UNIVERSIDADE DE ÉVORA

Mestrado em Matemática e Aplicações – Biénio 2006/2008

O comprimento descritivo mínimo na amostragem por transectos lineares

Dissertação apresentada por: **Fernando José Mão de Ferro Ceia**

para a obtenção do grau de Mestre em Matemática e Aplicações

Orientador: Professor Doutor Russell Gerardo Alpizar – Jara

UNIVERSIDADE DE ÉVORA

Mestrado em Matemática e Aplicações – Biénio 2006/2008

O comprimento descritivo mínimo na amostragem por transectos lineares

Dissertação apresentada por: **Fernando José Mão de Ferro Ceia**

para a obtenção do grau de Mestre em Matemática e Aplicações

Orientador: Professor Doutor Russell Gerardo Alpizar – Jara

À minha família
Aos meus pais

Agradecimentos

Esta dissertação não teria sido possível concluir sem a preciosa ajuda e compreensão de várias pessoas, a quem tenho a obrigação de prestar os meus agradecimentos.

Ao Professor Doutor Russell Alpizar – Jara, pelo facto de ter aceitado a sua orientação e ainda pela sua disponibilidade, amizade e rigor que demonstrou durante toda a sua elaboração.

Ao Professor Doutor Petri Kontakanen, que apesar de não nos conhecermos pessoalmente, se disponibilizou para fornecer uma preciosa ajuda, que sem a qual teria sido muito difícil a sua conclusão.

Aos Professores Doutora Manuela Neves e Doutor Carlos A. Braumann pelos comentários e sugestões que permitiram melhorar a apresentação deste trabalho e corrigir algumas gralhas.

Aos meus pais João Raimundo Ceia e Manuela da Conceição Mão de Ferro Cordas, pelo empenho que sempre tiveram para que esta fosse concluída e principalmente pelo seu carinho, amizade, ajuda e compreensão que sempre demonstraram durante toda a minha vida.

À minha esposa Susana Ceia e aos meus filhos Diogo e Sofia, pela sua compreensão e ajuda imprescindíveis, que sem eles teria sido impossível a sua conclusão.

À Mestre Fátima Morgado, pela sua disponibilidade em prestar ajuda, quando lhe foi solicitado.

Aos membros da Comissão Executiva da Escola EB23 José Régio Portalegre, pela sua disponibilidade para adequarem o meu serviço docente, à necessidade de tempo para a sua conclusão.

Resumo

Apresenta-se um breve resumo histórico da evolução da amostragem por transectos lineares e desenvolve-se a sua teoria.

Descrevemos a teoria de amostragem por transectos lineares, proposta por Buckland (1992), sendo apresentados os pontos mais relevantes, no que diz respeito à modelação da função de detecção.

Apresentamos uma descrição do princípio CDM (Rissanen, 1978) e a sua aplicação à estimação de uma função densidade por um histograma (Kontkanen e Myllymäki, 2006), procedendo à aplicação de um exemplo prático, recorrendo a uma mistura de densidades.

Procedemos à sua aplicação ao cálculo do estimador da probabilidade de detecção, no caso dos transectos lineares e desta forma estimar a densidade populacional de animais.

Analisamos dois casos práticos, clássicos na amostragem por distâncias, comparando os resultados obtidos.

De forma a avaliar a metodologia, simulámos vários conjuntos de observações, tendo como base o exemplo das estacas, recorrendo às funções de detecção semi-normal, taxa de risco, exponencial e uniforme com um co-seno. Os resultados foram obtidos com o programa DISTANCE (Thomas *et al.*, in press) e um algoritmo escrito em linguagem C, cedido pelo Professor Doutor Petri Kontkanen (Departamento de Ciências da Computação, Universidade de Helsínquia). Foram desenvolvidos programas de forma a calcular intervalos de confiança recorrendo à técnica *bootstrap* (Efron, 1978).

São discutidos os resultados finais e apresentadas sugestões de desenvolvimentos futuros.

The minimum description length in line transect sampling

Abstract

We present a brief historical note on the evolution of line transect sampling and its theoretical developments.

We describe line transect sampling theory as proposed by Buckland (1992), and present the most relevant issues about modeling the detection function.

We present a description of the CDM principle (Rissanen, 1978) and its application to histogram density estimation (Kontkanen and Myllymäki, 2006), with a practical example, using a mixture of densities.

We proceed with the application and estimate probability of detection and animal population density in the context of line transect sampling. Two classical examples from the literature are analyzed and compared.

In order to evaluate the proposed methodology, we carry out a simulation study based on a wooden stakes example, and using as detection functions half-normal, hazard rate, exponential and uniform with a cosine term. The results were obtained using program DISTANCE (Thomas *et al.*, in press), and an algorithm written in C language, kindly offered by Professor Petri Kontkanen (Department of Computer Science, University of Helsinki). We develop some programs in order to estimate confidence intervals using the bootstrap technique (Efron, 1978).

Finally, the results are presented and discussed with suggestions for future developments.

Abreviaturas

D - Densidade populacional.

A - Área onde se distribui a população de interesse.

a -Área amostrada.

N- Número de indivíduos, na área de interesse.

L - Comprimento do transecto.

$g(y)$ - Função de detectabilidade.

$\mathcal{D}$- Conjunto dados.

$\mathcal{H}$ – Conjunto de hipóteses.

CDM- Comprimento descritivo mínimo. (Da literatura inglesa MDL - *Minimum description length.*)

MVN- Máxima verosimilhança normalizada. (Da literatura inglesa NML- *Normalized maximum likelihood.*)

CE- Complexidade Estocástica. (Da literatura inglesa SC - *sthocastic complexity.*)

AIC- *Akaike information criterion.*

$R_{h_k}^n$ - Complexidade paramétrica de um histograma com K classes.

$f_{H_{CDM}}\left(x^n \mid C\right)$- Função densidade MVN, para um histograma com um conjunto de pontos de corte C.

$\hat{f}_{CDM}(0)$ - Estimador da densidade de probabilidade por CDM sobre o transecto.

$\hat{D}_{CDM}$ - Estimador da densidade populacional por CDM.

Índice

Índice de figuras

Índice de tabelas

1. Introdução

Nos últimos anos tem-se vindo a assistir a grandes alterações climáticas, que têm conduzido a uma degradação do meio ambiente, com graves consequências para os sistemas de organismos vivos. Neste sentido, a Ecologia têm vindo a assumir um papel preponderante na gestão, desenvolvimento e manutenção dos recursos naturais, vitais para a sociedade.

A análise destas problemáticas, entre outras metodologias, passa sempre pelo estudo aprofundado dos sistemas de organismos vivos (ecossistemas). Nestes estudos, uma das análises básicas, mas com grande importância para a compreensão do estado do ecossistema é a estimação do tamanho das populações ou a densidade populacional.

Este estudo, na maioria das situações é bastante dispendioso ou por vezes impossível, dado que a contagem de populações naturais está sujeita a condicionantes topográficas, tipo de organismo, habitat e recursos humanos necessários para o implementar. Podemos dar como exemplo a impossibilidade de contar todos os coelhos existentes no Alentejo, ou então como contar todos os elefantes duma savana africana?

Para dar resposta ao tipo questões anteriores, surgiram técnicas estatísticas que nos permitem estimar com margens de erro razoáveis, o tamanho das populações e desta forma, fornecer um indicador do estado do ecossistema.

Os métodos de amostragem de populações animais mais utilizados são captura – recaptura (Pollock *et al.,* 1990) e a amostragem por distâncias (Buckland *et al.*, 2001), sendo que a primeira não será alvo de estudo neste trabalho.

A amostragem por distâncias teve a sua origem no início do século XX, com técnicas bastante rudimentares, destacando-se a contagem de

pássaros no estado de Illinois (Forbes, 1907; Forbes e Gross, 1921). Uma metodologia baseada na contagem de objectos efectuada ao longo de uma estrada foi introduzida por Nice e Nice (1921). Os autores assumiram que os objectos eram todos contabilizados até uma determinada distância da mesma. Este conceito foi posteriormente aperfeiçoado por Kelker (1945), registando as distâncias à estrada dos objectos detectados.

A partir do final da década de 60 começaram a surgir vários estudos com suporte teórico, dos quais se destacam Gates *et al.* (1968), Seber (1973), Burnham e Anderson (1976) e Burnham *et al.* (1980).

Na década de 80 surgiram vários artigos que originaram a teoria mais utilizada actualmente nesta área, destacando-se Hayes e Buckland (1983), Buckland (1985) e Buckland (1992a). Neste último, é sugerido um modelo semi-paramétrico mais robusto, que resultava da combinação de uma função chave, (modelo paramétrico) ajustado por termos de uma série polinomial (modelo não paramétrico).

Com a publicação do livro, Buckland *et al.* (1993), foi também desenvolvido um software para análise dos dados segundo esta teoria, denominado DISTANCE (Thomas *et al.*, in press) tendo como base um outro denominado TRANSECT (Laake *et al.*, 1979). Em Buckland *et al.* (2001) é feita uma actualização desta obra, propondo-se novas metodologias para a estimação e detecção de animais.

Este trabalho surgiu, após ter sido realizada uma análise no DISTANCE de um conjunto de dados, recolhidos numa experiência didáctica realizada no pólo da Mitra da Universidade de Évora. Recorreu-se a um histograma de classes desiguais, tendo sido os extremos das mesmas, escolhidos de forma empírica até o histograma apresentar o padrão clássico obtido na amostragem por distâncias.

Esta metodologia, embora empírica e sem qualquer suporte teórico, forneceu alguns resultados interessantes, pelo que procedeu-se com uma

pesquisa no meio científico de teorias sobre estimação de uma função densidade, recorrendo a um histograma de classes desiguais.

Esta dissertação é constituída por seis capítulos, mais os anexos correspondentes a algumas rotinas implementadas no ambiente R (http://www.r-project.org/), necessárias para proceder às análises das simulações.

Do primeiro capítulo, onde se procede a uma breve introdução, passamos ao segundo, onde se apresentam os fundamentos teóricos da amostragem por transectos lineares. No terceiro capítulo é apresentada a teoria subjacente à estimação de uma função densidade, recorrendo a histogramas de classes desiguais por CDM e exemplos de aplicação.

No quarto capítulo procede-se à aplicação da metodologia referida no capítulo anterior, à estimação de uma função densidade no caso dos transectos lineares, e respectiva aplicação a dois exemplos clássicos na literatura.

No quinto capítulo são implementadas simulações, de diversos cenários tendo como base de partida o exemplo das estacas, recorrendo às funções de detectabilidade mais comuns na prática, procedendo-se de igual forma a uma comparação com os resultados obtidos pela teoria proposta por Buckland *et al.*, (2001).

Por fim, no sexto capítulo, apresentam-se as conclusões e propostas de desenvolvimentos futuros.

2. Amostragem por transectos lineares

Em ecologia procede-se ao estudo da distribuição e abundância de plantas e animais, e as suas interacções com o meio ambiente. Muitos estudos sobre populações biológicas, requerem o cálculo de estimadores da densidade populacional, ou o seu tamanho ou ainda, taxas de variação da população no tempo.

Um parâmetro de interesse é a densidade populacional, (número de indivíduos por unidade de área), que se denota por D. Este parâmetro e a dimensão da população N estão relacionados por $N=DA$, em que A representa a área onde se distribui a população de interesse.

Na amostragem por transectos lineares, consideramos uma população de dimensão N, numa determinada área de tamanho A. Coloca-se um conjunto de linhas distribuídas aleatoriamente no campo, e medem-se as distâncias dos objectos detectados à linha, quando esta é percorrida por alguma forma de locomoção. Com esta informação pretende-se inferir a partir do número de objectos observados numa dada área ao longo da linha, ou linhas, para uma região com área superior.

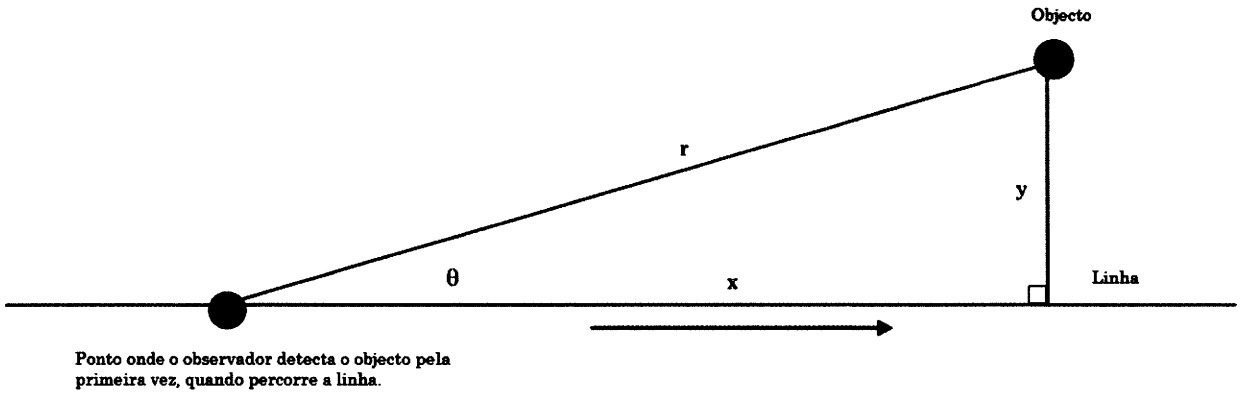

Figura 2-1: Os objectos são detectados ao longo da linha.

Na figura anterior, ilustra-se a metodologia inerente à amostragem por transectos lineares, em que a distância do objecto à linha pode ser calculada através da expressão $y = r \times sen(\theta)$, quando são medidas as distâncias radiais r e o ângulo de observação θ. A linha percorrida tem um comprimento L conhecido sendo que na prática são colocadas q linhas, $L_1, L_2, ..., L_q$, com $L = \sum_{i=1}^{q} L_i$. Objectos afastados da linha poderão não ser detectados, mas se as distâncias forem medidas de forma precisa, podemos obter estimadores fiáveis da densidade populacional.

2.1 Função de detectabilidade

Na metodologia da amostragem por distâncias, um conceito fundamental é a função de detectabilidade, denotada por $g(x)$; $x \geq 0$;

$g(x)$=probabilidade de detectar um objecto sabendo que está a uma distância perpendicular x, da linha central do transecto.

Geralmente a função decresce com o aumento da distância, mas $0 \leq g(x) \leq 1$. Em teoria assume-se que $g(0) = 1$, ou seja, os objectos na linha são sempre detectados com probabilidade 1.

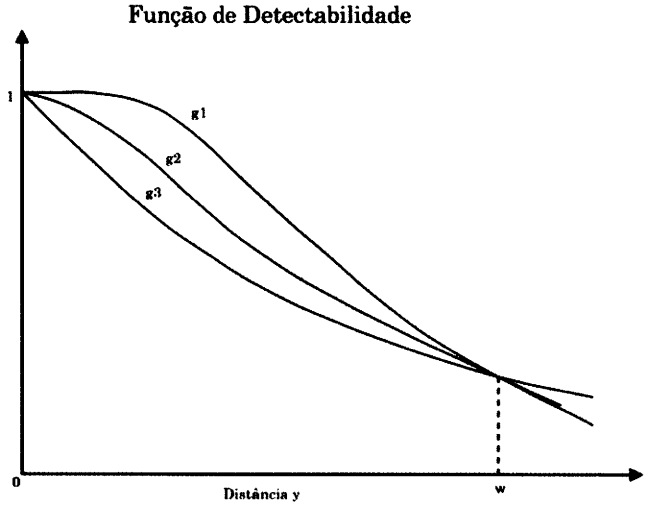

Figura 2-2: Exemplos de funções de detectabilidade.

Na figura anterior, apresentam-se as funções de detectabilidade $g1$, $g2$ e $g3$, exemplos que surgem na amostragem por distâncias. A forma específica destas funções será explicitada mais a frente. A abcissa w representa a distância a partir da qual a probabilidade de observar um objecto é muito pequena, ou a distância máxima à qual são detectados os objectos, $w = \max(x_i)$.

Em geral, numa amostragem por transectos lineares estabelece-se a priori o valor w, tal que $0 \leq x \leq w$, a partir do qual todas as distâncias observadas com valor superior são truncadas. Desta forma, resulta então que a área a ser observada é rectangular e igual a $2wL$.

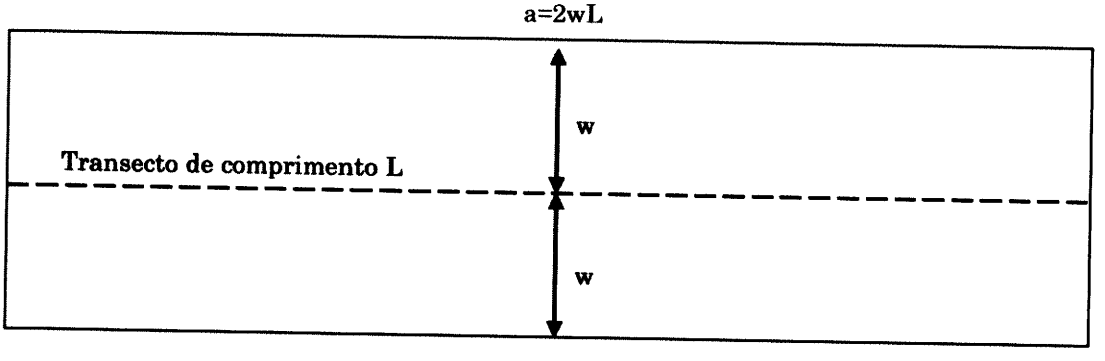

Figura 2-3: Área do transecto.

A truncagem dos dados pode ser vantajosa, na medida que se existirem "*outliers*", estes tornam a modelação da função g difícil. Uma regra para a truncagem consiste em eliminar 5% - 10% dos objectos detectados a maiores distâncias (Buckland *et al.*, 2001).

Toda a metodologia inerente à amostragem por distâncias assenta em fortes pressupostos, que caso venham a ser violados, os resultados obtidos poderão não ter nenhuma consistência. Estes são, por ordem de importância:

- Os indivíduos situados sobre o transecto, são detectados com probabilidade 1 ($g(0) = 1$);

- Os indivíduos são detectados antes de qualquer movimento relativo ao observador;

- As distâncias e ou ângulos são medidos com a máxima precisão possível.

2.2 Estimação da densidade populacional

Consideremos uma área A a ser estudada, com uma população de indivíduos ou objectos de interesse, dispostos segundo um processo estocástico; considere-se uma faixa de comprimento L e largura w, logo uma área a=$2wL$ é observada e se todos os objectos presentes nela forem enumerados resulta então que o estimador do número de objectos por unidade de área vem igual a $\frac{n}{2wL}$.

Na amostragem por transectos lineares define-se por P_a a proporção de objectos detectados na área analisada. Esta proporção é geralmente estimada através das distâncias perpendiculares, e um estimador da densidade populacional é dado por:

$$\hat{D} = \frac{n}{2wL\hat{P}_a}. \qquad [2\text{-}1]$$

Em que n representa o número de objectos detectados, L o comprimento do transecto, w metade da largura observada e $\hat{P}_a$ o estimador da proporção.

Um estimador da probabilidade de detectar um objecto na faixa de área a=$2wL$ é obtido por:

$$\hat{P}_a = \frac{\int_0^w \hat{g}(x)dx}{w}. \qquad [2\text{-}2]$$

Se substituir-mos [2-2] em [2-1], temos que:

$$\hat{D} = \frac{n}{2wL\dfrac{\int_0^w \hat{g}(x)dx}{w}} = \frac{n}{2L\int_0^w \hat{g}(x)dx}. \qquad [2\text{-}3]$$

O integral anterior torna-se então a quantidade crítica que se denota por μ.

Tem-se então:

$$\hat{D} = \frac{n}{2L\hat{\mu}}. \qquad [2\text{-}4]$$

A forma de estimar a quantidade $\frac{1}{\mu}$, resulta do facto, da função densidade de probabilidade das distâncias perpendiculares, condicionada se o objecto é detectado é obtida por:

$$f(x) = \frac{g(x)}{\int_0^w g(x)dx}. \qquad [2\text{-}5]$$

A figura seguinte serve para ilustrar que a probabilidade de detectar um objecto na faixa de área a, é dada pela proporção que representa a área debaixo da curva, relativamente à área do rectângulo.

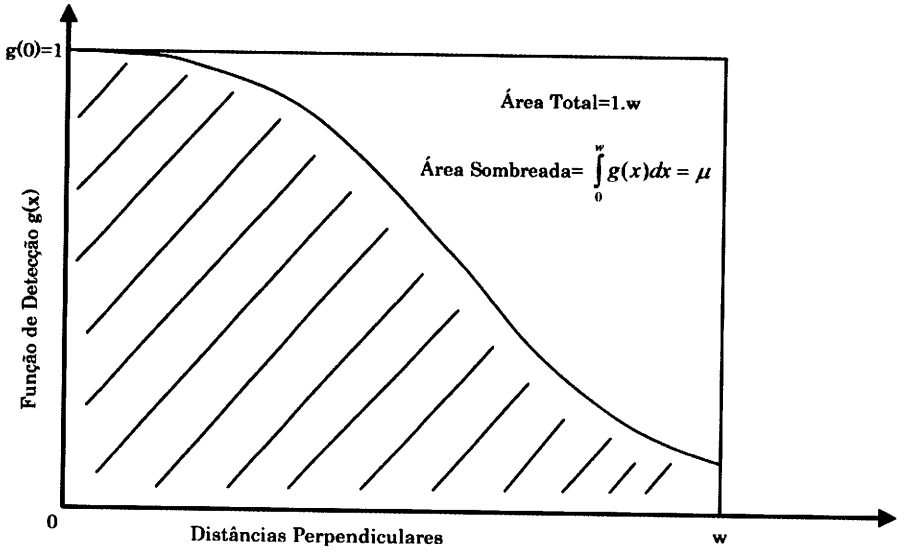

Figura 2-4:Cálculo da probabilidade de detecção na faixa de área a=2 wL .

O resultado [2-5], diz-nos que a função densidade de probabilidade (f.d.p.) é igual à função de detectabilidade $g(x)$, sujeita a uma mudança de escala de forma que $\int_0^w f(x)dx = 1$.

Por hipótese $g(0)=1$, logo tem-se:

$$f(0) = \frac{g(0)}{\int_0^w g(x)dx} = \frac{1}{\int_0^w g(x)dx} = \frac{1}{\mu}. \qquad [2\text{-}6]$$

Então:

$$\hat{D} = \frac{n}{2L\hat{\mu}} = \frac{n\hat{f}(0)}{2L}. \qquad [2\text{-}7]$$

Resulta então, que a questão chave na amostragem por transectos lineares assenta na correcta estimação do parâmetro $f(0)$.

2.3 Estimação semi-paramétrica da função de detecção

Como atrás foi exposto, o problema estatístico na estimação da densidade populacional, resume-se simplesmente à estimativa de $f(0)$ e por consequência, a uma correcta modelação de $g(x)$.

Para tal, várias formas foram apresentadas na literatura, modelos não paramétricos e paramétricos. Relativamente ao primeiro caso, temos os métodos de estimação da densidade pelo método do núcleo (Chen, 1996; Mack e Quang, 1998; Gerard e Schucany, 2002), a utilização de séries de Fourier (Crain *et al.*, 1979), o modelo *logspline* (Rendas, 2001; Rendas e Alpizar - Jara, 2005) e os polinómios hermíticos (Buckland, 1985).

No que diz respeito ao segundo caso, os modelos paramétricos foram introduzidos por Eberhardt (1978) com o modelo logístico reverso, Pollock (1978) com o modelo exponencial de série de potências, Burnham *et al.* (1980) com o modelo Beta e exponencial quadrático e Hayes e Buckland (1983) com o modelo das taxas de risco.

Em Buckland, (1992b), surge uma metodologia semi-paramétrica, utilizando uma função chave normal truncada ajustada por uma serie de polinómios de hermíticos.

Dado que actualmente o modelo com mais desenvolvimentos teóricos e de maior aplicabilidade, é o proposto por Buckland *et al.,* (2001), vamos apresenta-lo de seguida.

Seja $\varphi(x)$ uma função paramétrica que melhor se ajusta aos dados, função chave, que na literatura inglesa se designa por "*key function*", com k parâmetros. Se o ajustamento não for satisfatório, são adicionados termos de uma série não paramétrica, resultando então que a expressão da densidade de probabilidade vem igual a:

$$f(x)=\frac{\varphi(x)}{\gamma}\left[1+\sum_{i=1}^{m}a_i p_i(x_s)\right] \quad , x \geq 0 , \qquad [2\text{-}8]$$

onde $\varphi(x)$ tem k parâmetros, $f(x)$ tem $k+m$ parâmetros e $p_i(x_s)$ é uma função definida por:

$$p_i(x_s)=\begin{cases} x_s^i, \text{ se for um polinómio de ordem } i \\ H_i(x_s), \text{ se for um polinómio hermítico de ordem } i \\ \cos(i\pi x_s), \text{ se for uma série de Fourier (série de co-senos)} \end{cases} ;$$

x_s é um valor estandardizado de x. No caso da função chave ter apenas um parâmetro de escala, σ define-se $x_s=\frac{x}{\sigma}$. No caso de ser escolhido um ajustamento com termos de uma série de Fourier, $x_s=\frac{x}{w}$, isto com o objectivo de evitar eventuais problemas de convergência na estimação dos parâmetros, sendo estas as estandardizações implementadas no software DISTANCE;

$a_i = 0$, se o termo de ordem i não for utilizado no modelo, caso contrário é um parâmetro a ser estimado por métodos de máxima verosimilhança;

γ é uma função normalizadora que depende dos parâmetros.

Na tabela seguinte apresentam-se algumas combinações disponíveis no software DISTANCE.

Função Paramétrica	Série	Função Densidade
Uniforme	Co-senos	$\frac{1}{w}\left[1+\sum_{i=2}^{m} a_i \cos\left(\frac{i\pi x}{w}\right)\right]$
Uniforme	Polinómio	$\frac{1}{w}\left[1+\sum_{i=1}^{m} a_i \left(\frac{x}{w}\right)^{2i}\right]$
Semi-normal	Co-seno	$e^{-\frac{x^2}{2\sigma^2}}\left[1+\sum_{i=2}^{m} a_i \cos\left(\frac{i\pi x}{w}\right)\right]$
Semi-normal	Hermítico	$e^{-\frac{x^2}{2\sigma^2}}\left[1+\sum_{i=2}^{m} a_i H_i(x_s)\right]$, em que $x_s = \frac{x}{w}$
Taxa de risco	Co-seno	$\left(1-e^{-\left(\frac{x}{\sigma}\right)^{-b}}\right)\left[1+\sum_{i=2}^{m} a_i \cos\left(\frac{i\pi x}{w}\right)\right]$

Tabela 2-1- Algumas combinações disponíveis no software DISTANCE. Note-se que os parâmetros das funções devem satisfazer certas condições, para que o seu integral seja igual a um.

2.4 Critérios de escolha do modelo

Da teoria exposta, conclui-se que dispomos de um conjunto variado de curvas possíveis, para modelar a função de detecção. Por um lado isto dá-nos uma grande flexibilidade, mas por outro lado surge-nos outra questão; como seleccionar o melhor modelo? Em Buckland *et al.* (2001), Burnham e Anderson (1976) e Burnham *et al.* (1980) são apresentados quatro critérios que os modelos assumidos para a função de detecção deverão verificar. Estes por ordem de importância são, estimação robusta, critério de forma, eficiência e critério de informação de *Akaike*.

2.4.1 Estimação robusta

Uma vez que a verdadeira função de detecção não é conhecida, a não ser quando são efectuadas simulações em computador, os modelos que possuímos deverão ser facilmente adaptáveis às várias formas que o gráfico da função de detectabilidade possa apresentar.

O estimador obtido deverá ser suficientemente estável na presença de factores perturbadores, tais como a topologia do terreno, condições meteorológicas, estação do ano, etc., ou seja deverão obter-se estimadores pouco sensíveis. Na literatura anglo-saxónica, dá-se a este critério o nome de *pooling robustness*.

2.4.2 Critério de forma

Este critério pode ser matematicamente definido como $g'(0) = 0$, ou seja diz-nos que a função de detectabilidade deverá ter um extremo no ponto de abcissa zero, que em termos gráficos significa que $g(x)$ tem um "ombro" junto da origem. A violação deste critério implica que a robustez da estimativa da densidade populacional é posta em causa, senão mesmo impraticável.

2.4.3 Eficiência

No cálculo das estimativas da densidade populacional é necessário que o modelo seleccionado, além de robusto e satisfaça o critério de forma, deva igualmente ser eficiente, significa isto que as estimativas obtidas devem ter variância reduzida. Para tal os métodos de máxima verosimilhança devem ser utilizados, uma vez que fornecem resultados com boa precisão.

2.4.4 Critério de informação de Akaike (AIC)

Este critério fornece-nos um valor quantitativo, que nos permite definir uma hierarquia entre vários modelos calculados e escolher aquele que apresentar o menor valor. O AIC é definido recorrendo ao logaritmo da função de verosimilhança e ao número de parâmetros do modelo, tendo a seguinte expressão matemática:

$$AIC = -2\ln(L) + 2p. \qquad [2\text{-}9]$$

Em que $\ln(L)$, representa o logaritmo natural do máximo da função de verosimilhança, para os estimadores dos parâmetros e p, o número de parâmetros do modelo.

Desta forma, tenta-se encontrar um modelo que se ajuste de forma razoável aos dados, mas que não tenha um número muito elevado de parâmetros, ou seja, procura-se um modelo que seja parcimonioso.

3. Estimação da função densidade por CDM

No capítulo anterior foi exposta a teoria semi-paramétrica, actualmente utilizada na amostragem por distâncias, e na qual assenta o software DISTANCE, amplamente utilizado por biólogos e investigadores nas áreas da ecologia zoologia e botânica. No entanto esta teoria não é consensual, tendo surgido várias outras abordagens, entre as quais se destacam os métodos de estimação por núcleos (Chen, 1996; Mack e Quang, 1998; Gerard e Schucany, 2002) e *logsplines* (Rendas, 2001; Rendas e Alpizar-Jara, 2005). Pretende-se com este trabalho introduzir um novo conceito no campo da Amostragem por Distâncias, na estimação da função de detectabilidade, baseado numa teoria não paramétrica denominada por CDM (comprimento descritivo mínimo), da literatura inglesa, MDL (*minimum description length*) introduzida por Rissanen (1978).

Na teoria clássica (semi-paramétrica), usualmente procede-se a um agrupamento das distâncias em classes, geralmente de igual comprimento, para desta forma ter uma percepção do tipo de gráfico que terá a função de detectabilidade e desta forma, proceder ao ajuste de um modelo ao histograma apresentado.

Genericamente, a estimação de uma função densidade é um dos principais objectivos da inferência estatística. Perante um determinado conjunto de dados oriundos de uma função densidade desconhecida, a sua estimação por um histograma passa pela escolha de uma função por secções, que melhor represente os dados utilizando um determinado critério. Escolhendo um número suficiente de classes, podemos adaptar um histograma a um certo conjunto de funções densidade.

Na literatura surgiram vários métodos para a escolha do número de classes a considerar para a construção de histogramas, entre os quais podemos citar: Sturges (1926), Scott (1979) e Freedman e Diaconis (1981).

No entanto, todos assentam no mesmo princípio, consideram classes de igual comprimento e dão apenas uma referência para o número óptimo de classes de um determinado conjunto de dados.

Os histogramas obtidos por estas metodologias têm todos os mesmos problemas, se a distribuição dos dados não for uniforme dentro da classe é necessário aumentar o número de classes para capturar as zonas com elevada densidade, e desta forma utilizar um número exagerado de classes nas zonas de baixa densidade. Então a única forma de contornar este problema será considerar classes com comprimentos variáveis.

Para construir estes histogramas, é necessário encontrar um conjunto óptimo de pontos de corte e necessariamente o número de classes, o que nos conduz a um problema mais complexo. Para resolver este problema, consideramos todos os possíveis conjuntos de pontos de corte, como uma selecção de modelos. Nesta metodologia, escolhe-se um primeiro conjunto de pontos de corte e procura-se o modelo óptimo utilizando um critério de selecção de modelos. Esta abordagem é baseada em Teoria da Informação, mais concretamente em métodos de "codificação mínima" ou "complexidade mínima". Estes, procuram tornar a informação num formato o mais compacto possível e para tal baseiam-se no princípio de que, se pretendemos a melhor codificação possível, temos de capturar todas as regularidades presentes nessa informação (Grünwald, 2007).

Uma das formalizações deste princípio é o "comprimento descritivo mínimo - CDM" (Rissanen, 1978). A ideia chave em que assenta este princípio é a de que cada regularidade num determinado conjunto de dados, pode ser usada para os comprimir. De acordo com Rissanen o objectivo principal da inferência indutiva será, "espremer o mais possível os dados até ao máximo de regularidades possíveis". A principal tarefa consiste em extrair a informação significativa do ruído, interpretado como informação acidental.

As regularidades podem ser identificadas com a capacidade de comprimir os dados, o CDM diz-nos que para um dado conjunto de hipóteses $\mathcal{H}$ e conjunto de dados $\mathcal{D}$, devemos tentar encontrar a hipótese ou combinação de hipóteses, que melhor comprime o conjunto $\mathcal{D}$ (Grünwald, 2007).

Esta ideia pode ser aplicada a todos os problemas de inferência, mas tem sido principalmente aplicada a problemas de selecção de modelos.

A forma como o princípio CDM faz esta selecção é minimizando uma quantidade chamada **complexidade estocástica,** que representa o comprimento descritivo mínimo de um conjunto de dados relativamente a uma dada classe de modelos. Esta definição é baseada na função de máxima verosimilhança normalizada (MVN), introduzida por Shtarkov (1987) e Rissanen (1996). O lado prático da MVN envolve elevados recursos de computação, dado que é necessário o cálculo de um integral normalizado ou de uma soma denominada **complexidade paramétrica**, conceito que será explicado mais adiante.

3.1. Comprimento descritivo mínimo

O critério de selecção de modelos CDM, baseia-se na minimização da complexidade estocástica que vamos formalizar de seguida, baseando-nos em Kontkanen e Myllymäki (2006).

Seja $x^n = (x_1,...,x_n)$, uma amostra aleatória de dimensão n, em que cada x_i é um elemento de um espaço de observações X, temos então que $x^n \in X^n$.

Considere-se $\Theta \subset \Re^d$, onde d é um inteiro positivo. A classe de distribuições paramétricas indexadas pelos elementos de Θ é chamada de classe de modelos, isto é M é uma classes de modelos se $M = \{f(.|\theta) : \theta \in \Theta\}$.

Seja $\hat{\theta}(x^n)$ o estimador de máxima verosimilhança dependente de $x^n = (x_1,...,x_n)$. A densidade de máxima verosimilhança normalizada (MVN) é definida por:

$f_{MVN}(x^n \mid M) = \frac{f(x^n \mid \hat{\theta}(x^n), M)}{R_M^n}$, onde a constante de normalização R_M^n, vem igual a $R_M^n = \int_{x^n \in X^n} f(x^n \mid \hat{\theta}(x^n), M) dx^n$ (Shtarkov 1987).

A complexidade estocástica (CE), (*sthocastic complexity*- SC) do conjunto de dados x^n dada uma classe de modelos M, define-se recorrendo à densidade MVN por:

$$CE(x^n \mid M) = -\log f_{MVN}(x^n \mid M) \Leftrightarrow$$
$$\Leftrightarrow CE(x^n \mid M) = -\log\left[\frac{f(x^n \mid \hat{\theta}(x^n), M)}{R_M^n}\right] \Leftrightarrow$$
$$\Leftrightarrow CE(x^n \mid M) = -\left[\log f(x^n \mid \hat{\theta}(x^n), M) - \log R_M^n\right] \Leftrightarrow$$
$$\Leftrightarrow CE(x^n \mid M) = -\log f(x^n \mid \hat{\theta}(x^n), M) + \log R_M^n.$$

Apresentado o princípio CDM e a densidade MVN, vamos de seguida introduzir a definição de densidade MVN para um histograma.

3.2. Máxima verosimilhança normalizada para um histograma

Seja $x^n = (x_1,...,x_n)$ uma amostra de n observações no intervalo $[x_{\min}, x_{\max}]$ em que os extremos são o mínimo e o máximo da amostra. Sem perda de generalidade, assumimos que os dados estão ordenados por ordem crescente e que estão registados com uma precisão denotada por ε, em que $\varepsilon > 0$. Este parâmetro será escolhido em função do número de casas decimais significativas dos dados.

Isto significa que cada $x_i \in x^n \quad i = 1,...,n$ pertence ao conjunto X definido por $\mathrm{X} = \left\{ x_{\min} + j\varepsilon : j = 0,..., \frac{x_{\max} - x_{\min}}{\varepsilon} \right\}$.

Estes pressupostos permitem-nos simplificar a formulação matemática.

Seja $C = (c_1,...,c_{k-1})$ uma sequência crescente de pontos, dividindo o intervalo $\left[x_{\min} - \frac{\varepsilon}{2}, x_{\max} + \frac{\varepsilon}{2} \right]$ em K classes da seguinte forma:

$$\left[x_{\min} - \frac{\varepsilon}{2}, c_1 \right],]c_1, c_2], ...,]c_{k-1}, x_{\max} + \frac{\varepsilon}{2}].$$

Os pontos c_k são chamados pontos de corte do histograma. Vê-se facilmente que entre dois elementos consecutivos de X, existe apenas um ponto de corte. Por questões de simplicidade assume-se que os pontos de corte pertencem ao conjunto definido por:

$$C = \left\{ x_{\min} + \frac{\varepsilon}{2} + j\varepsilon : j = 0,..., \frac{x_{\max} - x_{\min}}{\varepsilon} - 1 \right\}.$$

Isto significa que cada ponto de corte corresponde ao ponto médio entre dois valores consecutivos de X.

Seja $c_0 = x_{\min} - \frac{\varepsilon}{2}, c_K = x_{\max} + \frac{\varepsilon}{2}$ e seja $A_k = c_k - c_{k-1}, k = 1,...,K$ as amplitudes das classes. Dado $\underline{\theta} \in \Theta$, um vector de parâmetros $\underline{\theta} = \left\{ (\theta_1,...,\theta_K) : \theta_i \geq 0, \sum_{i=1}^{K} \theta_i = 1 \right\}$ e uma sequência de pontos de corte C, define-se a densidade do histograma por, $f_H(x \mid \underline{\theta}, C) = \frac{\varepsilon \theta_k}{A_k}$, onde $x \in]c_{k-1}, c_k]$, como se pode observar na seguinte figura.

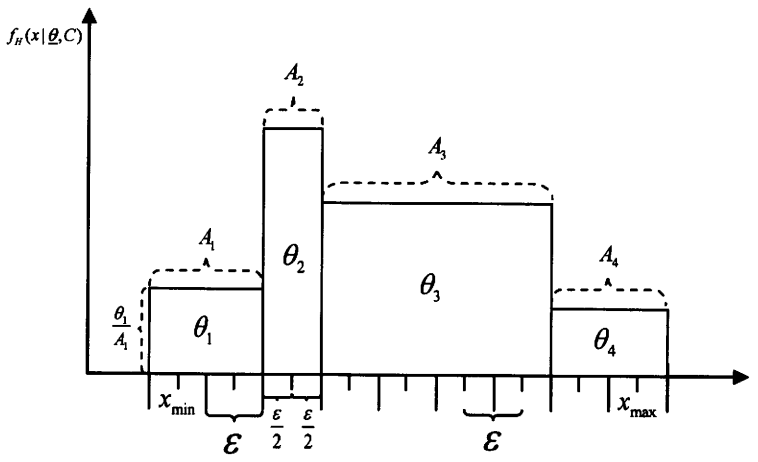

Figura 3-1: Densidade do histograma com classes desiguais.

Note-se que f_H não é uma função densidade, mas sim a probabilidade de um valor de X cair na classe $\left]x-\frac{\varepsilon}{2},x+\frac{\varepsilon}{2}\right]$.

Desta forma, a função de verosimilhança será dada por:

$\mathcal{L}_H(\underline{\theta},C\mid x^n)=\prod_{k=1}^{K}\left(\frac{\varepsilon\theta_k}{A_k}\right)^{h_k}$, onde h_k representa a frequência absoluta da k-ésima classe.

Para encontrarmos a densidade MVN para um histograma, temos de determinar os estimadores de máxima verosimilhança $\hat{\theta}(x^n)=\left(\hat{\theta}_1,...,\hat{\theta}_K\right)$ e calcular a complexidade paramétrica. É sabido que os estimadores de máxima verosimilhança, correspondem às frequências relativas $\hat{\theta}_k=\frac{h_k}{n}$ e desta forma temos:

$$\mathcal{L}_H(\hat{\theta}(x^n),C\mid x^n)=\prod_{k=1}^{K}\left(\frac{\varepsilon h_k}{A_k n}\right)^{h_k}. \qquad [3\text{-}1]$$

Seja $R^n_{h_k}$ a complexidade paramétrica para um histograma de K classes. Note-se que o integral da definição apresentada acima é substituído por uma soma sobre o espaço X^n tendo-se então:

$$R_{h_k}^n = \sum_{x^n \in X^n} \prod_{k=1}^{K} \left(\frac{\varepsilon h_k}{A_k n} \right)^{h_k} = \sum_{h_1+\ldots+h_K=n} \frac{n!}{h_1!\ldots h_K!} \prod_{k=1}^{K} \left(\frac{A_k}{\varepsilon} \right)^{h_k} \prod_{k=1}^{K} \left(\frac{\varepsilon h_k}{A_k n} \right)^{h_k} =$$

$$= \sum_{h_1+\ldots+h_K=n} \frac{n!}{h_1!\ldots h_K!} \prod_{k=1}^{K} \left(\frac{h_k}{n} \right)^{h_k}. \qquad [3\text{-}2]$$

O termo $\left(\frac{A_k}{\varepsilon} \right)^{h_k}$ resulta do facto de que uma classe de comprimento A_k conter exactamente $\frac{A_k}{\varepsilon}$ elementos do conjunto X e o coeficiente multinomial $\frac{n!}{h_1!\ldots h_K!}$, dá-nos o número de formas de distribuir n objectos por K conjuntos, contendo cada um respectivamente $h_1,\ldots,h_K$ objectos.

Resulta então que a densidade MVN para um histograma de K classes vem igual a:

$$f_{H_{MVN}}\left(x^n \mid C\right) = \frac{\mathcal{L}_H(\hat{\theta}(x^n), C \mid x^n)}{R_{h_k}^n}$$

$$f_{H_{MVN}}\left(x^n \mid C\right) = \frac{\prod_{k=1}^{K} \left(\frac{\varepsilon h_k}{A_k n} \right)^{h_k}}{\sum_{h_1+\ldots+h_K=n} \frac{n!}{h_1!\ldots h_K!} \prod_{k=1}^{K} \left(\frac{h_k}{n} \right)^{h_k}}. \qquad [3\text{-}3]$$

A complexidade estocástica para a classe C, resulta igual a:

$$CE_H(x^n \mid C) = -\log f_{H_{MVN}}(x^n \mid C) \Leftrightarrow$$

$$\Leftrightarrow CE_H(x^n \mid C) = -\log \left[\frac{f_H(x^n \mid \hat{\theta}(x^n), C)}{R_{h_k}^n} \right] \Leftrightarrow$$

$$\Leftrightarrow CE_H(x^n \mid C) = -\log \left[\frac{\prod_{k=1}^{K} \left(\frac{\varepsilon h_k}{A_k n} \right)^{h_k}}{R_{h_k}^n} \right] = \sum_{k=1}^{K} -h_k \left[\log(\varepsilon h_k) - \log(A_k n) \right] + \log R_{h_k}^n. \qquad [3\text{-}4]$$

Deduz-se então, que a equação anterior é a chave para aferir a qualidade de um histograma MVN, ou seja comparar diferentes conjuntos de pontos de corte para os dados.

Relativamente ao termo $\log R_{h_K}^n$, em Kontkanen *et al.* (2005), é apresentada uma aproximação cuja expressão é dada por:

$$\log R_{h_K}^n = \frac{K-1}{2}\log\frac{n}{2} + \log\left[\frac{\sqrt{\pi}}{\Gamma\left(\frac{K}{2}\right)}\right] + \frac{\sqrt{2}K\times\Gamma\left(\frac{K}{2}\right)}{3\Gamma\left(\frac{K}{2}-\frac{1}{2}\right)}\times\frac{1}{\sqrt{n}} + \\ + \left[\frac{3+K(K-2)(2K+1)}{36} - \frac{\Gamma^2\left(\frac{K}{2}\right)\times K^2}{9\Gamma^2\left(\frac{K}{2}-\frac{1}{2}\right)}\right]\times\frac{1}{n} + \mathrm{O}\left(\frac{1}{n^{\frac{3}{2}}}\right). \qquad [3\text{-}5]$$

Como o erro de [3-5] converge rapidamente para zero mesmo para valores de n reduzidos, a aproximação pode-se considerar bastante aceitável e de mais fácil tratamento.

3.3. Histogramas CDM óptimos

Neste ponto vamos abordar uma forma prática de encontrar o melhor histograma CDM, para um determinado conjunto de dados.

Em Kontkanen e Myllymäki (2006) é descrito um algoritmo de programação dinâmica, com o objectivo de encontrar o conjunto de pontos de corte, bem como o número de classes óptimos para um dado conjunto de dados, que apresentamos de seguida.

Seja $x^n = (x_1, x_2, ..., x_n)$ uma amostra de dimensão n. Sem perda de generalidade, consideramos $x_1 = x_{\min}$ e $x_n = x_{\max}$. Para escolher os pontos de corte para a amostra anterior, é considerado um conjunto inicial $\tilde{C}$,

constituído por todos os possíveis pontos de corte obtidos, colocando dois pontos possíveis, o mais próximo de cada valor da amostra x_i, como se ilustra em baixo.

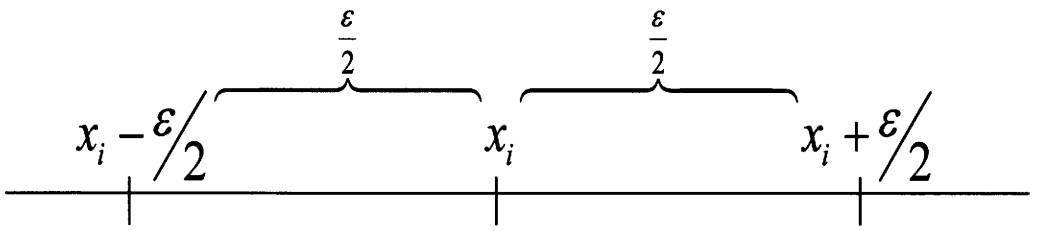

Figura 3-2: Definição dos possíveis pontos de corte.

Desta forma o conjunto $\tilde{C}$ será definido por :

$$\tilde{C} = \left\{ x_i - \frac{\varepsilon}{2} : x_i \in x^n \right\} \cup \left\{ x_i + \frac{\varepsilon}{2} : x_i \in x^n \right\} \setminus \left\{ x_{\min} - \frac{\varepsilon}{2}, x_{\max} + \frac{\varepsilon}{2} \right\} .$$

Note-se que os extremos da amostra são excluídos, uma vez que estão sempre incluídos em todos os conjuntos de pontos de corte.

Trata-se agora de encontrar o conjunto $C \subseteq \tilde{C}$, que minimize o critério [3-4], considerando que cada conjunto de possíveis pontos de corte, será um modelo.

Para tal é necessário ainda considerar que todos os modelos têm a mesma probabilidade de serem seleccionados, isto induz-nos que a distribuição será uniforme para todos os conjuntos de pontos de corte com a mesma dimensão. (Grünwald, 2005).

Para um histograma com B classes e com um possível conjunto de pontos de corte de dimensão $S = \frac{x_{\max} - x_{\min}}{\varepsilon} - 1$, temos $\binom{S}{B-1}$ formas de escolher os extremos das classes (pontos de corte), logo o critério para

comparar os diferentes conjuntos será obtido pela função definida da seguinte forma:

$$BCE(x^n, S, B, C) = CE(x^n \mid C) + \log\binom{S}{B-1} = \\ = \sum_{k=1}^{K} -h_k\left[\log(\varepsilon h_k) - \log(A_k n)\right] + \log R_{h_k}^n + \log\binom{S}{B-1}. \qquad [3\text{-}6]$$

Dado que o número de conjuntos de pontos de corte possíveis é muito grande, vamos apresentar de seguida um algoritmo que nos permite encontrar o conjunto que minimiza o critério [3-6].

Este algoritmo baseia-se num conceito de programação dinâmica, que utiliza uma fórmula recursiva para encontrar a solução óptima. (Bellman, 1957).

Consideremos sem perda de generalidade, que os elementos do conjunto $\tilde{C}$ estão ordenados por ordem crescente, da seguinte forma:

$\tilde{C} = \left\{\tilde{c}_1, ..., \tilde{c}_S\right\}, \tilde{c}_1 < ... < \tilde{c}_S$, consideremos também que $\tilde{c}_{S+1} = x_{\max} + \frac{\varepsilon}{2}$

Seja $\hat{H}_{B,s} = \min_{C \subseteq \tilde{C}} BCE(x^{n_s}, S, B, C)$ [3-7]

em que $x^{n_s} = \left(x_1, ..., x_{n_s}\right)$ corresponde à parte dos dados que se encontram na classe $\left[x_{\min}, \tilde{c}_s\right], s = 1, ..., S+1$. Desta forma, conclui-se que a expressão [3-7] é o valor óptimo de [3-6], quando os dados estão restringidos a x^{n_s}.

Consideremos um histograma com B classes, em que o conjunto de pontos de corte, será definido por $C = \left\{\tilde{c}_{s_1}, ..., \tilde{c}_{s_{B-1}}\right\}$, assumindo que os dados

estão todos no intervalo $\left[x_{\min}, \tilde{c}_{s_B}\right]$ para $\tilde{c}_{s_B} > \tilde{c}_{s_{B-1}}$. Podemos então escrever a função $BCE(x^{n_{s_B}}, S, B, C)$ a partir da função $BCE(x^{n_{s_{B-1}}}, S, B-1, C')$ de um histograma com B - 1 classes e pontos de corte $C' = \left\{\tilde{c}_{s_1}, ..., \tilde{c}_{s_{B-2}}\right\}$, da seguinte forma:

$$BCE(x^{n_{s_B}}, S, B, C) = BCE(x^{n_{s_{B-1}}}, S, B-1, C') - \left(n_{s_B} - n_{s_{B-1}}\right)\left(\log\left(\varepsilon\left(n_{s_B} - n_{s_{B-1}}\right)\right) - \log\left(\left(\tilde{c}_{s_B} - \tilde{c}_{s_{B-1}}\right)n\right)\right) +$$

$$+ \log\frac{R_{h_B}^{n_{s_B}}}{R_{h_{B-1}}^{n_{s_{B-1}}}} + \log\frac{\binom{S}{B-1}}{\binom{S}{B-2}}. \qquad [3\text{-}8]$$

Note-se que $\left(n_{s_B} - n_{s_{B-1}}\right)$ representa o número de dados que se encontram na B-ésima classe, $\left(\tilde{c}_{s_B} - \tilde{c}_{s_{B-1}}\right)$ a amplitude dessa classe e o coeficiente $\log\frac{\binom{S}{B-1}}{\binom{S}{B-2}}$ pode ser simplificado da seguinte forma:

$$\log\frac{\binom{S}{B-1}}{\binom{S}{B-2}} = \log\frac{\frac{S!}{(B-1)!(S-B+1)!}}{\frac{S!}{(B-2)!(S-B+2)!}} = \log\frac{(B-2)!(S-B+2)!}{(B-1)!(S-B+1)!} =$$

$$= \log\frac{(B-2)!(S-B+2)(S-B+1)!}{(B-1)(B-2)!(S-B+1)!} = \log\frac{(S-B+2)}{(B-1)}.$$

A fórmula recursiva resulta então:

$$\hat{H}_{B,s} = \min_{s'}\left\{\hat{H}_{B-1,s'} - \left(n_s - n_{s'}\right)\left(\log\left(\varepsilon\left(n_s - n_{s'}\right)\right) - \log\left(\left(\tilde{c}_s - \tilde{c}_{s'}\right)n\right)\right) + \log\frac{R_{h_B}^{n_s}}{R_{h_{B-1}}^{n_{s'}}} + \log\frac{(S-B+2)}{(B-1)}\right\}. \qquad [3\text{-}9]$$

em que $s' = B-1,...,s-1$.

O processo iterativo inicia-se com:

$$\hat{H}_{1,s} = -n_s\left(\log(\varepsilon n_s) - \log\left(\left(\tilde{c}_s - \left(x_{\min} - \frac{\varepsilon}{2}\right)\right)n\right)\right), s = 1,...,S+1. \qquad [3\text{-}10]$$

Iterativamente, o número de classes é aumentado sucessivamente de um em um e [3-9] é aplicado para $s = B,...,S+1$ até um número máximo de classes $B_{\max}$ ser atingido. O valor mínimo de $\hat{H}_{B,s}$ é então seleccionado para a solução final.

3.4. Exemplo

Para ilustrar a aplicabilidade da teoria apresentada, simulámos amostras de dimensões 100, 500, 1000 e 10000 da mistura de uma densidade semi-normal, com uma densidade normal obtida da seguinte forma:

Consideraram-se $Y_1 \sim semiN(\theta = 0.1)$ e $Y_2 \sim N(50;12)$

$$Z(y) = \sum_{i=1}^{2} p_i Y_i \ , \ p_1 = 0.6; p_2 = 0.4$$

Procedemos à aplicação do algoritmo de programação dinâmica descrito anteriormente, para encontrar o histograma CDM óptimo, para cada amostra gerada, sendo o parâmetro ε fixado em 0.1.

Na tabela seguinte apresenta-se um resumo dos resultados obtidos:

Dimensão da Amostra	*BCE*	B
100	655.29	2
500	3244.16	5
1000	6446.85	6
10000	64369.64	14

Tabela 3-1: BCE representa a complexidade estocástica óptima e B, o número de classes escolhido pelo algoritmo.

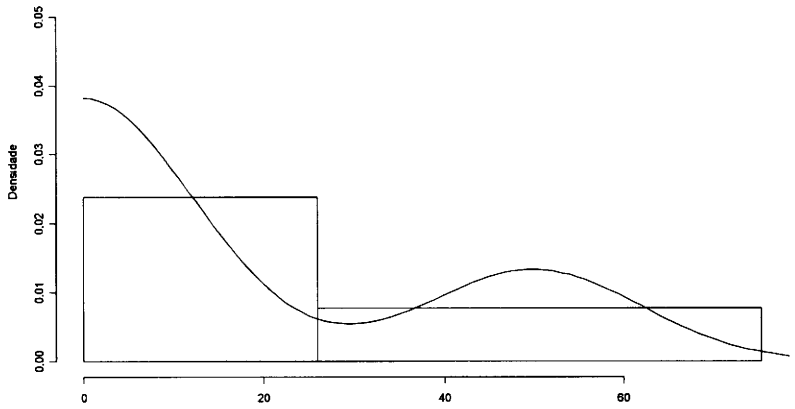

Figura 3-3: Histograma CDM óptimo para a amostra de dimensão 100.

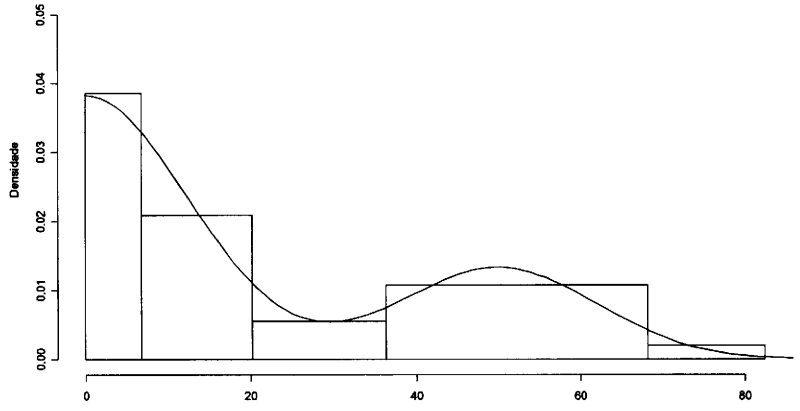

Figura 3-4: Histograma CDM óptimo para a amostra de dimensão 500.

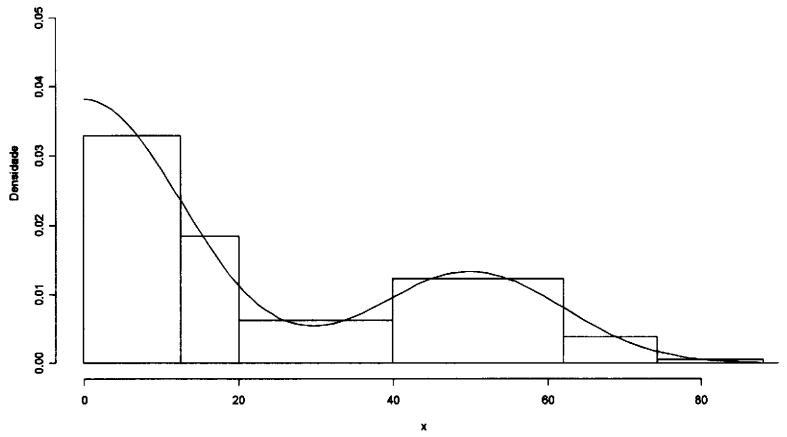

Figura 3-5: Histograma CDM óptimo para a amostra de dimensão 1000.

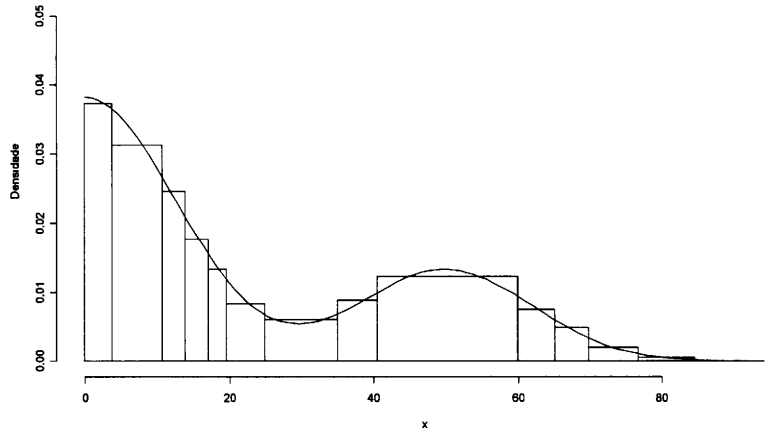

Figura 3-6: Histograma CDM óptimo para a amostra de dimensão 10000.

Verifica-se que à medida que a dimensão da amostra vai aumentando, o algoritmo coloca mais classes para capturar os pormenores da curva estimada. Isto mostra-nos que os histogramas de classes desiguais tem mais flexibilidade em se ajustarem aos dados, do que os de classes equidistantes.

4. Aplicação do CDM em transectos lineares

Da teoria da amostragem por transectos lineares introduzida no capítulo 2, vimos que o estimador da densidade populacional é obtido através da expressão:

$$\hat{D} = \frac{n\hat{f}(0)}{2L}. \qquad [4\text{-}1]$$

$$\text{Dado que } \hat{f}(0) = \frac{1}{\hat{\mu}} = \frac{1}{\text{Área sob curva de } g(x)} \qquad [4\text{-}2]$$

podemos deduzir uma expressão para $\hat{f}(0)$ e para $\hat{D}$, no caso CDM.

Seja $x^n = \{x_1,...,x_n\}$ um conjunto de n valores de distâncias, obtidas de uma amostragem a partir de um transecto de comprimento L e seja $\tilde{C} = \{c_1,...,c_{B+1}\}$, os extremos das B classes do histograma CDM óptimo para os dados anteriores.

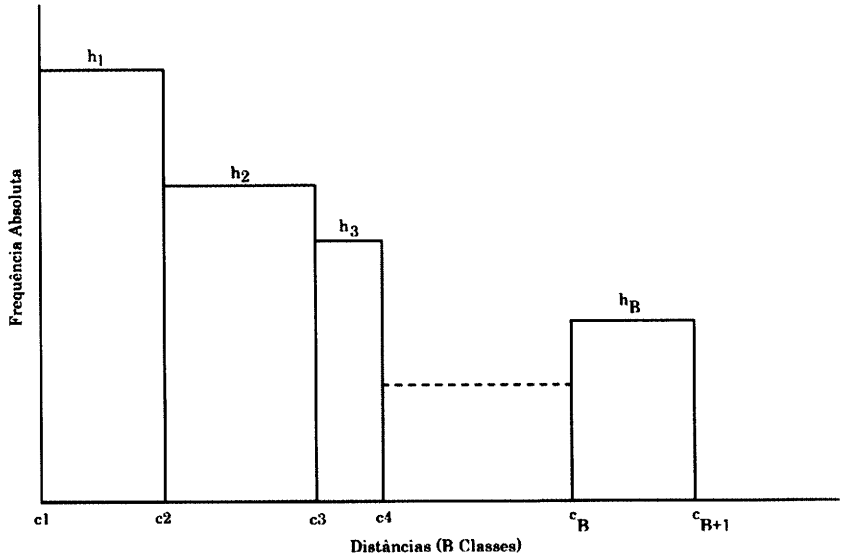

Figura 4-1: Histograma CDM com B classes para a amostra x^n.

Como se pode observar na figura 4.1, deduz-se então que a área do histograma é dada por:

$Área = \frac{1}{n}\sum_{j=1}^{B}(c_{j+1} - c_j) \times h_j$, sendo $(c_{j+1} - c_j)$ a amplitude da classe j, n a dimensão da amostra e h_j a frequência absoluta da classe j.

Note-se que a área considerada, resulta do histograma de frequências relativas, uma vez que $0 < g(x) \leq 1$, sendo g a função de detectabilidade.

Tem-se então:

$$\hat{f}_{CDM}(0) = \frac{1}{\frac{1}{n}\sum_{j=1}^{B}(c_{j+1}-c_j)\times h_j} = \frac{n}{\sum_{j=1}^{B}(c_{j+1}-c_j)\times h_j}. \qquad [4\text{-}3]$$

Resulta então que:

$$\hat{D}_{CDM} = \frac{n\hat{f}_{CDM}(0)}{2L} = \frac{n\times\frac{n}{\sum_{j=1}^{B}(c_{j+1}-c_j)\times h_j}}{2L} = \frac{n^2}{2L\sum_{j=1}^{B}(c_{j+1}-c_j)\times h_j}. \qquad [4\text{-}4]$$

Desta forma, conclui-se então que o estimador da densidade populacional, em amostragem por transectos lineares por CDM é dado por:

$$\hat{D}_{CDM} = \frac{n^2}{2L\sum_{j=1}^{B}(c_{j+1}-c_j)\times h_j} \qquad [4\text{-}5]$$

onde os c_j e h_j $j=1,...,B$ são obtidos como se descreve no capítulo anterior.

4.1. Intervalos de confiança

Uma forma de se obter uma estimativa do desvio padrão e intervalos de confiança, para os estimadores $\hat{f}_{CDM}(0)$ e $\hat{D}_{CDM}$ é recorrendo à técnica "bootstrap" (Efron 1978), em que se procede a uma reamostragem com reposição da amostra inicial, $x^n = \{x_1,...,x_n\}$ calculando para cada réplica, $x_i^{nBoot} = \{x_{1i}^{Boot},...,x_{ni}^{Boot}\}, i=1,...,T$ o valor do parâmetro de interesse.

No nosso caso, designemos por $\hat{f}_{CDM_i}^{Boot}(0)$, o valor de $\hat{f}_{CDM}(0)$ para a i-ésima replica bootstrap, obtido pela aplicação do algoritmo de programação

dinâmica descrito no capítulo anterior, a x_i^{nBoot}, mantendo o parâmetro ε utilizado na amostra inicial e por $\hat{D}^{Boot}_{CDM_t}$ o estimador de $\hat{D}_{CDM}$ para a mesma réplica.

Desta forma, temos para cada réplica um conjunto de pontos de corte óptimo $\tilde{C}_i^{Boot}$, calculando-se $\hat{f}^{Boot}_{CDM_t}(0)$ utilizando [4-3], e $\hat{D}^{Boot}_{CDM_t}$ segundo [4-5].

O estimadores da variância de $\hat{f}_{CDM}(0)$ e $\hat{D}_{CDM}$ podem ser obtidos a partir da variância dos estimadores $\hat{f}^{Boot}_{CDM_t}(0)$ e $\hat{D}^{Boot}_{CDM_t}$, respectivamente das várias réplicas bootstrap.

Recomenda-se um número de réplicas T, entre 200 e 1000, sendo que nos nossos cálculos utilizamos o último valor.

Formalizando, temos que:

$$\mathrm{var}\left[\hat{f}_{CDM}(0)\right] = \frac{1}{T-1}\sum_{i=1}^{T}\left(\hat{f}^{Boot}_{CDM_i}(0) - \overline{\hat{f}^{Boot}_{CDM}(0)}\right)^2, \qquad [4\text{-}6]$$

$$\overline{\hat{f}^{Boot}_{CDM}(0)} = \frac{1}{T}\sum_{i=1}^{T}\hat{f}^{Boot}_{CDM_t}(0), \qquad [4\text{-}7]$$

$$\mathrm{var}\left[\hat{D}_{CDM}\right] = \frac{1}{T-1}\sum_{i=1}^{T}\left(\hat{D}^{Boot}_{CDM_i} - \overline{\hat{D}^{Boot}_{CDM}}\right)^2, \qquad [4\text{-}8]$$

$$\overline{\hat{D}^{Boot}_{CDM}} = \frac{1}{T}\sum_{i=1}^{T}\hat{D}^{Boot}_{CDM_t}.$$

Os intervalos de confiança a 95% para $\hat{f}_{CDM}(0)$ e $\hat{D}_{CDM}$ podem ser obtidos utilizando o método dos percentis, ou assumindo que as distribuições $\left(\hat{f}^{Boot}_{CDM}(0)\right)^T$ e $\left(\hat{D}_{CDM}\right)^T$ são normais, resultando então:

$$\left]\hat{f}^{Boot}_{CDM}(0)_{[0.025]} \;;\; \hat{f}^{Boot}_{CDM}(0)_{[0.975]}\right[\quad \text{(Percentis)} \qquad [4\text{-}9]$$

$$\left]\hat{f}_{CDM}(0)-1.96\sqrt{\text{var}\left[\hat{f}_{CDM}(0)\right]} \;;\; \hat{f}_{CDM}(0)+1.96\sqrt{\text{var}\left[\hat{f}_{CDM}(0)\right]}\right[\text{(Normal)} \; [4\text{-}10]$$

$$\left]\hat{D}^{Boot}_{CDM\ [0.025]} \;;\; \hat{D}^{Boot}_{CDM\ [0.975]}\right[\quad \text{(Percentis)} \qquad [4\text{-}11]$$

$$\left]\hat{D}_{CDM}-1.96\sqrt{\text{var}\left[\hat{D}_{CDM}\right]} \;;\; \hat{D}_{CDM}+1.96\sqrt{\text{var}\left[\hat{D}_{CDM}\right]}\right[\text{(Normal)} \qquad [4\text{-}12]$$

4.2. Software

As simulações foram realizadas utilizando o software gratuito R versão (2.4.1), (ver anexos), permitindo efectuar programação no domínio da estatística, podendo ser obtida mais informação em **http://cran.r-project.org/**.

Quanto aos resultados dos estimadores relativos, à teoria semi-paramétrica descrita no capítulo 2, foram obtidos com o programa DISTANCE, (versão 5.2), encontrando-se informação detalhada em **http://www.ruwpa.st-and.ac.uk/distance/**.

Em Buckland *et al.* (2001), é sugerido a utilização das funções chave e ajustamentos, de acordo com a seguinte tabela:

Função Chave	Ajustamento
Semi-normal	Série de co-senos
Semi-normal	Série de polinómios hermíticos
Uniforme	Série de co-senos
Uniforme	Série de polinómios simples
Taxa de risco	Série de co-senos

Tabela 4-1: Combinações de funções chave e séries de ajustamento utilizadas no DISTANCE.

O modelo seleccionado é aquele que apresenta o menor AIC, definido em [2-9].

No caso da selecção do melhor conjunto de pontos de corte, para a metodologia CDM, foi utilizado o algoritmo de programação dinâmica MDL_Histogram, escrito por Kontkanen e Myllymäki (2006) em linguagem C, tendo sido alvo de algumas modificações, para poder ser executado no ambiente R. Este passou a ser designado por NMLF, tendo-se recorrido ao compilador de C, **lcc-win32 (http://www.cs.virginia.edu/~lcc-win32).**

Este programa funciona de forma nativa em ambiente DOS, recorrendo à linha de comandos do sistema operativo WINDOWS. Os parâmetros que têm que ser introduzidos no software são, o nome do ficheiro de dados, com extensão "TXT" ou "DAT", o número máximo de classes que podem ser consideradas, o valor de ε escolhido e um parâmetro $\delta \geq \varepsilon$, correspondente ao comprimento mínimo das classes que o algoritmo vai construir. O *output* dá-nos o valor da complexidade estocástica óptimo (BCE) definido em [3-7], o número de classes e os extremos das mesmas (pontos de corte).

4.3. Exemplos

4.3.1. Estacas de madeira

Um dos exemplos mais estudados na literatura respeitante à amostragem por transectos lineares, foi realizado por Laake (1978) nos EUA no estado do *Utah*, em que 150 estacas de madeira foram colocadas aleatoriamente numa área rectangular, de forma que a sua distribuição fosse uniforme. Vários observadores percorreram um transecto de 1000m de comprimento, registando as distâncias perpendiculares ao transecto das estacas que observavam. No nosso caso vamos considerar os dados registados pelo observador identificado pelo número 4.

Este exemplo tem a grande vantagem de se conhecer o verdadeiro valor da densidade (37.5 estacas/ha), o que nos permite avaliar os resultados obtidos.

Na tabela seguinte apresenta-se um resumo dos resultados.

Parâmetros			$f(0)$		D	
Metodologia	Critério Selecção	Valores Verdadeiros	0.11		37.5	
		Estimadores	$\hat{f}(0)$	Int.Conf. 95%	$\hat{D}$	Int.Conf. 95%
DISTANCE	AIC[1)]	324.16	0.118	]0.090 ; 0.155[	35.58	]27.18 ; 46.58[
CDM	BCE[2)]	304.66	0.112	]0.058 ; 0.262[	33.74	]17.55 ; 78.87[

Tabela 4-2: Estimadores de $f(0)$ e de D.

[1)] O valor de AIC foi calculado utilizando [2-9]. Os intervalos de confiança no caso CDM, foram calculados usando [4-9] e [4-11].

[2)] O valor de BCE é calculado utilizando [3-7], tendo-se considerado um valor de ε=0.1 e δ=0.2.

Nas figuras seguintes apresentam-se o histograma CDM e a representação gráfica da função densidade estimada pelo DISTANCE.

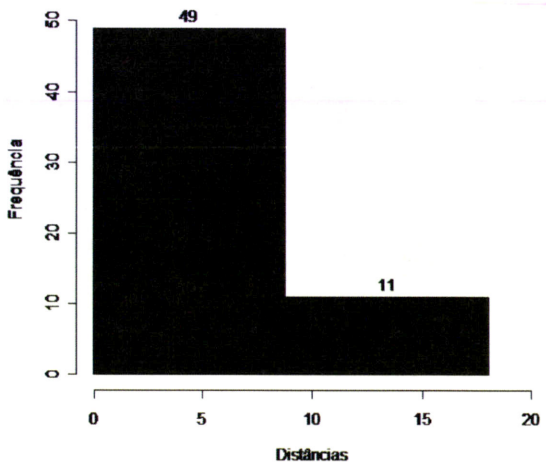

Figura 4-2: Histograma CDM para os dados das estacas (observador 4).

Têm-se apenas duas classes cujos extremos são, -0.05, 8.75 e 18.05.

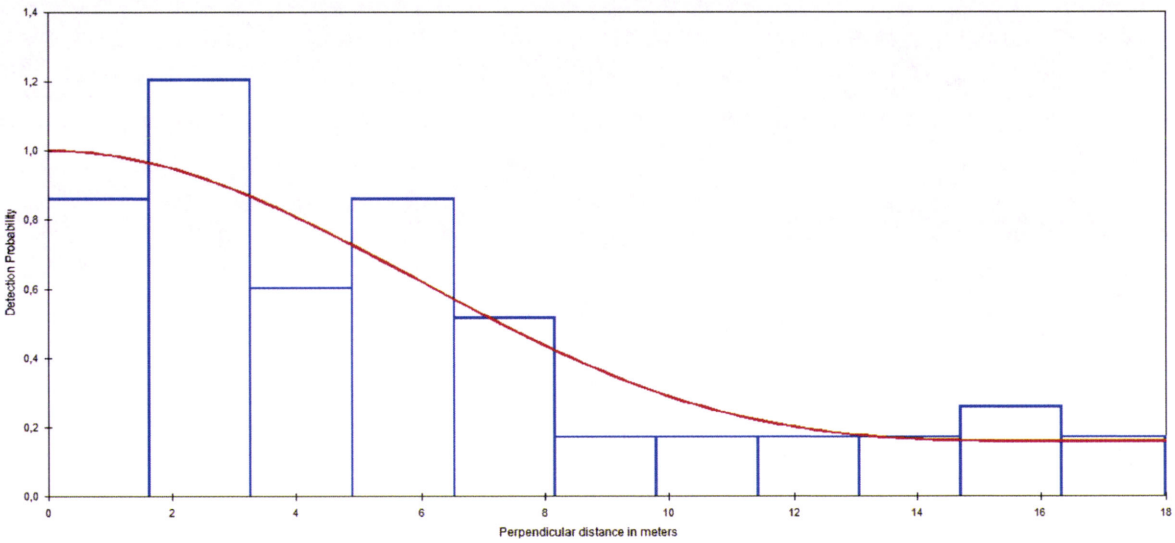

Figura 4-3: Representação gráfica da função densidade estimada pelo DISTANCE, para os dados das estacas (observador 4).

Dada a diferente natureza das duas metodologias, no caso CDM não existe uma expressão analítica da função densidade estimada, como no caso semi-paramétrico, mas sim uma selecção de um modelo que comprima os dados num histograma que se aproxima da densidade desconhecida dos mesmos.

Analisando os resultados expressos na tabela anterior, conclui-se que no caso não paramétrico, o estimador $\hat{f}(0)$ está mais próximo em termos de enviesamento do que o obtido com o DISTANCE, já o mesmo não acontecendo com o estimador da densidade que apresenta maior enviesamento para o CDM.

No respeitante à variabilidade, a metodologia semi-paramétrica leva vantagem para os dois estimadores, dado que os intervalos de confiança para esta tem uma amplitude menor, do que os obtidos na metodologia CDM.

Desta forma, pode-se concluir que a metodologia CDM pode ainda assim fornecer estimadores, que se aproximam bastante dos valores reais.

4.3.2. Ungulados africanos

Outro exemplo que se pode apresentar é o do estudo da densidade populacional dos ungulados no continente africano. Paul Hemingway, Hemingway, (1971), estudou estes animais e numa das suas amostragens registou 73 observações, ao longo de um transecto com 60Km. Os dados que vamos analisar provém de Rendas, (2001).

Metodologia	Critério Selecção	Estimadores	Parâmetros			
			$f(0)$		D	
			$\hat{f}(0)$	Int.Conf.95%	$\hat{D}$	Int.Conf.95%
DISTANCE	AIC[1)]	813.94	0.0063	]0.0053 ; 0.0075[	3.88	]3.26 ; 4.61[
CDM	BCE[2)]	452.22	0.0056	]0.0038 ; 0.023[	3.41	]2.36 ; 14.35[

Tabela 4-3: Estimadores de $f(0)$ e de D.

1) O valor de AIC foi calculado utilizando [2-9]. Os intervalos de confiança no caso CDM, foram calculados usando [4-9] e [4-11].

2) O valor de BCE é calculado utilizando [3-7], tendo-se considerado um valor de ε=0.5 e δ=0.8.

Dado que neste exemplo não se dispõe dos verdadeiros valores de $f(0)$ e de D, podemos afirmar que os estimadores obtidos estão próximos dos que se obtém pela metodologia semi-paramétrica.

5. Simulação

Para avaliar a metodologia CDM na amostragem por transectos lineares, procedeu-se a um conjunto de simulações, utilizando as funções de detectabilidade mais usuais, semi-normal, taxa de risco, exponencial negativa e uniforme com um co-seno. Para tal considerou-se como base para as simulações o exemplo das estacas já utilizado no capítulo 4. Supôs-se que dispúnhamos de uma população de 150 objectos, distribuída aleatoriamente numa dada área, considerando-se um transecto com 1km de extensão e em que a probabilidade de detecção P_a foi fixada em 46.6%, sendo que a distância w foi fixada nos 20m. A dimensão da amostra escolhida é de 70 objectos detectados ao longo do transecto, de modo que $P_a = \frac{n}{N} = \frac{70}{150} = 0.466$.

Geraram-se então 1000 amostras de cada uma das funções de detectabilidade referidas anteriormente, procedendo-se à análise das mesmas pelo algoritmo CDM e pelo DISTANCE, estando os resultados expressos mais à frente.

5.1. Cenários de simulação

Dado que $P_a = \frac{n}{N} = \frac{\mu}{w}$, $\mu = \int_0^w g(x)dx$ e $f(0) = \frac{1}{\mu}$, deduz-se que:

$P_a = \frac{\frac{1}{f(0)}}{w} = \frac{1}{wf(0)}$, como $P_a = \frac{70}{150}$ e $w = 20$, resulta que

$$\frac{1}{20f(0)} = \frac{70}{150} \Leftrightarrow f(0) = \frac{1}{20 \times \frac{70}{150}} \Leftrightarrow f(0) = \frac{3}{28}. \qquad [5\text{-}1]$$

Logo de forma trivial, resulta que $\mu = \frac{1}{f(0)} = \frac{28}{3}$. [5-2]

Como primeira etapa determinaram-se os parâmetros das funções de detectabilidade consideradas, de forma que fossem cumpridos os pressupostos apresentados anteriormente.

5.1.1. Semi-normal

Considerando como função de detectabilidade a semi-normal, tem-se que:

$g(x) = e^{-\frac{x^2}{2\sigma^2}}$ ou então $g(x) = e^{-\frac{x^2\theta^2}{\pi}}$ em que o parâmetro $\theta = \frac{\sqrt{\frac{\pi}{2}}}{\sigma}$, σ - desvio padrão.

Desta forma a função densidade de probabilidade das distâncias, dos objectos detectados vem igual a $f(x) = \frac{2\theta}{\pi} e^{-\frac{x^2\theta^2}{\pi}}$. [5-3]

Dado [5-1] e considerando $f(x)$ definida em [5-3], tem-se que

$f(0) = \frac{2\theta}{\pi}$, logo $\frac{2\theta}{\pi} = \frac{3}{28} \Leftrightarrow \theta = \frac{3\pi}{56}$. [5-4]

Resulta então que $\sigma = \frac{\sqrt{\frac{\pi}{2}}}{\theta} = \frac{\sqrt{\frac{\pi}{2}}}{\frac{3\pi}{56}} = \frac{56\sqrt{\frac{\pi}{2}}}{3\pi}$. [5-5]

Define-se então, a função de detectabilidade semi-normal por:

$g(x) = e^{-\frac{x^2\left(\frac{3\pi}{56}\right)^2}{\pi}} \Leftrightarrow g(x) = e^{-\frac{\left(\frac{3\pi}{56}x\right)^2}{\pi}}$. [5-6]

A função densidade $f(x)$, resulta então igual a:

$$f(x)=\frac{g(x)}{\mu}=\frac{3}{28}e^{-\frac{\left(\frac{3\pi}{56}x\right)^2}{\pi}}\ ,x\in[0,20]. \qquad [5\text{-}7]$$

5.1.2. Taxa de risco

Um modelo para a função de detecção proposto por Hayes e Buckland (1983), é a família de funções taxa de risco definidas por $g(x)=1-e^{\left(-\frac{x}{\sigma}\right)^{-b}}$, em que b representa o parâmetro de forma e σ o parâmetro de escala. (Buckland *et al.*, 1992).

Implementamos este modelo nas simulações, fixando-se o parâmetro de forma $b=3$ e calculou-se o valor do parâmetro de escala, de forma que $\int_0^{20}1-e^{\left(-\frac{x}{\sigma}\right)^{-3}}dx=\frac{28}{3}$, tendo-se obtido o valor $\sigma=7.23952$.

Desta forma, a expressão para a função de detecção taxa de risco, vem igual a:

$$g(x)=1-e^{\left(-\frac{x}{7.23952}\right)^{-3}} \qquad [5\text{-}8]$$

A função densidade $f(x)$ resulta então igual a:

$$f(x)=\frac{g(x)}{\mu}=\frac{1-e^{\left(-\frac{x}{7.23952}\right)^{-3}}}{\frac{28}{3}}\Leftrightarrow$$

$$\Leftrightarrow f(x)=\frac{3}{28}-\frac{3}{28}e^{\left(-\frac{x}{7.23952}\right)^{-3}}\ ,x\in[0,20]. \qquad [5\text{-}9]$$

5.1.3. Exponencial negativa

Gates *et al.* (1968) propuseram um modelo para a função densidade de probabilidade das distâncias perpendiculares, aplicado apenas a dados sem serem truncados e desagrupados, que seguia uma exponencial negativa da forma $f(x)=\lambda e^{-\lambda x}$.

Desta forma, têm-se então que $f(0)=\lambda$. Dada a simplicidade do modelo, por vezes resulta que os estimadores da densidade obtidos são bastante enviesados e imprecisos. De forma a testar a performance da metodologia CDM, implementamos também este modelo nas nossas simulações, com a seguinte parametrização:

Como $f(0)=\lambda$, tem-se que $\lambda=\frac{3}{28}$.

Desta forma a função de detectabilidade vem igual:

$$g(x)=e^{-\frac{3}{28}x} \qquad [5\text{-}10]$$

e do mesmo modo que nos casos anteriores,

$$f(x)=\frac{g(x)}{\mu}=\frac{e^{-\frac{3}{28}x}}{\frac{28}{3}} \Leftrightarrow$$

$$\Leftrightarrow f(x)=\frac{3}{28}e^{-\frac{3}{28}x} \quad , x\in[0,20]. \qquad [5\text{-}11]$$

Refira-se que para este modelo, a condição $f'(0)=0$ não é verificada, mas no entanto continua a ser bastante utilizado na prática.

5.1.4. Uniforme com um co-seno

Da teoria semi-paramétrica exposta no capítulo 2, surge o conceito de função chave, com ajustamentos por uma serie de Fourier. Um dos modelos propostos, utiliza para a função chave a distribuição uniforme com ajustamento de termos de uma série de co-senos, cuja expressão é:

$$f(x)=\frac{1}{w}\left[1+\sum_{i=1}^{m} a_i \cos\left(i\pi x_s\right)\right] \qquad [5\text{-}12]$$

em que x_s representa os valores das distâncias estandardizados, definido por $x_s=\frac{x}{w}$, sendo esta a formulação utilizada para todos os modelos implementados no DISTANCE.

Como vamos utilizar apenas um termo da série de ajustamento, a expressão [5-12], resume-se a $f(x)=\frac{1}{w}\left[1+a\cos\left(\frac{\pi x}{w}\right)\right]$. Tendo em conta os parâmetros utilizados para as simulações, vem que:

$$f(x)=\frac{1}{20}\left[1+a\cos\left(\frac{\pi x}{20}\right)\right]. \qquad [5\text{-}13]$$

Como $f(x)=\frac{g(x)}{\mu}$ e $\mu=\frac{28}{3}$, resulta que a função de detectabilidade vem igual a:

$$g(x)=\mu f(x)=\frac{28}{3}\frac{1}{20}\left[1+a\cos\left(\frac{\pi x}{20}\right)\right]\Leftrightarrow g(x)=\frac{7}{15}\left[1+a\cos\left(\frac{\pi x}{20}\right)\right].$$

Como $g(0)=1$, temos que $\frac{7}{15}[1+a]=1\Leftrightarrow a=\frac{8}{7}$.

Tem-se então, que a expressão para a função de detectabilidade Uniforme com um co-seno, resulta igual a:

$$g(x)=\frac{7}{15}\left[1+\frac{8}{7}\cos\left(\frac{\pi x}{20}\right)\right] \quad , \; x\in[0,20]. \qquad [5\text{-}14]$$

5.2. Estatísticas utilizadas na avaliação dos resultados

Para avaliar a performance dos estimadores da densidade populacional e do parâmetro $f(0)$, calcularam-se para cada conjunto as estatísticas descritivas apresentadas de seguida, denotando θ um parâmetro genérico e N o número de amostras geradas para cada função de detecção.

Desta forma, define-se:

Enviesamento - $env(\hat{\theta})=E(\hat{\theta})-\theta$;

Enviesamento Relativo (em percentagem) - $envR(\hat{\theta})=\frac{E(\hat{\theta})-\theta}{\theta}\times 100\%$;

Valor Médio - $\bar{\hat{\theta}}=\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$;

Variância - $\mathrm{var}(\hat{\theta})=\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i-\bar{\hat{\theta}}_i\right)^2$;

Raiz quadrada do erro quadrático médio -

$$RMSE(\hat{\theta})=\sqrt{\mathrm{var}(\hat{\theta})+\left[env(\hat{\theta})\right]^2};$$

Erro Padrão - $\sigma_{\hat{\theta}}=\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\hat{\theta}_i-\bar{\hat{\theta}}_i\right)^2}$;

Coeficiente de Variação (em percentagem) - $CV(\%)=\frac{\sigma_{\hat{\theta}}}{\hat{\theta}}\times 100\%$.

5.3. Selecção dos modelos

Nas simulações com o programa DISTANCE, implementaram-se dois esquemas de selecção, **selecção 1** e **selecção 2**. Na selecção 1, foram escolhidas as funções chave associadas a cada função de detecção simulada e na selecção 2 utilizaram-se as combinações abaixo descriminadas:

Selecção 2:

semi-normal +Série de co-senos;

semi-normal +Série de polinómios hermíticos;

Uniforme+ Série de co-senos;

Uniforme+ Série de polinómios;

taxa de risco+ Série de co-senos;

Exponencial Negativa+ Série de polinómios.

No que diz respeito às simulações com o algoritmo CDM, utilizou-se um valor para o parâmetro ε de 0.1 e de 0.2 para δ.

5.4. Resultados com a função de detecção semi-normal

Metodologia	$E\left[\hat{f}(0)\right]$	$envR(.)$	var(.)	$RMSE(.)$	$CV(\%)$
CDM	0.101	-5.091527	0.00053323	0.03189317	22.86
DISTANCE Selecção 1	0.108	1.382442	0.00019111	0.01390351	12.80
DISTANCE Selecção 2	0.112	4.830088	0.00056491	0.02432484	21.22

Tabela 5-1: Resultados para a função semi-normal.

Selecção 1:

semi-normal+Série de co-senos;

semi-normal+Série de polinómios;

semi-normal+Série de polinómios hermíticos.

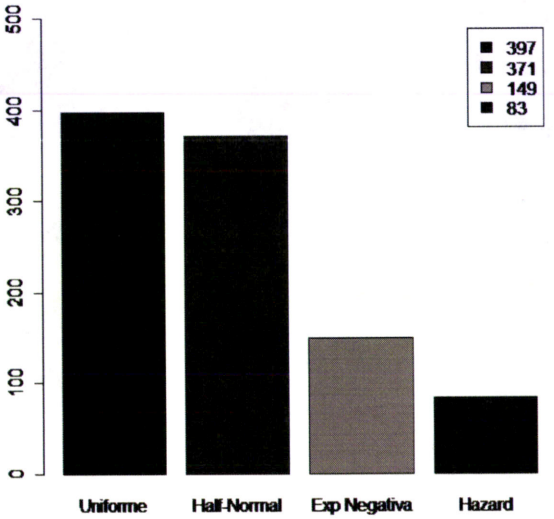

Figura 5-1: Distribuição de frequências dos modelos seleccionados, quando as amostras provêm de uma semi-normal para a selecção 2.

5.5. Resultados com a função de detecção taxa de risco

Metodologia	$E\left[\hat{f}(0)\right]$	$envR(.)$	var(.)	$RMSE(.)$	$CV(\%)$
CDM	0.103	-3.044149	0.00049206	0.02242115	21.53
DISTANCE Selecção 1	0.112	4.647070	0.00025904	0.01684727	14.37
DISTANCE Selecção 2	0.121	12.95799	0.00051271	0.02656073	18.71

Tabela 5-2: Resultados para a função taxa de risco.

Selecção 1:

taxa de risco+Série de co-senos;

taxa de risco +Série de polinómios;

taxa de risco +Série de polinómios hermíticos.

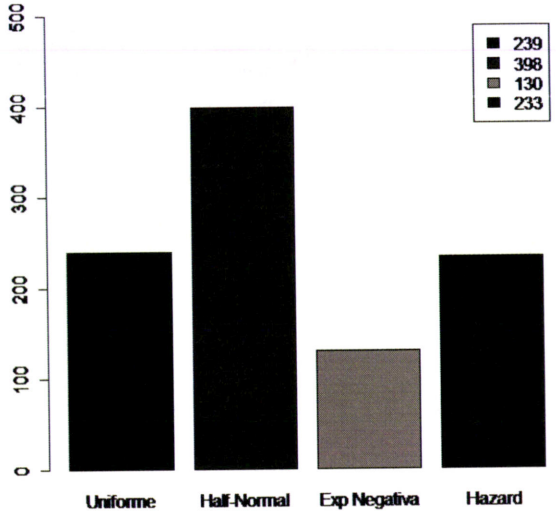

Figura 5-2: Distribuição de frequências dos modelos seleccionados, quando as amostras provêm de uma função taxa de risco para a selecção 2.

5.6. Resultados com a função de detecção exponencial negativa

Metodologia	$E\left[\hat{f}(0)\right]$	$envR(.)$	var(.)	$RMSE(.)$	$CV(\%)$
CDM	0.0906	-15.41002	0.00054843	0.0286538	25.84
DISTANCE Selecção 1	0.119	11.388927	0.00051634	0.0257921	19.09
DISTANCE Selecção 2	0.115	8.029565	0.00122257	0.0360081	30.40

Tabela 5-3: Resultados para a função exponencial negativa.

Selecção 1:

Exponencial negativa+Série de co-senos;

Exponencial negativa +Série de polinómios;

Exponencial negativa +Série de polinómios hermíticos.

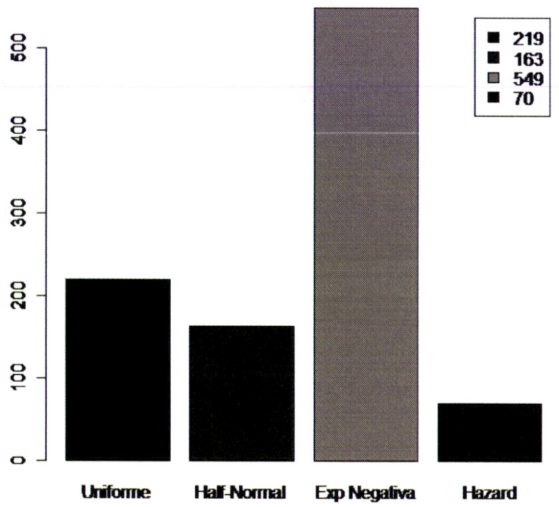

Figura 5-3: Distribuição de frequências dos modelos seleccionados, quando as amostras provêm de uma Exponencial Negativa para a selecção 2.

5.7. Resultados com a função de detecção uniforme com um termo de uma série de co-senos

Com esta função de detecção, implementamos três simulações com o modelo 1. Na primeira simulação consideramos um termo, na segunda dois termos e na terceira três termos. Destaca-se ainda que nesta função, para a metodologia CDM, considerou-se ε=0.1 e δ=0.1.

Metodologia	$E\left[\hat{f}(0)\right]$	$envR(.)$	var(.)	$RMSE(.)$	$CV(\%)$
CDM	0.0984	-8.128268	0.0008559	0.03052587	29.73
DISTANCE Selecção 1-um termo	0.0982	8.3122918	0.0000247	0.01020313	5.06
DISTANCE Selecção 1-dois termos	0.1008	5.8904658	0.0000972	0.0117076	9.78
DISTANCE Selecção 1-três termos	0.1019	4.8004161	0.0001483	0.01322167	11.95
DISTANCE Selecção 2	0.1058	-1.242646	0.0003679	0.01922805	18.13

Tabela 5-4: Resultados para a função uniforme.

Selecção 1:

Uniforme+Série de co-senos;

Uniforme +Série de polinómios;

Uniforme +Série de polinómios hermíticos.

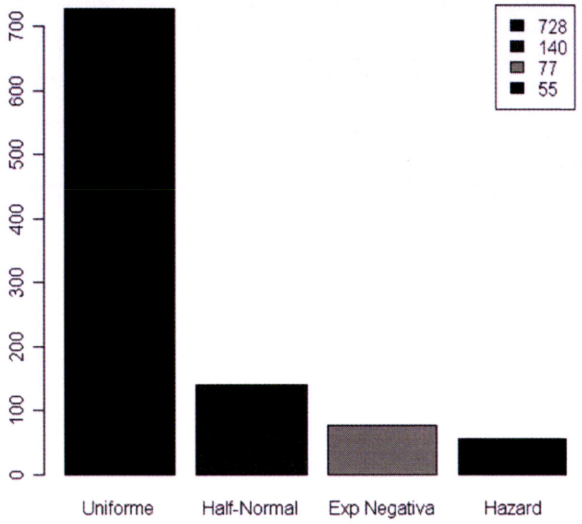

Figura 5-4: Distribuição de frequências dos modelos seleccionados, quando as amostras provêm de uma uniforme com um co-seno para a selecção 2.

6. Conclusões e trabalho futuro

Neste trabalho pretendeu-se introduzir um novo conceito na estimação do parâmetro $f(0)$ e consequentemente na estimação da densidade populacional por transectos lineares.

Após análise dos resultados das várias simulações efectuadas, observa-se que a metodologia CDM, fornece sempre estimadores do parâmetro considerado com um enviesamento negativo, facto que nos indica que existe um sobrestimação da área efectiva de amostragem w. Salienta-se que os resultados obtidos com a selecção 1, são sempre melhores do que os que se obtêm com a selecção 2, para cada uma das metodologias, facto que não é estranho, dado que os modelos incluídos na primeira selecção são do mesmo tipo dos que se implementaram nas simulações.

Na prática, este não é o quadro real, uma vez que o utilizador não tem conhecimento do tipo de modelo de onde provêm os dados que recolheu, optando por um conjunto de modelos sugeridos na literatura, correspondendo nos nossos cenários à selecção 2.

Procedendo a uma análise mais detalhada, verifica-se que com a função de detecção semi-normal, que os resultados obtidos com as metodologias CDM e DISTANCE são bastante comparáveis, a primeira subestima e a segunda sobrestima, ambas em cerca de 5% em valores absolutos, existindo uma ligeira vantagem para esta última.

No que diz respeito à função de detecção taxa de risco, a metodologia CDM fornece um estimador com maior precisão, do que a metodologia semi-paramétrica. Em termos globais existe uma ligeira vantagem para o CDM.

No que diz respeito à função de detecção exponencial negativa, o enviesamento é pior para o CDM, mas a variabilidade é melhor. Em termos globais existe uma vantagem relativamente à metodologia alternativa.

Relativamente à função de detecção uniforme com um co-seno, a metodologia semi-paramétrica leva vantagem nas várias simulações efectuadas, sendo notório o efeito de parcimónia, ou seja quando o número de parâmetros do modelo aumenta, diminui o enviesamento mas verifica-se um incremento da variabilidade.

Em termos gerais conclui-se que a metodologia apresentada, é uma alternativa à semi-paramétrica, no caso das funções de detectabilidade semi-normal, taxa de risco e exponencial negativa, com a vantagem de não ser necessário presumir que modelo se ajustará melhor aos dados.

No que diz respeito a desenvolvimentos futuros, sugere-se a criação de um software amigável para o utilizador, que contenha ajustamento, selecção de modelos e estimação de parâmetros com a metodologia proposta.

Propõe-se efectuar uma comparação de performance com outras metodologias não paramétricas, como estimação pelos métodos dos núcleos (Chen 1996) e *logsplines* (Rendas, 2001; Rendas e Alpizar-Jara, 2005).

Por fim, estabelecer uma conexão com a teoria de selecção e ponderação de modelos (Morgado, 2008).

Bibliografia

Buckland, S. T. (1985). *Perpendicular distance models for line transect sampling*. Biometrics **54**, 1221-1237.

Buckland, S. T. (1992a). *Fitting density functions using polynomials*. Applied Statistics **41**, 63-76.

Buckland, S. T. (1992b). *Maximum likelihood fitting of the Hermite and simple polynomials densities*. Applied Statistics **41**, 241-266.

Buckland, S. T., Anderson, D., Burnham, K. P. e Laake, J. L. (1993). *Distance Sampling: Estimating Abundance of Biological Populations*. Chapman and Hall, London.

Buckland, S. T., Anderson, D., Burnham, K. P., Laake, J. L., Borchers, D. e Thomas, L. (2001). *Introduction to distance sampling – estimating animal abundance of biological populations*. Oxford University Press, Oxford.

Burnham, K. e Anderson, D. (1976). *Mathematical models for nonparametric inferences from line transect data*. Biometrics **32**, 325-336.

Burnham, K., Anderson, D. e Laake, J. L. (1980). *Estimation of density from line transect sampling of biological populations*. Wildlife Monographs **72**, 1-202.

Bellman, R., (1957). *Dynamic Programming*. Princeton University Press. Dover paperback edition (2003).

Chen, S. X. (1996). *Studying school size effects in line transect sampling using the kernel method*. Biometrics **52**, 1283-1294.

Crain, B., Burnham, K., Anderson, D. e Laake, J. (1979). *Nonparametric estimation of population density for line transect sampling using Fourier series*. Biometrical Journal **21**, 731- 748.

Eberhardt, L. (1978). *Transect methods for population studies.* Journal of Wildlife Management **42**, 1-31.

Efron, B. (1978). *Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods.* Biometrika **v68**, 589-599.

Forbes, S. A. (1907). *An ornithological cross-section of Illinois in autumn.* Illinois natural history survey bulletin **7**, 305-335.

Forbes, S. A. e Gross, A. O. (1921). *The orchard of Illinois summer.* Illinois natural history survey bulletin **14**, 1-8.

Freedman, D. e Diaconis, P. (1981). *On the histogram as a density estimator: L_2 theory.* Probability Theory and Related Fields. **57(4)**, 453-476.

Gates, C. E. Marshal, W. H. e Olson, D. P. (1968). *Line transect method of estimating grouse population densities.* Biometrics **24**, 135-145.

Gerard, P. e Schucany, W. (2002). *Combining Population Density Estimates in Line Transect Sampling Using the Kernel Method.* Journal of Agricultural, Biological, and Environmental Statistics, Vol. **7**, No. **2**, 233-242.

Grünwald, P. D. (2005). *Minimum description length tutorial.* MIT Press.

Grünwald, P. D. (2007). *The minimum description length principle.* MIT Press.

Hayes, R. J. e Buckland, S. T. (1983). *Radial distance models for line transect method.* Biometrics **39**, 29-42.

Hemingway, P. (1971). *Field trials of the line transect method of sampling large populations of herbivores.* The Scientific Management of Animal and Plant Communities for Conservation. Blackwell Scientific Publications, Oxford, England, 405-411.

Kelker, G. H. (1945). *Measurements and interpretation of forest that determine populations of managed deer.* Tese de Doutoramento, University of Michigan, Ann Arbor, MI, USA.

Kontkanen, P. e Myllymäki, P. (2006). *Information –Theoretically Optimal Histogram Density Estimation.* Helsinki Institute for Information Techonology.

Kontkanen, P., Myllymäki, P. e Wetting, H. (2005). *NML Computation Algorithms for Tree-Structured Multinomial Bayesian Networks.* Helsinki Institute for Information Techonology.

Laake, J. L. (1978). Line *transect estimators robust to animal movement.* Dissertação de mestrado, Utah State University, Logan, pp-55.

Laake, J. L., Burnham, K. P. e Anderson, D. R. (1979). User manual of program TRANSECT. Utah State University Press, Logan, pp-26.

Mack, P., e Quang, X. (1998). *Kernel Methods in Line and Point Transect Sampling.* Biometrics, **54**, 606-619.

Morgado, M.F.R. (2008). *Selecção de modelos em amostragem por distâncias.* Dissertação de Mestrado, Universidade de Évora, pp-76.

Nice, M. M. e Nice, L. B. (1921). *The roadside census.* Wilson bulletin **33**, 113-123.

Pollock, K. H. (1978). *A family of density estimators for line transect sampling.* Biometrics **34**, 475-478.

Pollock, K. H., Nichols, J., Brownie, C. e Hines (1990). *Statistical inference for capture-recapture experiments.* Wildlife Monographs, 107.

Rendas, L. (2001). *Estimação da Densidade Populacional em Amostragem por Transectos Lineares com Recurso ao Modelo Logspline.* Dissertação de Mestrado, Universidade de Évora, pp-77.

Rendas, L. e Alpizar-Jara, R. (2005). *O modelo logspline aplicado aos transectos lineares.* Em Estatística Jubilar. (Braumann, C.A., Infante, P., Oliveira, M.M., Alpizar - Jara, R. e Rosado, F., eds.), 629-640, Edições SPE, Portugal.

Rissanen, J. (1978). *Modeling by shortest data description.* Automatica **14**, 465-471.

Rissanen, J. (1996). *Fisher information and stochastic complexity.* IEEE Transactions on Information Theory **42 (1)**, 40–47.

Seber G. A. F. (1973). *The Estimation of Animal Abundance.* Hafner, New York.

Scott, D. W. (1979). *On Optimal and Data-Based Histograms.* Biometrics **66**, 605 -610.

Shtarkov, Y. M. (1987). *Universal sequential coding of single messages.* Problems of Information Transmission **23**, 3–17.

Sturges, H. A. (1926). *The choice of a class interval.* Journal American Statistical Association, 65–66.

Thomas, L., S.T. Buckland, E.A. Rexstad, J. L. Laake, S. Strindberg, S. L. Hedley, J. R.B. Bishop, T. A. Marques, e K. P. Burnham. In press. Distance software: design and analysis of distance sampling surveys for estimating population size. Journal of Applied Ecology.

DOI: 10.1111/j.1365-2664.2009.01737.x

Anexos

Anexo 1

Rotina para gerar as amostras da mistura de densidades.

```
set.seed(2)
u<-c(1:1)
f<-c(1:1)
g<-c(1:1)
N<-10000
p1<-0.6
p2<-0.4
u<-runif(N)
j<-1
k<-1
for(i in 1:N)
{
if (u[i]<p1)
      {
      f[j]<-1
      j<-j+1
      }
      else
      {
      g[k]<-2
      k<-k+1
      }
}
u
n1<-length(f)
n2<-length(g)

theta1<-0.1
miu<-50
sigma<-12

d1<-rhalfnorm(n1,theta1)
d2<-rnorm(n2,miu,sigma)
d<-c(d1,d2)
d

write.table(d,"C:/Users/Fernando/Documents/dadosmixhalf10000.txt",dec=
".",row.names=FALSE,col.names=FALSE,sep="")

system("C:/Users/Fernando/Documents/NML_histogram
C:/Users/Fernando/Documents/dadosmixhalf10000.txt 20 0.1 0.2", intern
= TRUE, wait = TRUE,show.output.on.console = TRUE)

cut<-c(-
0.050000,3.750000,10.550000,13.750000,16.950000,19.550000,24.950000,34
.950000,40.550000,59.950000,65.150000,69.750000,76.750000,84.750000,94
.450000)

h<-hist(d,breaks=cut,freq=FALSE,main="Histograma
Mistura",xlab="x",ylab="Densidade",plot=TRUE,ylim=c(0,0.055))
```

```
mx<-function(x) p1*dhalfnorm(x,theta1)+p2*dnorm(x,miu,sigma)
curve(mx,0,90,add=TRUE)
```

Anexo 2

Rotina para gerar as amostras da função de detecção *taxa de risco.*

```
ext<-"txt"
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)
}

write.table(t,paste(paste("c:/AMSHZ/amostrahz",i,sep=""),ext,sep="."),
dec=".",sep="\n",quote=FALSE,row.names=FALSE,col.names=FALSE)#Ficheiro
auxiliar
}
```

Anexo 3

Cálculo dos pontos de corte para as amostras *taxa de risco.*

```
ext<-"txt"
cmd<-"txt 8 0.1 0.2"
nsim<-1000

for (k in 1:nsim)
{
system(paste(paste("c:/NMLF
c:/AMSHZ/amostrahz",k,sep=""),cmd,sep="."),intern=TRUE,wait=TRUE,show.
output.on.console=FALSE,invisible=TRUE)}
```

Anexo 4

Cálculo dos estimadores de f(0) para as amostras *taxa de risco* pelo CDM.

```
f0<-c(1:1)
n<-c(1:1)
A<-c(1:1)
D<-c(1:1)
area<-c(1:1)
width<-c(1:1)
dif<-c(1:1)
ext<-"txt"
nsim<-1000

haz<-function(x){1-exp(-(x/sigma)^-b)}
sigma<-7.23952
b<-3

cut<-
read.table("c:/cuthazard.txt",dec=".",sep=",",fill=TRUE,header=FALSE,c
ol.names=c("c1","c2","c3","c4","c5","c6","c7","c8","c9","c10","c11","c
12","c13"))#Leitura dos pontos de corte para cada amostra

cutvec<-as.matrix(cut)
cutvec

for(k in 1:nsim)
{
#cutvec[k,"c1"]<-0
ams<-
read.table(paste(paste("c:/AMSHZ/amostrahz",k,sep=""),ext,sep="."),dec
=".",sep="",header=FALSE,fill=TRUE)
amsd<-as.matrix(ams)
amsd
n[k]<-length(na.exclude(cutvec[k,]))
c<-n[k]-1

h<-
hist(amsd,breaks=na.exclude(cutvec[k,]),freq=TRUE,col="red",labels=TRU
E,plot=FALSE,include.lowest=TRUE)
h

      for(j in 1:c)
      {
            width[j]<-(cutvec[k,j+1]-cutvec[k,j])
            A[j]<-width[j]*h$counts[j]
      }
      area[k]<-sum(A)

barplot((h$counts/70),width,space=0,xlim=c(0,20),ylim=c(0,1))
curve(haz,0,20,col = 5, lty = 3, lwd = 5, add = TRUE)

f0[k]<-1/(area[k]/70)
}
mean(f0)
envie<-((f0-(3/28))/(3/28))*100
par(mfrow=c(2,2))
```

```
hist(envie,main="Histograma do Enviesamento de
f(0)Hazard",xlab="Enviesamento em %",ylab="Frequência")

mean(envie)
sd(envie)
plot(dif)
hist(f0)
hist(dif)

rsme<-sqrt(var(f0)+(mean(f0)-(3/28))^2)
rsme
```

Anexo 5

Cálculo dos estimadores de f(0) pelo DISTANCE para as amostras *semi-normal*, com a selecção 1.

```
ext<-"txt"
f0<-c(1:1)
D<-c(1:1)
file<-"c:/ESTACAS/ResultadosHalfNormal.txt"

#HALF NORMAL

titulo<-"SEMI-NORMAL"
write.table(titulo,file,append=T,quote=FALSE,row.names=FALSE,col.names
=FALSE)


for(i in 1:1000)
{
system(paste(paste("C:/ESTACAS/mcds 0,
c:/ESTACAS/INPUTHF/inputhf",i,sep=""),ext,sep="."),intern=FALSE,wait=T
RUE,show.output.on.console=TRUE,invisible=TRUE)
stats<-
read.table("c:/ESTACAS/STATS.txt",dec=".",sep="",fill=TRUE,header=FALS
E)

f0[i]<-stats[8,"V6"]    #f0 para a amostra i
D[i]<-stats[39,"V6"]     # Densidade populacional
}

f0m<-as.matrix(f0)

mf0<-mean(na.exclude(f0m))
mf0
vrf0<-var(f0m)
vrf0
envie<-((na.exclude(f0m)-(3/28))/(3/28))*100
hist(envie)
envf0<-mean(envie)
envf0
sd(envie)
rmse<-sqrt(var(f0m)+(mean(f0m)-(3/28))^2)
rmse
```

```
mD<-mean(na.exclude(D))
mD
vrD<-var(D)
vrD
envieD<-((na.exclude(D)-(37.5))/(37.5))*100
envD<-mean(envieD)
envD
sd(envieD)
rmseD<-sqrt(var(D)+(mean(D)-(37.5))^2)
rmseD
nomes<-c("f0","varf0","Envf0","RMSEf0","D","varD","EnvD","RMSED")
```

Anexo 6

Cálculo dos estimadores de f(0) pelo DISTANCE para as amostras *semi-normal*, com a selecção 2 e distribuição de frequências para os modelos seleccionados.

```
ext<-"txt"
f0<-c(1:1)
model<-c(1:1)
D<-c(1:1)
for(i in 1:1000)
{
system(paste(paste("C:/ESTACAS/mcds 0,
c:/ESTACAS/INPUTHFCOMB/inputhfcomb",i,sep=""),ext,sep="."),intern=FALS
E,wait=TRUE,show.output.on.console=TRUE,invisible=TRUE)
stats<-
read.table("c:/ESTACAS/STATS.txt",dec=".",sep="",fill=TRUE,header=FALS
E)

f0[i]<-stats[8,"V6"]    #f0 para a amostra i
model[i]<-stats[15,"V6"] # Modelo Seleccionado
D[i]<-stats[39,"V6"]# Densidade populacional
}
f0

f0m<-as.matrix(f0)
mean(na.exclude(f0m))
var(f0m)
envie<-((na.exclude(f0m)-(3/28))/(3/28))*100
hist(envie)
mean(envie)
sd(envie)
rmse<-sqrt(var(f0m)+(mean(f0m)-(3/28))^2)
rmse

mean(na.exclude(D))
var(D)
envieD<-((na.exclude(D)-(37.5))/(37.5))*100

mean(envieD)
sd(envieD)
```

```
rmseD<-sqrt(var(D)+(mean(D)-(37.5))^2)
rmseD

h<-hist(model,breaks=c(0,1,2,3,4))
h
barplot((h$counts),axisnames=T,names.arg=c("Uniforme","Semi-
normal","Exp
Negativa","Hazard"),ylim=c(0,500),legend=h$counts,col=c(1,2,3,4),
main="Modelos Seleccionados Semi-normal")
```