do modelo. A tendência é para a escolha de modelos que sejam semelhantes ao verdadeiro, e não para a escolha do verdadeiro modelo. Porém, no caso da simulação da *half-normal*, entre 82% e 92% das amostras o modelo seleccionado tem apenas um parâmetro (à semelhança do modelo verdadeiro) mas no caso da simulação da *hazard-rate*, só entre 18% e 47% é que é seleccionado um modelo com 2 parâmetros.

Outra conclusão a que se chegou, através das simulações prende-se com o enviesamento das estimativas: nota-se que entre 62% e 82% dos casos, o modelo seleccionado pelo AIC apresenta estimativas mais enviesadas do que as dadas por modelos não seleccionados.

A metodologia ponderação de modelos que consiste em estimar os parâmetros usando para tal a ponderação de vários modelos considerados *a priori* foi comparada com a selecção de modelos tradicional em que é seleccionado apenas um único modelo (selecção tradicional) a partir do qual é efectuada a estimação dos parâmetros. Da comparação das duas metodologias resulta que a ponderação de modelos apresenta, em geral, resultados ligeiramente melhores em relação ao outro método, em termos de enviesamento das estimativas determinadas. Os resultados obtidos para as duas funções simuladas (*half-normal* e *hazard-rate*) não são concordantes. Dos quatro cenários analisados, aquele em que existe uma maior diferença entre a ponderação de modelos e a selecção de modelos tradicional, é onde não se truncam os dados e o modelo exponencial negativo faz parte do conjunto de modelos considerados *a priori*, para o caso da *half-normal*. Já no caso da *hazard-rate* o melhor cenário para a ponderação de modelos, é exactamente aquele que Buckland *et al.* (2001) defende para a selecção de modelos tradicional, ou seja, truncar os dados e excluir do conjunto de modelos a função exponencial negativa.

No que respeita ao enviesamento de $\hat{f}(0)$ verificou-se que existe mais enviesamento positivo no caso da *hazard-rate* do que no caso da *half-normal*. Quer na selecção de modelos tradicional, quer na ponderação de modelos, observou-se, em média um enviesamento de $\hat{f}(0)$ entre 11.8% e 19.1% em três dos cenários estudados, no caso da *hazard-rate* (a excepção é notada no caso de dados não truncados e exclusão do modelo exponencial negativo) e apenas num cenário da *half-normal* (que

corresponde aos dados truncados e inclusão do modelo exponencial negativo). No entanto, em relação ao desvio-padrão verifica-se que este é menor no caso da *hazard-rate* do que no caso da *half-normal*.

Por fim, apesar de, em média, a percentagem de enviesamento de $\hat{f}(0)$ nunca ultrapassar os 20%, em nenhum dos casos estudados, verificou-se nalguns cenários uma elevada percentagem de amostras em que a percentagem de enviesamento não se encontrava no intervalo $[-20, 20[$. Uma dessas situações observa-se na simulação da *hazard-rate*, no cenário onde se truncam os dados e se considera o modelo exponencial negativo, no qual mais de 40% das amostras apresentam percentagens de enviesamento fora do intervalo $[-20, 20[$.

APÊNDICE A

A.1 Dedução da fórmula da variância condicional

A dedução aqui apresentada é semelhante à que se pode encontrar em Burnham & Anderson (2002:158).

Na dedução da fórmula serão utilizadas as seguintes definições:

I - $var(\sum_{i=1}^n X_i) = \sum_{i=1}^n var X_i + 2 \sum_{1 \leq i < j \leq n} cov(X_i, X_j)$, para qualquer vector aleatório (X_1, \dots, X_n) .

II - A correlação das variáveis aleatórias X e Y é o número definido por

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}.$$

III - Segundo Buckland *et al.* (1997)

$$var(\hat{\theta}_i) = var(\hat{\theta}_i | \beta_i) + \beta_i^2,$$

onde $\hat{\theta}_i$ são identicamente distribuídos com esperança θ e β_i é o possível enviesamento verificado quando se estima θ a partir do modelo i , isto é, $\theta_i = \theta + \beta_i$. Buckland *et al.* (1997) supõem ainda que β_i tem uma distribuição não especificada com $E(\beta_i) = 0$ e $E(\hat{\theta}_i | \beta_i) = 0$.

Suponha-se que ω_i da fórmula (3.1) são constantes conhecidas.

Então, usando I, obtém-se

$$\begin{aligned} \widehat{var}(\hat{\theta}) &= \widehat{var}\left(\sum_{i=1}^R \omega_i \hat{\theta}_i\right) = \sum_{i=1}^R \widehat{var}(\omega_i \hat{\theta}_i) + \sum_{i=1}^R \sum_{k \neq i}^R cov(\omega_i \hat{\theta}_i, \omega_k \hat{\theta}_k) \\ \widehat{var}(\hat{\theta}) &= \sum_{i=1}^R \omega_i^2 \widehat{var}(\hat{\theta}_i) + \sum_{i=1}^R \sum_{k \neq i}^R \omega_i \omega_k cov(\hat{\theta}_i, \hat{\theta}_k) = \\ &= \sum_{i=1}^R \omega_i^2 \left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] + \sum_{i=1}^R \sum_{k \neq i}^R \omega_i \omega_k \rho_{ik} \sqrt{\widehat{var}(\hat{\theta}_i)} \sqrt{\widehat{var}(\hat{\theta}_k)} = \\ &= \sum_{i=1}^R \omega_i^2 \left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] + \\ &+ \sum_{i=1}^R \sum_{k \neq i}^R \omega_i \omega_k \rho_{ik} \sqrt{\left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] \left[\widehat{var}(\hat{\theta}_k | f_{2k}) + (\hat{\theta}_k - \hat{\theta})^2 \right]} = \end{aligned}$$

Assumindo $\rho_{ik} = \rho$ obtém-se:

$$\begin{aligned}
& \sum_{i=1}^R \omega_i^2 \left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] + \\
& + \sum_{i=1}^R \sum_{k \neq i}^R \omega_i \omega_k \rho \sqrt{\left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] \left[\widehat{var}(\hat{\theta}_k | f_{2k}) + (\hat{\theta}_k - \hat{\theta})^2 \right]} = \\
& = \sum_{i=1}^R \omega_i^2 \left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] - \rho \sum_{i=1}^R \omega_i^2 \left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] + \\
& + \rho \sum_{i=1}^R \omega_i^2 \left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] + \\
& + \rho \sum_{i=1}^R \sum_{k \neq i}^R \omega_i \omega_k \sqrt{\left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] \left[\widehat{var}(\hat{\theta}_k | f_{2k}) + (\hat{\theta}_k - \hat{\theta})^2 \right]} = \\
& = (1 - \rho) \sum_{i=1}^R \omega_i^2 \left[\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2 \right] + \rho \left[\sum_{i=1}^R \omega_i \sqrt{\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2} \right]^2.
\end{aligned}$$

Assumindo correlação perfeita, ou seja, $\rho = 1$, tem-se

$$\widehat{var}(\hat{\theta}) = \left[\sum_{i=1}^R \omega_i \sqrt{\widehat{var}(\hat{\theta}_i | f_{2i}) + (\hat{\theta}_i - \hat{\theta})^2} \right]^2$$

A.2 Dedução da variância de \hat{D}

Na dedução da variância aproximada usou-se a definição apresentada em Casella e Berger (2002) e que consiste no seguinte:

Sejam T_1, \dots, T_j variáveis aleatórias com médias τ_1, \dots, τ_j . Suponha-se que existe uma função diferenciável $h(\underline{T})$ (um estimador de um parâmetro qualquer) para a qual se quer uma estimativa aproximada da variância. Seja $h'_i(\underline{\tau}) = \frac{\partial}{\partial t_i} h(\underline{t})|_{t_1=\tau_1, \dots, t_j=\tau_j}$. Então pode aproximar-se a variância de $h(\underline{T})$ por:

$$var_{\underline{\tau}} h(\underline{T}) \approx \sum_{i=1}^j [h'_i(\underline{\tau})]^2 var_{\underline{\tau}} T_i + 2 \sum_{i>p} h'_i(\underline{\tau}) h'_p(\underline{\tau}) cov_{\underline{\tau}}(T_i, T_p)$$

Assumindo que n e $\hat{f}(0)$ são variáveis aleatórias independentes com esperança $E(n)$ e $E[\hat{f}(0)]$, respectivamente e que $\hat{f}(0)$ é não enviesado, a variância aproximada de \hat{D} é:

$$\begin{aligned}
 \text{var}(\hat{D}) &= \text{var}\left(\frac{n\hat{f}(0)}{2L}\right) = \frac{1}{(2L)^2} \text{var}(n\hat{f}(0)) \approx \\
 &\approx \frac{1}{(2L)^2} \left\{ [E[\hat{f}(0)]]^2 \text{var}(n) + [E(n)]^2 \text{var}(\hat{f}(0)) + \right. \\
 &\quad \left. + 2E(n)E[\hat{f}(0)]\text{cov}(n, \hat{f}(0)) \right\} = \\
 &= \frac{1}{(2L)^2} \left\{ [E[\hat{f}(0)]]^2 \text{var}(n) + [E(n)]^2 \text{var}[\hat{f}(0)] \right\} = \\
 &= \frac{1}{(2L)^2} \left[\hat{f}^2(0) \text{var}(n) + [E(n)]^2 \text{var}[\hat{f}(0)] \right]
 \end{aligned}$$

Supondo $\hat{E}(n) = n$, um estimador de $\text{var}(\hat{D})$ é:

$$\begin{aligned}
 \widehat{\text{var}}(\hat{D}) &= \frac{1}{(2L)^2} \left[\hat{f}^2(0) \widehat{\text{var}}(n) + [\hat{E}(n)]^2 \widehat{\text{var}}[\hat{f}(0)] \right] = \\
 &= \frac{n^2 \hat{f}^2(0)}{(2L)^2} \left[\frac{\widehat{\text{var}}(n)}{n^2} + \frac{\widehat{\text{var}}[\hat{f}(0)]}{\hat{f}^2(0)} \right] = \\
 &= \hat{D}^2 \left\{ \frac{\widehat{\text{var}}(n)}{n^2} + \frac{\widehat{\text{var}}[\hat{f}(0)]}{[\hat{f}(0)]^2} \right\}
 \end{aligned}$$

APÊNDICE B

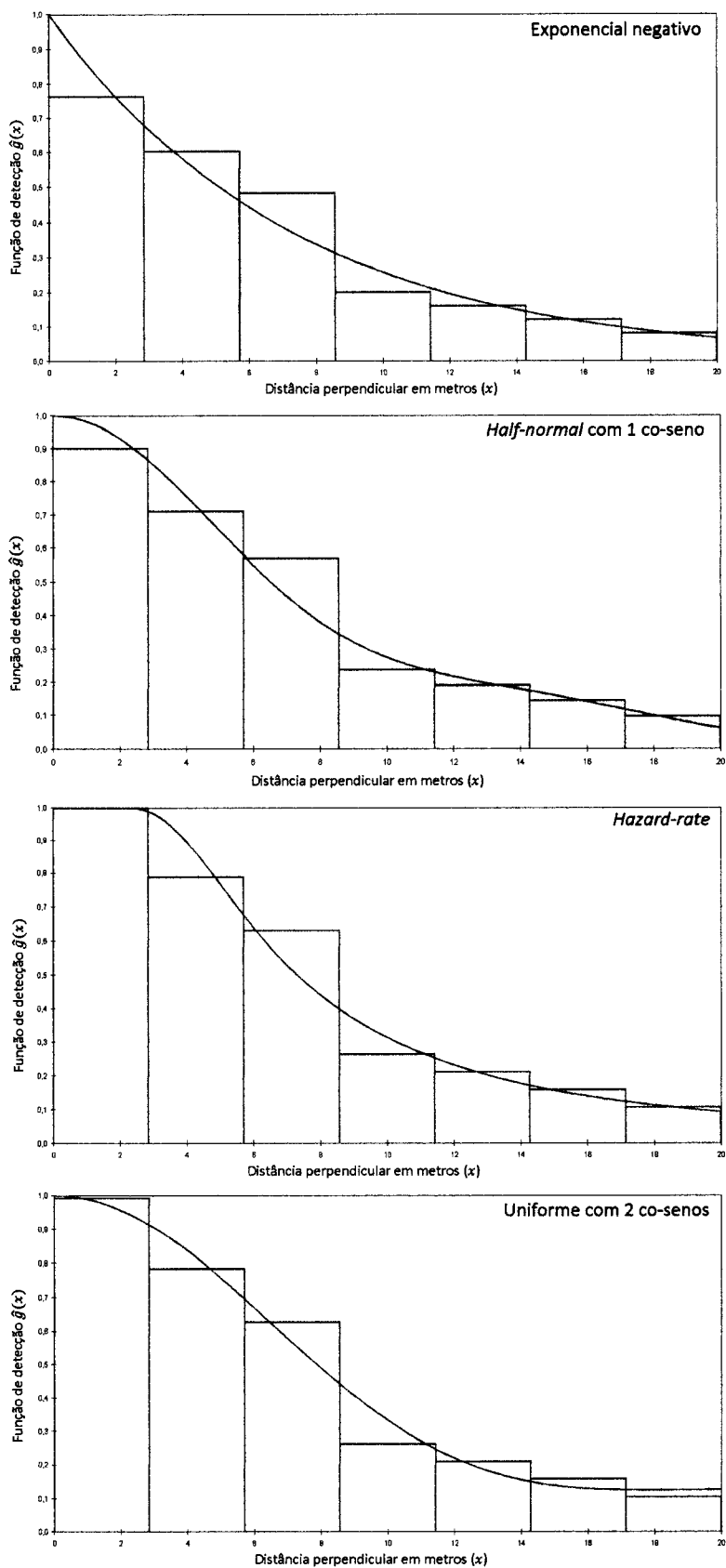


Figura B.1 Modelo ajustado aos dados sobre o histograma para o qual se obtém um melhor ajustamento. Dados relativos ao observador 4 (Laake, 1978).

Quadro B.1. *Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série respeitantes à simulação da função densidade de probabilidade half-normal.*

Função chave	Expansão em série	Nº de ajustamentos	CENÁRIOS			
			Sem truncar e com EN	Sem truncar e sem EN	Truncar e com EN	Truncar e sem EN
HALF-NORMAL	---	---	34.90 %	41.80 %	10.09 %	13.02 %
	Polinómio simples	1	0.20 %	0.20 %	0.00 %	0.00 %
		2	0.20 %	0.40 %	0.00 %	0.00 %
	Co-seno	1	0.20 %	1.80 %	1.01 %	2.22 %
		2	0.00 %	0.00 %	0.00 %	0.20 %
	UNIFORME	---	---	0.00 %	0.00 %	3.53 %
Polinómio simples		1	14.20 %	14.40 %	4.74 %	4.84 %
		2	0.20 %	0.30 %	0.00 %	0.00 %
		3	1.00 %	1.20 %	0.00 %	0.00 %
Polinómio Hermite		1	0.00 %	0.00 %	3.94 %	4.14 %
		2	0.00 %	0.00 %	0.10 %	0.10 %
Co-seno		1	20.90 %	22.20 %	45.41 %	60.14 %
		2	1.10 %	2.30 %	1.72 %	2.93 %
		3	0.50 %	0.50 %	0.00 %	0.00 %
HAZARD-RATE	---	---	10.80 %	12.30 %	5.55 %	8.48 %
	Polinómio simples	1	0.50 %	1.60 %	0.00 %	0.00 %
	Co-seno	1	0.40 %	1.00 %	0.00 %	0.10 %
EXPONENCIAL NEGATIVA	---	---	12.10 %	---	23.82 %	---
	Polinómio simples	1	2.30 %	---	0.00 %	---
	Co-seno	1	0.20 %	---	0.00 %	---
		2	0.30 %	---	0.10 %	---

Quadro B.2. *Percentagem de amostras correspondentes à selecção de cada modelo do tipo função chave + expansão em série respeitantes à simulação da função densidade de probabilidade hazard-rate.*

Função chave	Expansão em série	Número de ajustamentos	CENÁRIOS			
			Sem truncar e com EN	Sem truncar e sem EN	Truncar e com EN	Truncar e sem EN
HALF-NORMAL	---	---	30.20%	35.10%	8.29%	9.20%
	Polinómio simples	1	0.20%	0.20%	0.00%	0.00%
	Polinómio Hermite	1	0.10%	0.10%	0.00%	0.00%
	Co-seno	1	0.60%	3.10%	0.00%	0.20%
		2	1.70%	1.70%	0.10%	0.20%
UNIFORME	---	---	0.00%	0.00%	2.53%	2.53%
	Polinómio simples	1	3.80%	3.80%	4.25%	4.55%
		2	0.00%	0.00%	0.10%	0.10%
		3	4.00%	4.20%	0.30%	0.30%
	Polinómio Hermite	1	0.00%	0.00%	3.34%	3.44%
		2	0.00%	0.00%	0.20%	0.20%
	Co-seno	1	6.90%	7.20%	51.16%	59.45%
		2	5.80%	7.70%	3.24%	4.45%
		3	0.30%	0.40%	0.00%	0.00%
	HAZARD-RATE	---	---	33.60%	36.00%	14.05%
Polinómio simples		1	0.00%	0.40%	0.00%	0.00%
Polinómio Hermite		1	0.00%	0.00%	0.10%	0.10%
Co-seno		1	0.10%	0.10%	0.00%	0.10%
EXPONENCIAL NEGATIVA		---	---	11.60%	---	12.04%
	Polinómio simples	1	1.00%	---	0.00%	---
	Polinómio Hermite	1	0.00%	---	0.00%	---
	Co-seno	1	0.10%	---	0.00%	---
		2	0.00%	---	0.30%	---

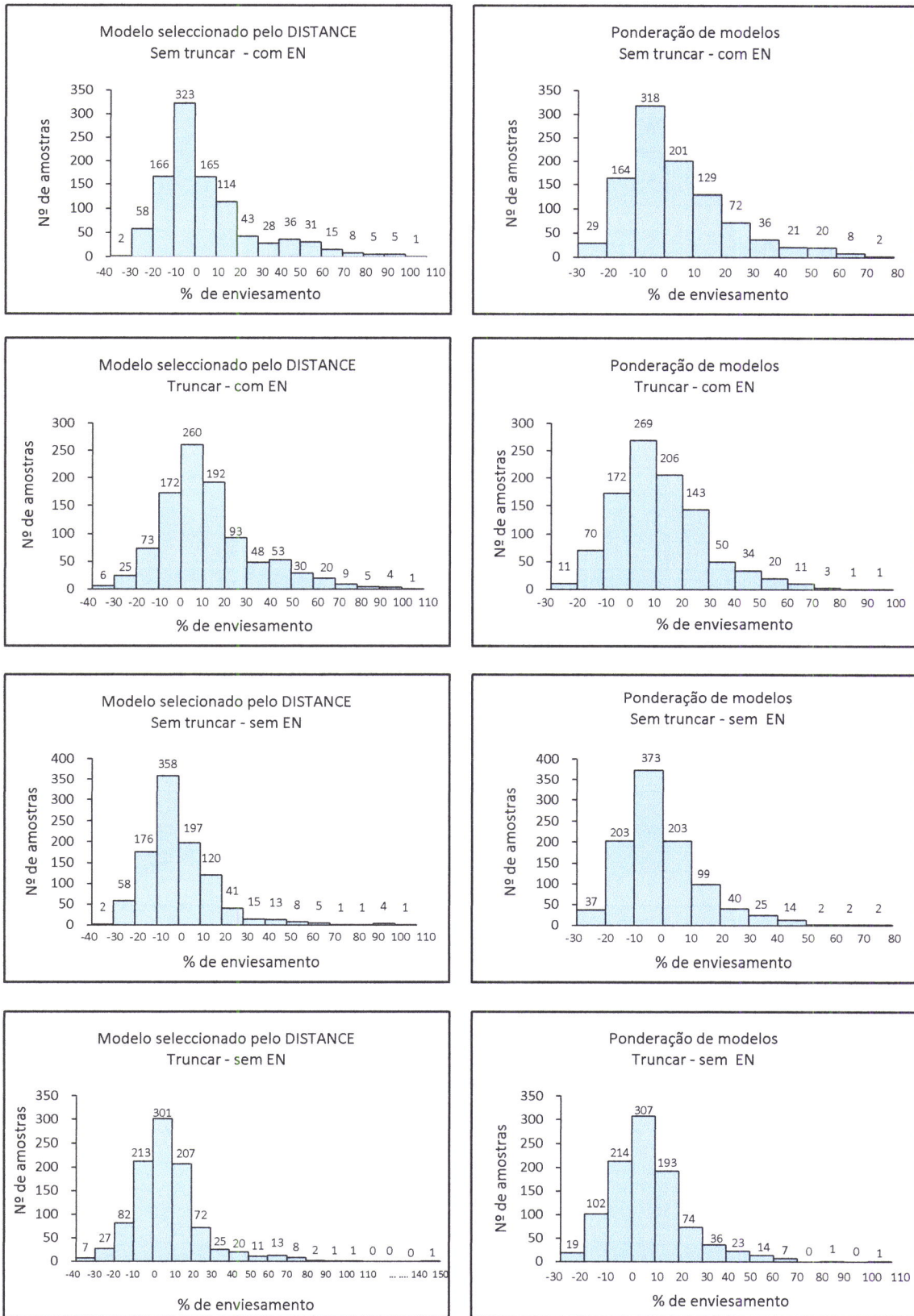


Figura B.2 Distribuições das percentagens de enviesamento de $\hat{f}(0)$, considerando vários cenários, no caso da simulação da função densidade de probabilidade *half-normal*.

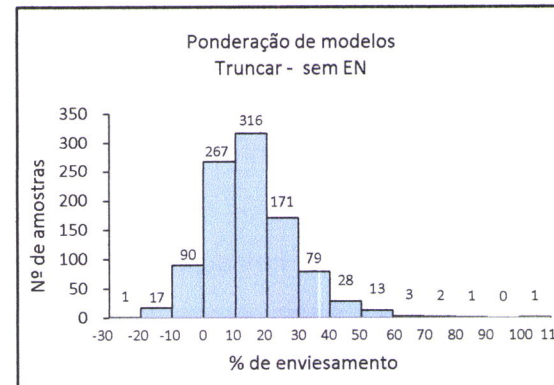
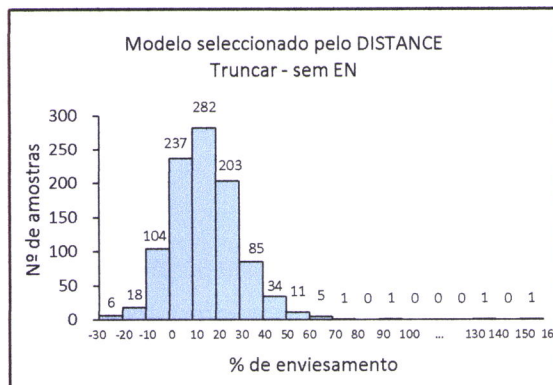
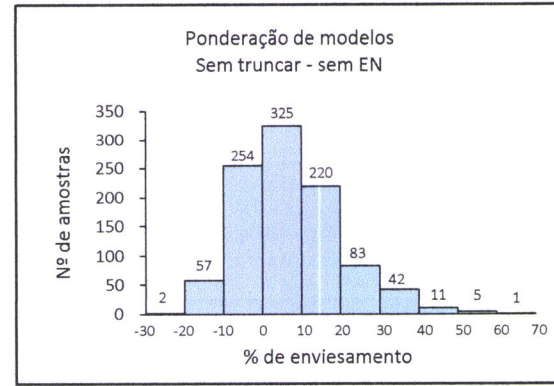
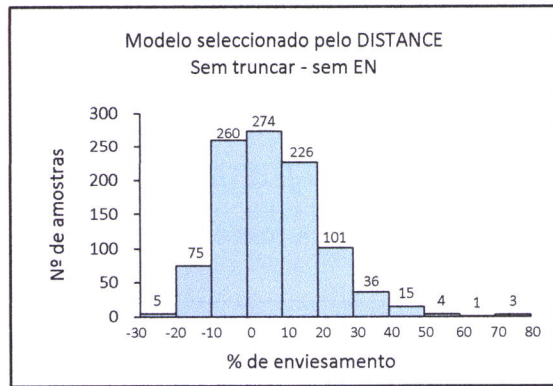
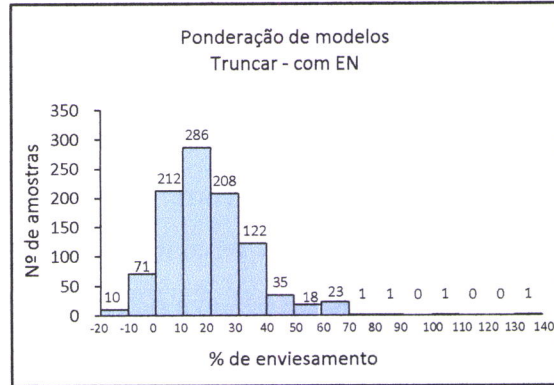
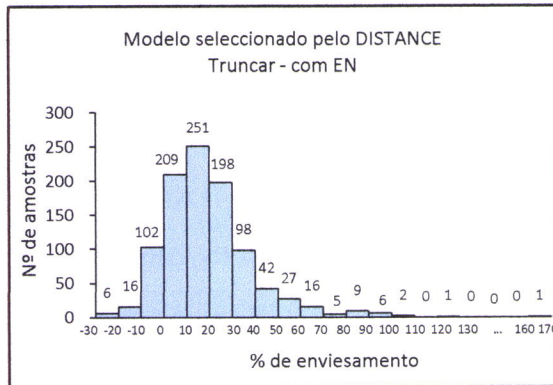
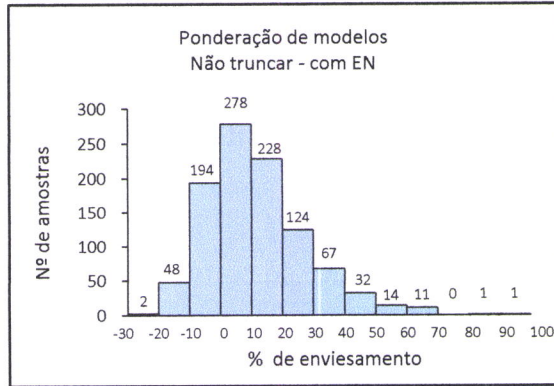
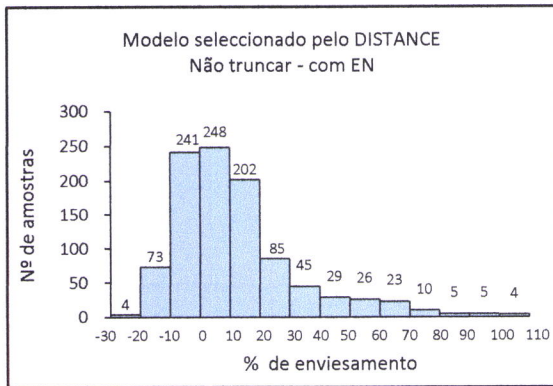


Figura B.3 Distribuições das percentagens de enviesamento de $\hat{f}(0)$, considerando vários cenários, no caso da simulação da função densidade de probabilidade *hazard-rate*.

APÊNDICE C

C.1 Etapas realizadas desde a simulação até à análise dos dados

1 – Simulação, no software R, das amostras provenientes das funções densidade de probabilidade *half-normal* e *hazard-rate*. Na simulação da primeira utilizou-se a função `rhalfnorm(n, theta)`, disponível no software, enquanto que para a segunda foi necessário recorrer ao método da rejeição.

Criou-se um ficheiro para cada amostra.

2 – De forma a correr automaticamente, com o software DISTANCE, todos os ficheiros que contêm as amostras criou-se uma rotina no MS-DOS que permitiu executar o comando MCDS.exe do DISTANCE.

3 – Antes e durante a análise de algumas amostras, pelo DISTANCE, foram gerados os seguintes erros: “*forrtl: error (65): floating invalid*” que não permitia que a análise com o DISTANCE fosse realizada, e “*Error: Negative variance estimate for f(0). Invalid variance.*” e “*Error: One or more estimated CDF is less than zero. Value has been set to zero.*” em que apesar da análise ser realizada, os resultados não eram aceitáveis.

De forma a poder eliminar essas amostras do estudo foi necessário criar duas rotinas que identificavam as amostras cujos outputs continham erros. Essas rotinas foram criadas e corridas com o Cygwin (emulador de aplicações para Linux): uma detectou as amostras que geraram erros antes de ser realizada a análise com o DISTANCE e outra encontrou as amostras cuja análise continha erros.

4 – Após a eliminação das amostras que continham erros, criou-se uma rotina, também usando o Cygwin que permitiu retirar dos ficheiros *stats* gerados para cada amostra pelo DISTANCE, a informação necessária, no que diz respeito aos AIC’s, modelos seleccionados para cada amostra e respectivos parâmetros de interesse.

5 – A escolha do modelo, em cada amostra, foi feita para cada classe de modelos, usando o DISTANCE. Para efectuar a escolha de um modelo de entre as classes consideradas criou-se uma rotina no software R, que realizou essa escolha para cada amostra. Foi também utilizado o software R para determinar os pesos de Akaike e o valor dos parâmetros usando a metodologia da ponderação de modelos.

6 – Finalmente a informação sobre a selecção de modelos tradicional e sobre a ponderação de modelos (valores dos parâmetros e modelos seleccionados) foi tratada no software EXCEL.

C.2 Algumas rotinas de programação utilizadas

1) Rotinas utilizadas no MS-DOS para executar o MCDS.exe do DISTANCE.

```
REM este e' um exemplo
@echo off
@echo a trabalhar com o directorio "dados"
@echo =====
@del /s /q trabalho\exec_status*
@cd dados
@FOR %%F IN (*) DO ..\goFile1 %%F
@cd ..

@echo off
@echo a trabalhar com %1
del /s /q ..\output\%1\*
rd /s /q ..\output\%1
md ..\output\%1
del /q ..\trabalho\entrada.txt
copy %1 ..\trabalho\entrada.txt
REM
REM aplicar o software a "config.txt"
CALL ..\MCDS.exe 0, ..\config.txt 2> ..\trabalho\exec_status2_%1.txt
REM comando para processar um ficheiro
REM ANTES: ..\MCDS.exe ..\dados\%1\*
REM comando para copiar o resultado para a pasta individual
copy ..\output\out.txt ..\output\%1
copy ..\output\log.txt ..\output\%1
copy ..\output\stats.txt ..\output\%1
REM
@echo .
```

- 2) Rotina utilizada no Cygwin para procurar erros nos outputs das amostras corridas no DISTANCE.

```
#!/bin/bash
echo "dar o nome de uma colecao"
COL=$1
echo " "
echo "a trabalhar com a colecao "$COL
if [ -z $COL ]; then
    echo "FALTA o nome da colecao"
    exit 1
fi
echo " "
echo "a procura de log files"
find ./$COL -name *log.txt > _LOGS
cat /dev/null > _ERRORS
echo " "
echo "a procurar erros em cada log"
for x in `cat _LOGS`; do
    grep -H Error $x >> _ERRORS
done
ls -l _ERRORS
exit 0
```

- 3) Rotinas utilizadas no Cygwin para procurar as amostras que não correm no DISTANCE, devido a erros produzidos pelo software.

```
#!/bin/bash
echo "dar o nome de uma colecao"
COL=$1
echo " "
echo "a trabalhar com a colecao "$COL
if [ -z $COL ]; then
    echo "FALTA o nome da colecao"
    exit 1
fi
echo " "
echo "a procura de exec_status files"
find ./$COL/trabalho -name exec_status2* > _EXEC_STATUS
cat /dev/null > _ERRORS_AMOSTRA
echo " "
echo "a procurar erros em cada log"
for x in `cat _EXEC_STATUS`; do
    grep -H error $x >> _ERRORS_AMOSTRA
done
ls -l _ERRORS_AMOSTRA
exit 0
```

- 4) Rotinas utilizadas no Cygwin para separar dados contidos no ficheiro stats do DISTANCE.

```

#!/bin/bash
echo "dar o nome de uma colecao"
COL=$1
echo " "
echo "a trabalhar com a colecao "$COL
if [ -z $COL ]; then
    echo "FALTA o nome da colecao"
    exit 1
fi
rm -f $COL/output/stats.txt
echo " "
D1=Dados_$COL
echo "a preparar colecao de dados "$D1
echo " "
echo "a procura de stats files"
find ./$COL -name *stats.txt > _STATS
# para cada parametro a observar no stats.txt
cat /dev/null > _PARAMS_$D1
cat /dev/null > _AIC_$D1
cat /dev/null > _FO_$D1
cat /dev/null > _AICC_$D1
cat /dev/null > _SERIE_$D1
cat /dev/null > _AJUST_$D1
cat /dev/null > _D_$D1
cat /dev/null > _N_$D1
echo " "
echo "a procurar informagco em cada log"
for x in `cat _STATS`; do
    grep " 2 1 " $x | head -n 1 >> _PARAMS_$D1
    grep " 2 2 " $x | head -n 1 >> _AIC_$D1
    grep " 2 4 " $x | head -n 1 >> _FO_$D1
    grep " 2 7 " $x | head -n 1 >> _AICC_$D1
    grep " 2 14 " $x | head -n 1 >> _SERIE_$D1
    grep " 2 16 " $x | head -n 1 >> _AJUST_$D1
    grep " 4 2 " $x | head -n 1 >> _D_$D1
    grep " 4 3 " $x | head -n 1 >> _N_$D1
#wc -l _PARAMS_$D1
done
exit 0

```

- 5) Rotina do software R para gerar amostras provenientes de uma função densidade de probabilidade *half-normal* e apresentar os ficheiros de dados de modo a serem lidos pelo DISTANCE.

```

library("fdrtool")
set.seed(123)
N<-100
for(i in 1:N){
    theta<-3*pi/56
    n<-70

```

```

hn<-rhalfnorm(n,theta)
m<-matrix(round(hn, digits=6),n,1)
st<-rep("Stratum A",n)
sa<-rep(4,n)
tn<-rep(i,n)
tl<-rep(1,n)
x<-cbind(st, sa, tn, tl, m)
print(x)
#tab<-write.table(x,file="c:/Users/Utilizador/Documents/
Mestrado/Tese/SimDefinitival/Amostras/HalfN.txt",
sep="\t", row.names=FALSE, col.names=FALSE, append=T,
qmethod="double", quote=FALSE)
fich<-paste("c:/Users/Utilizador/Documents/Mestrado/
Tese/SimDefinitival/Amostras/",i,"HN.txt",sep="")
print(fich)
y<-write.table(x,file=fich,sep="\t", row.names=FALSE,
col.names=c("Stratum", "Area", "LineNumber",
"TransectLength", "PerpDistance"), qmethod="double",
quote=FALSE)
}

```

- 6) Rotina do software R para gerar amostras provenientes de uma função densidade de probabilidade *hazard-rate* e apresentar os ficheiros de dados de modo a serem lidos pelo DISTANCE.

```

set.seed(123)
N<-1000
for(i in 1:N){
  haz<-function(x){1-exp(-(x/sigma)^-b)}
  sigma<-7.23952
  b<-3
  w<-20
  u<-runif(1,0,1)
  u1<-w*u
  u2<-runif(1,0,1)
  p<-if(u2<=haz(u1))
  print(u1)
  t<-p
  t
  while(length(t)<70){
    u<-runif(1,0,1)
    u1<-w*u
    u2<-runif(1,0,1)
    p<-if(u2<=haz(u1))
    print(u1)
    t<-append(t,p)
  }
  table<-write.table(t,file="c:/Users/Utilizador/
Documents/Mestrado/Tese/SimHazard/Amostras/Haz.txt",
sep="\t", row.names=FALSE, col.names=FALSE, append=F,
qmethod="double")
A<-as.matrix(read.table("c:/Users/Utilizador/Documents/
Mestrado/Tese/SimHazard/Amostras/Haz.txt", header=FALSE,
sep="\t", col.names=FALSE, dec="."))

```

```

print(A1)
m<-matrix(round(A1, digits=6),length(t),1)
st<-rep("Stratum A",length(t))
sa<-rep(4,length(t))
tn<-rep(i,length(t))
tl<-rep(1,length(t))
dist<-cbind(st, sa, tn, tl, m)
print(dist)
#tab<-write.table(dist,file="c:/Users/Utilizador/
Documents/Mestrado/
Tese/SimHazard/Amostras/Hazard.txt", sep="\t",
row.names=FALSE, col.names=FALSE, append=T,
qmethod="double", quote=FALSE)
fich<-paste("c:/Users/Utilizador/Documents/Mestrado/
Tese/SimHazard/
Amostras/",i,"HZ.txt",sep="")
print(fich)
y<-write.table(dist,file=fich,sep="\t", row.names=FALSE,
col.names=c("Stratum", "Area", "LineNumber",
"TransectLength", "PerpDistance"), qmethod="double",
quote=FALSE)
}

```

C.3 Ficheiro de comandos do DISTANCE que são executados com o MCDS.exe

```

C:\Users\Utilizador\Documents\Mestrado\Tese\SimDefinitiva1\EN1000\
output\out.txt
C:\Users\Utilizador\Documents\Mestrado\Tese\SimDefinitiva1\EN1000\
output\log.txt
C:\Users\Utilizador\Documents\Mestrado\Tese\SimDefinitiva1\EN1000\
output\stats.txt
None
None
None
Options;
Type=Line;
Length /Measure='Kilometer';
Distance=Perp /Measure='Meter'/width=20;
Area /Units='HECTARES';
Object=Single;
SF=1;
Selection=Sequential;
Lookahead=1;
Maxterms=2;
Confidence=95;
Print=Selection;
End;
Data /Structure=Flat;
Fields=STR_LABEL, STR_AREA, SMP_LABEL, SMP EFFORT, DISTANCE;
Infile=C:\Users\Utilizador\Documents\Mestrado\Tese\SimDefinitiva1\EN10
00\
trabalho\entrada.txt /
NoEcho;

```

```
End;  
Estimate;  
Distance;  
Density=SAMPLE;  
Encounter=SAMPLE;  
Print=All;  
Detection=SAMPLE;  
Estimator /Key=NE /Adjust=CO /Criterion=AIC;  
Estimator /Key=NE /Adjust=POLY /Criterion=AIC;  
Estimator /Key=NE /Adjust=HERM /Criterion=AIC;  
Monotone=Strict;  
Pick=AIC;  
GOF;  
VarN=Poisson;  
End;
```

BIBLIOGRAFIA

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, in B.N. Petrov, and F.Csaki, (eds). Akademiai Kiado, Budapest. 267-281.
- Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64, 912-923.
- Azevedo, M. (2004). *Teses, relatórios e trabalhos escolares: sugestões para estruturação da escrita* (4ª ed.). Lisboa: Universidade Católica.
- Azzalini, A. (1996). *Statistical inference: based on the likelihood*. Londres: Chapman and Hall.
- Barnett, V. (1999). *Comparative statistical inference* (2ª ed.). Chichester: Wiley.
- Buckland, S. T., Anderson, D.R., Burnham, K. P., Laake, J. L., Borchers, D. L. & Thomas, L. (2001). *Introduction to distance sampling: estimating abundance of biological populations*. Oxford, UK: Oxford University Press.
- Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model Selection: an integral part of inference. *Biometrics*, 53, 603-618.
- Burnham, K. P., & Anderson, D. R. (2001). Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, 28, 111-119.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: a practical information-theoretic approach* (2ª ed.). New York: Springer.

- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261-304.
- Burnham, K. P., Anderson, D. R., & Laake, J. L. (1980). Estimation of density from line transect sampling of biological populations. *Wildlife Monographs*, 72.
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2^a ed.). Pacific Grove, CA: Duxbury.
- Ellison, A. M. (1996). An introduction to bayesian inference for ecological research and environmental decision-making. *Ecological Applications*, 6, 1036-1046.
- Fisher, R. A. (1935a). The fiducial argument in statistical inference. *Annals of Eugenics*, 6, 391-398.
- Fisher, R. A. (1935b). The logic of inductive inference. *Journal of the Royal Statistical Society*, 98, 39-54.
- Fisher, R. A., (1959). *Statistical methods and scientific inference*. (2^a ed.). London: Oliver e Boyd.
- Guthery, F. S., Peterson, M. J., & Lusk, J. J. (2005). Information theory in wildlife science: critique and viewpoint. *Journal of Wildlife Management*, 69, 457-465.
- Hampel, F. (1998). *On the foundations of statistics: a frequentist approach*. Research report no. 85. Seminar für Statistik, Eidgenössische Technische Hochschule. Zürich.
- Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society, serie B* 41, 190-195.
- Hurvich, C. M., & Tsai, C-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.

- Johnson, J. B., & Omland, K. S. (2004). Model selection in ecology and evolution. *Trends in Ecology and Evolution*, 19, 101-108.
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22, 79-86.
- Laake, J. L. (1978). *Line transect estimators robust to animal movement*. Tese de mestrado, Utah State University.
- Lebreton, J-D., Burnham, K. P., Clobert, J., & Anderson, D. R. (1992). Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. *Ecological Monograph*, 62, 619-678.
- Lee, P. M. (1997). *Bayesian statistics: an introduction*. (2^a ed.). London: Arnold
- Lehmann, E. L. (1986). *Testing staistical hypotheses* (2^a ed.). New York: Springer.
- Lindsey, J. K. (1999). *Applying generalized linear models*. New York: Springer.
- Lukacs, P. M., Thompson, W. L., Kendall, W. L. Gould, W. R., Doherty JR, P. F., Burnham, K. P., & Anderson, D. R. (2007). Concerns regarding a call for pluralism of information theory and hypothesis testing. *Journal of Applied Ecology* 44, 456-460.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2^a ed.). Londres: Chapman and Hall.
- Paulino, C. D., Turkman, M. A. A.; & Murteira, B. (2003). *Estatística Bayesiana* (2^a ed.). Lisboa: Fundação Calouste Gulbenkian.
- Rissanen, J. (1989). Stochastic complexity in statistical inquiry. *World Scientific, series in Computer Science 15*. Singapore.

- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* 27, 379-423 e 623-656.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, Series B*, 64, 583-639.
- Stefanski, L. A., Alpizar-Jara, R., & Pollock, K. H. (manuscrito em revisão). Assessing the effects of measurement error in line transect sampling.
- Stephens, P. A., Buskirk, S. W., & Martínez del Rio, C. (2007). Inference in ecology and evolution. *Trends in Ecology and Evolution*, 22, 192-197.
- Stephens, P. A., Buskirk, S. W., Hayward, G. D., & Martínez del Rio, C. (2005). Information theory and hypothesis testing: a call for statistical pluralism. *Journal of Applied Ecology*, 42, 4-12.
- Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* A7, 13-26.
- Thomas, L., Laake, J. L., Strindberg S., Marques, F. F. C., Buckland, S. T., Borchers, D. L., Anderson, D. R., Burnham, K. P., Hedley, S. L., Pollard, J. H., Bishop, J. R. B. & Marques, T. A. (2006). *Distance 5.0 Release 2*. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK. <http://www.ruwpa.st-and.ac.uk/distance/>