UNIVERSIDADE DE ÉVORA

Mestrado em Matemática Aplicada - Biénio 1999/2001

Estimação da Densidade Populacional em Amostragem por Transectos Lineares com Recurso ao Modelo Logspline

Uma dissertação apresentada por: **Luís Manuel Pinto Rendas**

para a obtenção do grau de Mestre em Matemática Aplicada

Orientador: Prof. Doutor Russell Gerardo Alpizar-Jara

Esta dissertação não inclui as críticas e sugestões feitas pelo júri.

UNIVERSIDADE DE ÉVORA

Mestrado em Matemática Aplicada - Biénio 1999/2001

Estimação da Densidade Populacional em Amostragem por Transectos Lineares com Recurso ao Modelo Logspline

Uma dissertação apresentada por: **Luís Manuel Pinto Rendas**

para a obtenção do grau de Mestre em Matemática Aplicada

Orientador: Prof. Doutor Russell Gerardo Alpizar-Jara

Esta dissertação não inclui as críticas e sugestões feitas pelo júri.

Conteúdo

Glossário de Termos e Abreviaturas

Parâmetros, funções e abreviaturas:

D	...	densidade populacional (número de objectos por unidade de área)
N	...	tamanho da população na área em estudo ($N = D \times a$)
P_a	...	probabilidade de detectar objectos na área de estudo
$g(x)$	...	probabilidade de detecção de um objecto, sabendo-se que se encontra a uma distância x, perpendicular à linha do transecto;
$g(0)$	...	probabilidade de detectar um objecto sabendo-se que está sobre a linha
$f(y)$	...	função densidade de probabilidade das distâncias
$f(0)$	...	valor da função densidade de probabilidade sobre a linha
μ	...	faixa efectiva de amostragem
$f_s(y)$		estimador *logspline* da função densidade
$p[.]$	...	probabilidade
$s^k(x)$	...	função *spline* de ordem k
$B_i^k(x)$	...	B-*spline* de ordem k
G	...	espaço das funções $s^k(x)$ de classe C^2
$\phi(y, \theta)$	...	família das funções *logspline*
l	...	função de máxima verosimilhança
H	...	matriz hessiana
S	...	função score
h	...	função *logspline* obtida a partir da amostra $-x_n, ..., -x_1, x_1, ..., x_n$
$f.d.p.$	...	função densidade de probabilidade
$A.I.C.$	...	*Akaike's information criterion*
$I.C.$	...	intervalo de confiança
σ	...	desvio-padrão
Var	...	variância

Ep ... erro padrão
EQM ... erro quadrático médio
$E(.)$ ... esperança matemática
v ... viés
$(\hat{\cdots})$ ... estimador
A ... intervalo $]0, \delta[$, $\delta > 0$

Dados e constantes conhecidas:

k ... número de linhas do transecto
l_j ... comprimento da j-ésima linha, $j = 1, ..., k$
L ... comprimento total da linha ($L = \sum l_j$)
n ... número total de objectos observados
x_i ... distância perpendicular do i-ésimo objecto detectado à linha, $i = 1, ..., n$
r ... distância entre o observador e o objecto
w ... distância máxima de observação em relação ao transecto ou distância de truncatura
a ... área de estudo ($a = 2wL$)

Resumo

Apresentamos abordagem sobre a evolução histórica da amostragem por transectos lineares e desenvolve-se a teoria que lhe é subjacente.

Descrevemos a metodologia de estimação em amostragem por transectos lineares mais utilizada actualmente que foi proposta por Buckland (1992a). São discutidos os seus aspectos mais relevantes no que respeita aos pressupostos utilizados e à escolha de uma função de detecção adequada.

Fazemos a revisão de uma teoria recente denominada *logspline density estimation,* desenvolvida por Koo & Stone (1986a e 1986b) e Stone (1990), que permite estimar o logaritmo de uma função densidade de probabilidade utilizando-se *splines* cúbicos, estimação pelo método da máxima verosimilhança e adição e delecção de nós seleccionados pelas estatísticas de Rao e Wald, respectivamente.

Fazemos uma pequena adaptação que permite aplicar esta teoria ao cálculo do estimador da probabilidade de detecção $f(0)$, no contexto dos transectos lineares, e consequentemente estimar a densidade populacional de animais.

São analisados dois exemplos práticos: os dados das estacas de madeira e dos ungulados africanos descritos e estudados por Burnham *et al.* (1980). Comparamos os resultados obtidos utilizando a metodologia *logspline* com a utilizada no programa DISTANCE.

Avaliamos a metodologia das *logsplines* aplicadas aos transectos lineares através de um conjunto de simulações de populações animais, utilizando-se seis funções de detecção e seis diferentes dimensões de populações. Os cálculos foram efectuados através dos programas DISTANCE e POLSPLINE e desenvolvemos pequenos programas que permitiram gerar e formatar os dados, calcular as medidas utilizadas e gerar amostragens por bootstrap para calcular os intervalos de confiança, no caso da estimação por *logsplines.*

Discutimos os resultados obtidos e apontamos perspectivas de desenvolvimento futuro.

Palavras chave: transectos lineares, funções de detecção, estimação por *logsplines*, método da máxima verosimilhança.

Abstract

We present a brief historical note of line transect sampling and its underlying theory.

We describe the method, which is commonly used at the present time, proposed by Buckland (1992a). The most relevant features of the line transect methodology are discussed in terms of assumptions used and the choice of an adequate detection function.

We review a recent theory called *logspline* density estimation, developed by Koo & Stone (1986a e 1986b) and Stone (1990) which allows to estimate the logarithm of density probability function using cubic *splines*, maximum likelihood estimation and addition and deletion of knots selected by the Rao and Wald statistics.

We made a slight adjustment that allows us to apply this theory to the value of $f(0)$ estimator in the area of line transects and therefore to estimate the population density.

Here two practical examples are presented: the wooden stake data and the African ungulate data analised by Burnham *et al.* (1980).

We compare the results obtained by using the *logspline* method with the one used by program DISTANCE .

The *logspline* method applied to line transects is evaluated through a set of simulation scenarios, six detection functions and six different population dimensions. Are used Programs DISTANCE and POLSPLINE were used and small programs were developed. These enabled us to measure and generate samples by bootstrap to calculate the confidence intervals in *logspline* density estimation.

Finally the results obtained are discussed and we point out perspectives of future development.

Key-words: line transects, detection functions, *logspline* density estimation, maximum likelihood method.

Agradecimentos

Este trabalho não teria sido possível sem a ajuda de várias pessoas a quem sinto a necessidade de prestar o meu agradecimento.

Ao Doutor Russell Alpizar-Jara por ter aceite a orientação do trabalho e também pela sua disponibilidade, entusiamo, rigor e amizade demonstrados ao longo destes dois anos.

Às amigas Elsa Amaro e Ana Bela Santos pela ajuda imprescindível que me prestaram. Sem elas teria sido muito difícil levar a bom termo este trabalho.

Ao Comandante Ferreira da Silva pela sua compreeensão, disponibilidade e incentivo.

Aos investigadores Dr. Ronaldo Dias, Dr. Lucio Barabesi, Dr. Charles Kooperberg e Dr. Len Thomas pela sua disponibilidade em ajudar-nos neste trabalho.

Ao Mestre Tiago Marques pela sua disponibilidade e ajuda.

À professora Zulmira Andrade pelo seu auxílio nas pequenas peculiaridades da língua de Camões.

Ao Prof. Dr. Miguel Moreira pela sua ajuda nos meandros do LATEX.

Finalmente, os mais importantes. Dedico à minha família, em particular , à Elsa, Inês e Laura este trabalho. Sem elas, este não existiria nem teria sentido.

Capítulo 1

Introdução

1.1 Amostragem por distâncias

Nas últimas décadas, a par de um desenvolvimento sem precedentes das ciências biológicas, foi-se também assistindo a uma degradação acentuada do meio ambiente e dos ecossistemas com graves consequências para a diversidade, quantidade e qualidade das espécies selvagens existentes no planeta.

A Ecologia assume uma importância crescente na actualidade. A gestão, o desenvolvimento e a manutenção dos recursos naturais tornam-se vitais na sociedade, tanto a nível económico como social e também a longo prazo, uma questão de sobrevivência para o planeta e consequentemente, para o homem.

O estudo destas problemáticas passa sempre por uma análise detalhada do estado dos ecossistemas. Um dos aspectos mais básicos a ser estudado, mas não menos importante, é o tamanho das populações ou a respectiva densidade populacional[1].

Dado ser difícil e por vezes, impossível, uma contagem exaustiva das populações naturais, devido a aspectos físicos (como contar os peixes num lago?) ou aspectos económicos (seria extremamente dispendioso contar todas as perdizes existentes no Alentejo) utilizam-se técnicas estatísticas que permitem estimar com margens de erro aceitáveis o tamanho dessas populações que servem como indicadores para monitorizar o estado dos ecossistemas.

Os mais utilizados actualmente são os métodos de captura-recaptura e de amostragem por distâncias. Os primeiros não serão tratados neste trabalho. Para se ter uma ideia geral sobre o mesmo, ver por exemplo Pollock *et al* (1990). Relativamente ao segundo trata-se de um conjunto de técnicas para estimar a

[1] Entede-se por densidade populacional o número de indivíduos por unidade de área

abundância de populações baseadas nas distâncias a que os objectos se encontram de uma linha (transectos lineares) ou ponto (transectos pontuais).

Por razões de espaço e conveniência prática, os métodos de estimação apresentados neste trabalho estarão baseados nos transectos lineares mas são igualmente válidos para os transectos pontuais, com os correspondentes ajustes.

A amostragem por distâncias teve um grande desenvolvimento nas últimas três décadas aplicando-se a uma grande diversidade de espécies e habitats (*e.g.* Buckland *et al.*, 2001). Embora o seu aparecimento date do início do século XX (Forbes, 1907; Forbes and Gross, 1921 e Nice and Nice, 1921) ainda eram técnicas bastante empíricas.

Só a partir do fim da década de 60 surgiram os primeiros estudos científicos com fundamento teórico:

- Gates *et al.* (1968) fez a primeira abordagem rigorosa, utilizando a estimação por máxima verosimilhança e a distribuição exponencial negativa como curva de detecção, sendo hoje em dia já pouco utilizada.
- Seber (1973) apresentou um conjunto de modelos paramétricos mais plausíveis, tendo sido os seus fundamentos teóricos aprofundados por Ramsey (1979).
- Burnham & Anderson (1976) deram-nos o trabalho fundamental para a forma como a técnica é utilizada actualmente.
- Burnham *et al.* (1980) escreveram uma obra de referência, utilizando um modelo com séries de Fourier e o software TRANSECT para os cálculos, o que facilitou a utililização do método por um conjunto alargado de investigadores.

A partir daí, começou a ser publicada uma grande quantidade de artigos dos quais se destacam Buckland (1985), Hayes & Buckland (1983), Buckland (1992a) pelo facto de que originaram a teoria mais utilizada actualmente nesta área.

Comparando os vários modelos existentes, Buckland (1992a) sugeriu que um modelo mais robusto (ver secção 2.4) resultava da combinação de uma função chave (modelo paramétrico) ajustada com termos polinomiais (modelo não paramétrico).

Com o livro Buckland *et al.* (1993) foi desenvolvido, em paralelo, o software DISTANCE, baseado originalmente no programa TRANSECT, constituindo estes a base para a maioria das aplicações práticas em que se utiliza amostragem por distâncias. A actualização deste livro feita por Buckland *et al.* (2001) destacando-se um novo capítulo sobre o desenho e métodos de campo bem como a análise de novas metodologias para a detecção de animais.

1.2 Motivação

O trabalho apresentado aborda, fundamentalmente, dois temas: os transectos lineares e a estimação da função densidade de detecção por *logsplines*.

Esta escolha deve-se a vários factores. Por um lado, é uma forma de aplicar os conhecimentos base em Matemática Pura aos que foram obtidos na parte curricular do mestrado na área da Estatística e Modelos Biológicos. Por outro lado, pretende divulgar um pouco no nosso país os temas supracitados já que são poucas as investigações e informação sobre eles, embora tenha surgido recentemente alguma informação sobre amostragem por distâncias no âmbito de cadeiras de mestrado ministradas na Universidade de Évora, na Faculdade de Ciências da Universidade de Lisboa e no Instituto Superior de Agronomia na Universidade Técnica de Lisboa. A tese de mestrado de Marques (2002) na área da amostragem por distâncias, é também uma referência nesta área, em língua portuguesa.

Também foi importante o facto de não ser do nosso conhecimento a existência, na literatura, de aplicações da teoria de *logsplines* aos transectos lineares, o que constituiu por si só, um desafio e uma motivação suplementares para este trabalho.

Finalmente, terá sido importante o interesse do autor e do orientador pelos assuntos relacionados com a Ecologia e a preservação das espécies, para os quais este trabalho pretende ser um pequeno contributo.

1.3 Estrutura da tese

O objectivo deste trabalho consiste na aplicação da estimação de densidades por *logsplines* ao modelo de transectos lineares. Após esta breve introdução apresenta-se no capítulo 2 os desenvolvimentos teóricos essenciais relativos a amostragem por distâncias e em particular ao modelo de transectos lineares. Será dada uma importância relevante à estimação das funções de detecção, suporte fundamental do nosso trabalho.

No capítulo 3, são explicados com algum detalhe, os aspectos necessários ao desenvolvimento da teoria de *logsplines*. No capítulo 4 descreve-se a forma de aplicar a teoria das *logsplines* à estimação de $f(0)$ no contexto dos transectos lineares. Apresenta-se ainda o software que se irá utilizar e aplica-se a metodologia a dois exemplos práticos.

No capítulo 5 procedemos a um conjunto alargado de simulações com os objectivos de, por um lado, comparar a metodologia *logspline* com a de Buckland(1992a) e, por outro, avaliar a sua *perfomance*.

No último capítulo discutiremos a adequação do método e apresentaremos as conclusões do trabalho, apontando alguns tópicos para futuras investigações nesta área.

Capítulo 2

Amostragem por transectos lineares

O presente trabalho irá apenas incidir sobre o modelo de transectos lineares que tem por origem uma ideia muito simples: são traçados aleatoriamente transectos sobre a área a observar, que depois serão percorridos por um ou mais indivíduos que irão detectar os objectos de estudo medindo as distâncias a que estes se encontram da linha. Esta medição é feita, geralmente, utilizando-se a distância de observação r e o ângulo de observação θ para se obter x, a distância perpendicular à linha do transecto percorrido, de forma que $x = rsen(\theta)$. A figura 2.1 ilustra este procedimento.

Se considerarmos que o comprimento total dos transectos é L e que a distância máxima a que se observam n objectos é w, o número esperado de objectos por unidade de área poderia ser estimado por

$$\widehat{D} = \frac{n}{2wL}. \tag{2.1}$$

Mas este estimador é enviesado porque assume que todos os objectos na área de estudo são observados com probabilidade igual a 1. No entanto, apenas uma proporção dos objectos é detectada. Se denotarmos por P_a essa proporção, e esta puder ser estimada através dos dados por $\widehat{P}_a$, podemos escrever que:

$$\widehat{D} = \frac{n}{2wL\widehat{P}_a}. \tag{2.2}$$

Nas secções seguintes explicitaremos a forma e significado de $\widehat{P}_a$ bem como a teoria que lhe está subjacente.

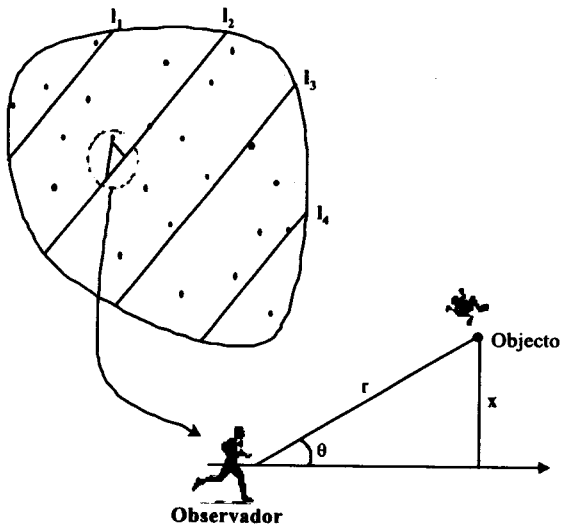

Figura 2.1: Os objectos são detectados ao longo de cada um dos transectos. Para cada um deles é medida a sua distância perpendicular ao transecto, geralmente obtida através da relação, $x = r \sin \theta$.

2.1 Pressupostos

A teoria relativa à amostragem por distâncias, neste caso específico, aos transectos lineares, assenta num conjunto de pressupostos. Os mais importantes são (*e.g.* Seber, 1973 e Buckland *et al.*, 2001):

1. Os objectos que estão sobre a linha do transecto são todos detectados.($g(0)$[1] = 1).

   Embora pareça uma situação bastante óbvia, se pensarmos na detecção de animais por via aérea ou marítima, apercebemo-nos que não será tão fácil detectar todos os animais ou objectos que se encontram sobre a linha do transecto. A quebra deste pressuposto acarreta consequências graves para a qualidade da estimação de D como veremos mais adiante. Actualmente, uma parte significativa da investigação nesta área desenvolve métodos que permitem a obtenção de estimadores mesmo quando $g(0) < 1$. (*e.g.* Alpizar-Jara, 1997; Borchers *et al.*, 1998 e Chen, 2000) .

[1] Representa a probabilidade de detecção de individuos que estão sobre a linha do transecto percorrido.

2. Não há movimento dos objectos provocado pela observação nem são observados mais de uma vez.

   A quebra deste pressuposto pode ser a causa principal de enviesamento nalguns estudos de amostragem por distâncias, essencialmente naqueles que realizam percursos a pé. Para além disso torna-se evidente que a aproximação dos objectos ao observador provoca uma sobrestimação da densidade enquanto que o afastamento causa uma subestimação. Uma das formas de resolver o problema é com a utilização de mais do que uma plataforma de observação que permitirá uma melhor estimativa da probabilidade de se avistarem os objectos antes de se movimentarem. (e.g. Buckland & Turnock ,1992).

3. As observações são eventos independentes.

   Embora seja importante, já que permite que a função de verosimilhança seja "tratável" em termos matemáticos, é muitas vezes violado (Buckland *et al.*, 2001). As principais causas da quebra deste pressuposto prendem-se com o facto dos animais poderem estar em grupos, o que facilmente se resolve considerando os próprios grupos como uma unidade de observação (geralmente o seu centro geométrico) existindo literatura extensa sobre o assunto (*e. g.* Ramsey *et al.*, 1987; Quinn, 1985 e Quang, 1991).

4. As distâncias são medidas sem erros.

   Segundo Buckland *et al* (1993), os erros são desprezáveis desde que sejam pequenos e com média zero. Actualmente, a tecnologia existente, nomeadamente o G.P.S. (global position system) permite medições com um grau de exactidão bastante razoável. No entanto, continua a ser investigado o problema e a desenvolverem-se técnicas no sentido de se minimizarem os efeitos dos erros nas estimativas (*e.g.* Alpizar-Jara, 1997; Chen, 1998 e Marques, 2002).

Para alguns autores, deveriam ser considerados outros pressupostos, enquanto outros põem frequentemente em causa a existência de alguns deles. Por exemplo, Seber (1973) considera sete pressupostos e Burnham *et al* (1980) consideram quatro pressupostos e formalizam duas situações físicas abstractas que lhes permitam o desenvolvimento da teoria. A título de exemplo refira-se que, por vezes, também são considerados os seguintes:

- a curva de detecção tem derivada nula em zero, o que na literatura anglo-saxónica se denomina por *shoulder condition*, permitindo um "comportamento suave" da curva de detecção perto da origem. Este pressuposto não deixa de ser uma conveniência matemática uma vez que esta condição só facilita a estimação de $f(0)$, permitindo uma maior estabilidade nos estimadores por máxima verosimilhança. A metodologia proposta por este trabalho revelou que a utilização de *logsplines* fornece boas estimativas quando esta condição não se cumpre.

- Os transectos são colocados aleatoriamente;

- Os objectos nunca são contados mais do que uma vez, numa mesma sessão de amostragem.

2.2 Estimação da densidade populacional

A teoria geral que se vai apresentar baseia-se, essencialmente, nos trabalhos realizados por Gates *et al.* (1968) e Seber (1982).

Faremos, seguidamente, uma outra abordagem através de uma interpretação gráfica do conceito de função de detecção.

Tinhamos visto anteriormente que um estimador da densidade populacional é dado por

$$\widehat{D} = \frac{n}{2wL\widehat{P}_a}. \tag{2.3}$$

Vejamos como estimar a probabilidade de detectar objectos na área de estudo, $\widehat{P}_a$. Consideremos um rectângulo de comprimento L e largura dx. Então a probabilidade de um objecto estar em $(x, x+dx)$ é dada por

$$p[\text{objecto estar em } (x, x+dx)] = \frac{2Ldx}{a} \tag{2.4}$$

e, por definição de função detecção,

$$p[\text{objecto ser observado} \mid \text{objecto está em } (x, x+dx)] = g(x). \tag{2.5}$$

Utilizando a definição de probabilidade conjunta,

$$p[\text{objecto ser observado em } (x, x+dx)] = \frac{2Lg(x)dx}{a} \tag{2.6}$$

podemos então definir a probabilidade de se observar um objecto ao longo do transecto como

$$P_a = \frac{2L}{a}\int_0^w g(x)dx \tag{2.7}$$

em que w representa a distância máxima de observação em relação ao transecto ou então uma distância de truncatura.

Se considerarmos

$$\mu = \int_0^w g(x)dx \tag{2.8}$$

obtemos a expressão

$$P_a = \frac{2L\mu}{a}. \tag{2.9}$$

Atendendo novamente à definição de probabilidade condicionada,

$$\begin{aligned} f(x)dx &= p[\text{objecto estar em } (x, x+dx) \mid \text{objecto é observado}] = \\ &= \frac{p[\text{objecto ser observado em } (x, x+dx)]}{p[\text{objecto ser observado}]} = \\ &= \frac{\frac{2Lg(x)dx}{a}}{P_a} = \frac{g(x)}{\mu}dx. \end{aligned} \tag{2.10}$$

Logo, a função densidade de probabilidade é dada por

$$f(x) = \frac{g(x)}{\mu}. \tag{2.11}$$

Repare-se que, atendendo ao significado de μ, $f(x)$ e $g(x)$ apresentam as mesmas características, sendo a equação anterior uma mudança de escala de modo a garantir que

$$\int_0^w f(x)dx = 1.$$

Como, pelo *pressuposto 1*, $g(0) = 1$, teremos,

$$f(0) = \frac{g(0)}{\mu} = \frac{1}{\mu} \tag{2.12}$$

ou seja,

$$\mu = \frac{1}{f(0)}. \tag{2.13}$$

Voltando à expressão inicial e substituindo pelos resultados obtidos anteriormente ficamos com:

$$\widehat{D} = \frac{n}{2wL\widehat{P}_a} = \frac{n}{2wL\frac{2L}{\widehat{af}(0)}} = \frac{n\widehat{f}(0)}{2L}. \tag{2.14}$$

Deste modo, um estimador da densidade populacional é definido pelo estimador da f.d.p. da detecção de individuos sobre a linha do transecto percorrido.

Vejamos agora uma forma mais intuitiva de obtermos o estimador da densidade.

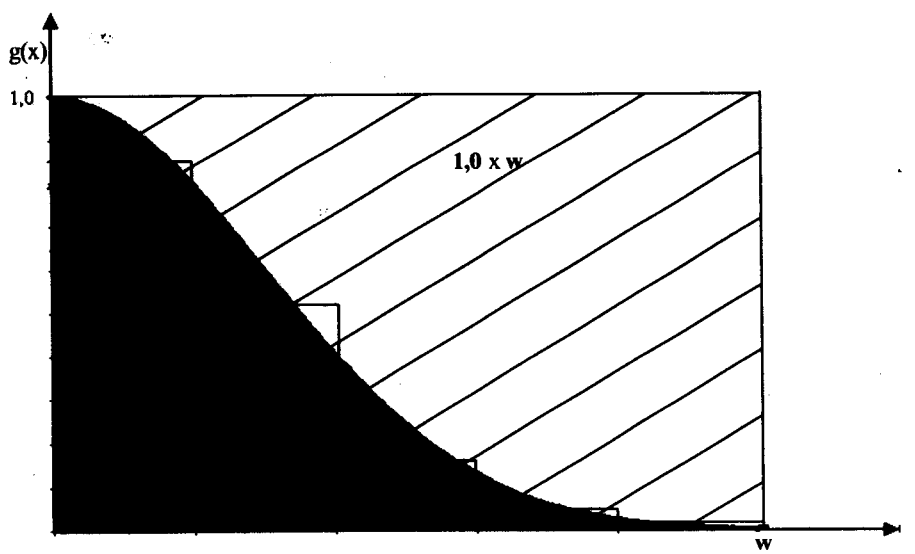

Figura 2.2: Sendo a área a tracejado igual a w, e a área a sombreado igual a μ deduz-se que a probabilidade de detectar um objecto dado que se enconta na área observada é $\frac{\mu}{w}$.

Considerando o histograma representativo das distâncias observadas (figura 2.2) e tendo sido feita uma mudança de escala conveniente, de forma que a curva corresponda à probabilidade de detecção como função das distâncias x a que os objectos se encontram do transecto, ou seja, $g(x)$, obtém-se uma área determinada por

$$\mu = \int_0^w g(x)dx \tag{2.15}$$

Atendendo também ao facto de que, se avistássemos todos os objectos no transecto, a sua probabilidade de detecção seria, evidentemente igual a 1, obter-se-ia uma área sombreada determinada por

$$\mu = 1 \times w \tag{2.16}$$

como se encontra ilustrado na figura 2.2.

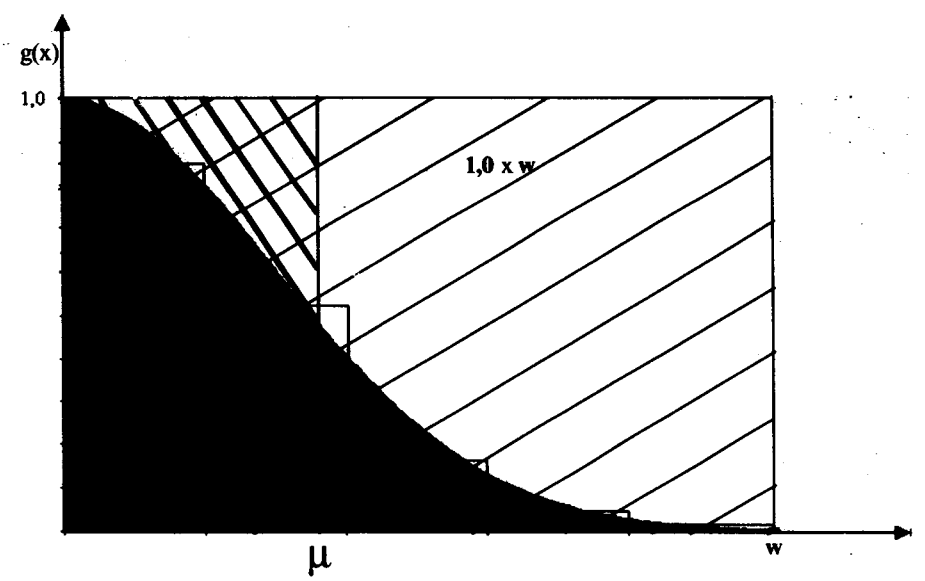

Figura 2.3: A área do rectângulo de comprimento 1 e largura μ é igual à área sombreada, podendo μ ser interpretado como uma faixa efectiva de amostragem.

Atendendo ao que se ilustra na figura 2.3, repare-se que existe um rectângulo com largura μ e comprimento 1, ou seja, com área igual a μ.

Logo, a probabilidade de detectar um objecto dado que se encontra a uma distância inferior ou igual a w, do transecto é dada por

$$P_a = \frac{\mu}{w} = \frac{1}{f(0)w}; \tag{2.17}$$

ou seja,

$$\widehat{D} = \frac{n\widehat{f}(0)}{2L}. \tag{2.18}$$

Este facto permite também interpretar o valor de μ como uma faixa efectiva de amostragem, ou seja, se fossem avistados todos os objectos até uma distância μ, estaríamos, na prática, a detectar em média o mesmo número de animais que o total observado.

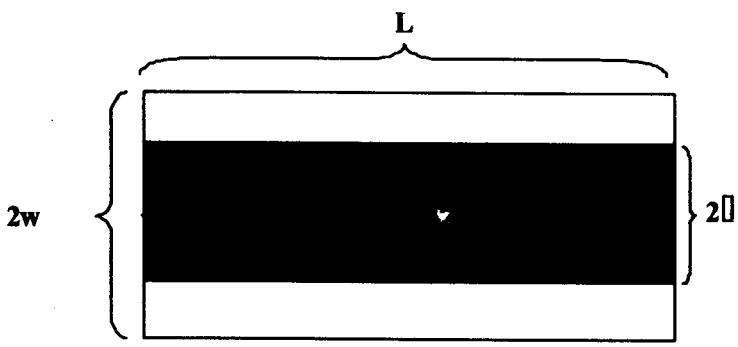

Figura 2.4: Esquema relativo a uma área a estudar. A zona a sombreado representa a região em que, em média, se avistariam todos os animais.

Outra forma interessante de analisarmos a importância de μ, é repararmos na sua relação com a, a área total observada. Considerando o rectângulo de largura $2w$ e comprimento L com área a, a medida 2μ corresponderia à largura de um rectângulo com o mesmo comprimento, onde teríamos avistado em média todos os objectos de estudo que aí se encontram, como se ilustra na figura 2.4.

2.3 Variância das estimativas e intervalos de confiança

A variância do estimador da densidade pode ser estimada, fundamentalmente por dois processos: através de uma expansão em série de Taylor (e.g. Casella &

Berger; 2002) ou por intermédio de processos de reamostragem. Relativamente ao primeiro, assumindo a não existência de correlação entre as duas componentes da estimação, n e $f(0)$, obtém-se

$$\widehat{Var}(\widehat{D}) = \widehat{D}^2 \left\{ \frac{\widehat{Var}(n)}{n^2} + \frac{\widehat{Var}\left[\hat{f}(0)\right]}{\left[\hat{f}(0)\right]^2} \right\} \tag{2.19}$$

ou, de forma equivalente, utilizando o coeficiente de variação,

$$\widehat{Var}(\widehat{D}) = \widehat{D}^2 \left\{ [cv(n)]^2 + \left[cv(\hat{f}(0))\right]^2 \right\} \tag{2.20}$$

O cálculo da variância de n depende do número de réplicas dos transectos. Se for realizado apenas um transecto, a variância terá de ser estimada a partir de um modelo assumido para a distribuição espacial dos objectos. Como, geralmente, se trata de uma distribuição de Poisson ter-se-á,

$$\widehat{Var}(n) = \widehat{E}(n) = n. \tag{2.21}$$

No caso de se terem utilizado k transectos obtém-se a variância por

$$\widehat{Var}(n) = \frac{L \sum_{i=1}^{k} l_i \left(\frac{n_i}{l_i} - \frac{n}{L}\right)^2}{k-1} \tag{2.22}$$

em que n_i é o número de objectos detectados em cada linha de comprimento l_i.

A variância do estimador de $f(0)$ é geralmente obtida através da estimativa da matriz de informação de Fisher (ver a expressão 2.33). Uma descrição mais pormenorizada pode ser encontrada em Buckland *et al* (2001).

Obtendo-se $\widehat{D}$ e $\widehat{Var}(\widehat{D})$, um intervalo de confiança a $100(1-2\alpha)\%$ é dado por

$$\widehat{D} \pm z_\alpha \sqrt{\widehat{Var}(\widehat{D})} \tag{2.23}$$

onde z_α é obtido através da distribuição normal reduzida.

Segundo Burnham *et al* (1980), como a distribuição de $\widehat{D}$ é positivamente enviesada, obtém-se um intervalo com melhor cobertura assumindo que $\widehat{D}$ tem distribuição log-normal. Então, o intervalo de confiança a $100(1-2\alpha)\%$ será dado por

$$\left]\frac{\widehat{D}}{C}, \widehat{D} \times C\right[\tag{2.24}$$

onde

$$C = \exp\left\{ z_\alpha \sqrt{\ln\left[1 + \frac{\widehat{Var}(\hat{D})}{\hat{D}^2}\right]} \right\}. \tag{2.25}$$

Relativamente aos processos de reamostragem para obter estimativas da variância, são referidos na literatura os métodos *jackknife* e *bootstrap*. Na amostragem por distâncias utiliza-se com maior frequência o *bootstrap* não paramétrico (*e.g.* Davison *et al.*, 1997). Encontra-se uma boa descrição destes métodos em Buckland *et al* (2001: pág 83).

2.4 Função de detecção e função densidade das distâncias

Como vimos o problema estatístico relativo à estimação da densidade populacional resume-se à estimativa de $f(0)$, o que requer um cuidado muito especial com a estimação e modelagem de $g(x)$.

Vários autores sugeriram uma abordagem paramétrica para a estimação de $f(0)$ apresentando um conjunto de modelos plausíveis, como por exemplo, "the power law model" e o modelo logístico reverso (Eberhardt, 1978), "exponential power series model" (Pollock, 1978), o modelo Beta e exponencial quadrático (Burnham *et al*, 1980) e o "hazard rate model" (Hayes and Buckland, 1983).

Por outro lado, no contexto dos transectos lineares, a abordagem não paramétrica foi introduzida por duas vias: a utilização de séries e os métodos de *Kernel (kernel density estimation)*. Crain *et al.* (1979) proposeram um modelo baseado em séries de Fourier e Buckland (1985) introduziu os polinómios hermíticos enquanto que Quang (1991) e Chen (1995) aplicaram os métodos de *Kernel* aos transectos pontuais e lineares, respectivamente.

No entanto, ambos os modelos revelaram algumas fraquezas. Os modelos paramétricos permitem uma grande precisão quando o modelo seleccionado é o melhor para a modelação dos dados observados, mas revela-se bastante pobre quando isso não acontece. Inversamente, os modelos não paramétricos oferecem boas aproximações às verdadeiras densidades mas não têm a eficiência do modelo paramétrico quando este é o o mais apropriado.

Uma forma de contornar o problema é considerar um modelo semi-paramétrico para a estimação de $f(0)$. Buckland (1992a) propôs um procedimento geral semi-paramétrico onde apresentou um modelo chave normal truncado ajustado pelos

primeiros termos de uma série polinomial de Hermite. Barabesi (2000) apresenta um modelo semi-paramétrico baseado na estimação da densidade por uma função de verosimilhança local.

Devido ao seu maior desenvolvimento, em termos práticos e teóricos, vamos utilizar neste trabalho o método de Buckland *et al* (2001).

Neste método selecciona-se uma função paramétrica $\gamma(x)$ que melhor se ajusta aos dados, e caso o ajustamento seja pouco perfeito , são adicionados alguns termos de uma série não paramétrica, ou seja, a função densidade de probabilidade é definida por,

$$f(x) = \frac{\gamma(x)}{\lambda}\left[1 + \sum_{j=1}^{m} a_j p_j(x_s)\right], \tag{2.26}$$

sendo $\gamma(x)$, a denominada *função-chave,* uma função paramétrica com k parâmetros ;

$$p_j(x_s) = \begin{cases} x_s^j, \text{se for um polinómio} \\ \hbar_j(x_s) \text{ se for um polinómio hermítico} \\ \cos(j\pi x_s) \text{ se for uma série de cosenos} \end{cases} ; \tag{2.27}$$

x_s é um valor standartizado de x (que visa evitar problemas de convergência aquando dos cálculos numéricos);

$a_j = 0$ se o termo de ordem j de $p_j(x_s)$ não for utilizado no modelo, ou então é estimado pelo método da máxima verosimilhança;

λ é uma função normalizada dos parâmetros.

Vejamos na tabela 2.1 algumas combinações possíveis de utilizar no Programa DISTANCE (ver secção 4.2):

Tabela 2.1 - Algumas combinações utilizadas por DISTANCE.

Função-Chave	Série	Função $f(x)$
Uniforme	Coseno	$\frac{1}{w}\left[1+\sum_{j=1}^{m} a_j \cos\left(\frac{j\pi x}{w}\right)\right]$
Uniforme	Polinómio	$\frac{1}{w}\left[1+\sum_{j=1}^{m} a_j \left(\frac{x}{w}\right)^{2j}\right]$
Normal	Coseno	$e^{-\frac{x^2}{2\sigma^2}}\left[1+\sum_{j=2}^{m} a_j \cos\left(\frac{j\pi x}{w}\right)\right]$
Normal	Hermite	$e^{-\frac{x^2}{2\sigma^2}}\left[1+\sum_{j=2}^{m} a_j h_j(x_s)\right]$, onde $x_s=\frac{x}{\sigma}$
"Hazard rate"	Coseno	$\left(1-e^{-\left(\frac{x}{\sigma}\right)^{-b}}\right)\left[1+\sum_{j=2}^{m} a_j \cos\left(\frac{j\pi x}{w}\right)\right]$

2.5 Método da máxima verosimilhança

Baseados em Buckland *et al* (2001), vamos fazer uma breve descrição desse método para dados não agrupados.

Consideremos a função de verosimilhança para as distâncias detectadas $x_1, x_2, ..., x_n$, condicionadas em n:

$$L(\underline{\theta}) = \prod_{i=1}^{n} f(x_i) \tag{2.28}$$

onde x_i é a i-ésima observação e $\underline{\theta}$ é um vector com $k+m$ componentes em que $\theta_1, \theta_2, ..., \theta_k$ são os parâmetros da *função-chave* e $\theta_{k+1}, \theta_{k+2}, ..., \theta_{k+m}$ são os parâmetros (coeficientes) dos termos ajustados. Tem-se, então,

$$l = \ln\left[L(\underline{\theta})\right] = \sum_{i=1}^{n} \ln\left[f(x_i)\right] = \sum_{i=1}^{n} \ln\left[f(x_i)\lambda\right] - n\ln\lambda \tag{2.29}$$

derivando em ordem a θ_j, obtém-se,

$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{n} \frac{\partial \ln [f(x_i)]}{\partial \theta_j} =$$

$$= \sum_{i=1}^{n} \left\{ \frac{1}{f(x_i).\lambda} \times \frac{\partial [f(x_i)\lambda]}{\partial \theta_j} \right\} - \frac{n}{\lambda} \times \frac{\partial \lambda}{\partial \theta_j}, \; j = 1, ..., k+m \quad (2.30)$$

onde

$$\frac{\partial [f(x_i)\lambda]}{\partial \theta_j} = \begin{cases} \gamma(x_i) \left[\sum_{j'=1}^{m} a_{j'} \frac{\partial p_{j'}(x_{is})}{\partial x_{is}} \right] \frac{\partial x_{is}}{\partial \theta_j} + \\ + \left[1 + \sum_{j'=1}^{m} a_{j'} p_{j'}(x_{is}) \right] \frac{\partial \gamma(x_i)}{\partial \theta_j}, \; 1 \leq j \leq k \\ \gamma(x_i) p_{j-k}(x_s), \; j > k \end{cases} \quad (2.31)$$

$$\frac{\partial p_{j'}(x_{is})}{\partial x_{is}} = \begin{cases} j \times p_{j-1}(x_{is}), \text{ com } p_0(x_{is}) = 1 \text{ para um} \\ \qquad \text{polinómio simples ou hermítico} \\ -j\pi \sin(j\pi x_s) \text{ para o modelo de séries de Fourier} \end{cases} \quad (2.32)$$

Quando

$k = 1$ e $x_{is} = \frac{x_i}{\theta_1}$,

$$\frac{\partial x_{is}}{\partial \theta_1} = -\frac{x_i}{\theta_1^2};$$

quando

$k = 1$ e $x_{is} = \frac{y_i}{w}$,

$$\frac{\partial x_{is}}{\partial \theta_1} = 0.$$

As equações

$$\frac{\partial l}{\partial \theta_j} = 0, \; j = 1, ..., k+m$$

serão resolvidas por métodos numéricos, geralmente utilizando o método de Newton-Raphson (ver, por exemplo, Stoer, J. (1996)).

Para se efectuar uma mudança entre polinómios simples e hermíticos é necessário redefinir $p_j(x_s)$, $j = 1, ..., m$. Se a mudança for entre polinómios e séries de Fourier, $\partial p(x_s)/\partial x_s$ também terá de ser redefinida.

Se for requerida uma *função-chave* diferente das apresentadas terão de ser implementados outros métodos para calcular as derivadas parciais, nomeadamente integração numérica.

A matriz de informação de Fisher pode ser estimada a partir da matriz Hessiana $H(\underline{\hat{\theta}})$ em que cada elemento é dado por

$$H_{jh}(\underline{\hat{\theta}}) = \frac{1}{n}\left[\sum_{i=1}^{n} \frac{\partial \ln[f(x_i)]}{\partial \hat{\theta}_j} \times \frac{\partial \ln[f(x_i)]}{\partial \hat{\theta}_h}\right]. \tag{2.33}$$

2.6 Critérios de escolha do modelo

O modelo definido em(2.26), com a utilização de uma função paramétrica ajustada pela adição de um determinado número de termos de uma série, embora possua grande flexibilidade, não é de todo o único, existindo outros que poderão, eventualmente, ser mais apropriados. Põe-se, então, o problema de como iremos escolher esse modelo. A observação atenta do histograma representativo das distâncias observadas é uma das formas mais simples, mas nem por isso menos importante de se ter uma ideia geral da curva representativa da função de detecção que queremos obter.

No entanto, outros aspectos importantes não poderão ser descurados. Referimos seguidamente alguns deles. Os critérios para se obter uma estimação robusta também se encontram, com bastante detalhe, em Burnham *et al.* (1980) e Buckland *et al.* (2001).

2.6.1 Modelo robusto

Dado que a verdadeira forma da função de detecção não é conhecida, excepto quando se procede a simulações, é desejável possuirmos um modelo que se adeque ao maior número possível de formas que o gráfico da função possa ter.

Por outro lado, a probabilidade de detecção pode depender de outros factores para além das distâncias observadas, como por exemplo, o relevo do terreno, as condições climatéricas, as características dos observadores, a estação do ano, etc.

O estimador obtido deve ser "insensível" a estes factores, ou seja, deve-se obter um valor aproximado ao que se obteria em circunstâncias consideradas óptimas,

de um ponto de vista teórico. Esta particularidade é referida na literatura por *pooling robust.*

2.6.2 Coerência do modelo

A função de detecção deve, gráficamente, estar de acordo com os pressupostos que fundamentaram o desenvolvimento do modelo teórico. Por um lado, deve decrescer à medida que aumenta a distância de avistamento e, portanto, $g'(x) < 0$ para $0 < x < w$. Por outro lado, a função de detecção deverá ter, perto da origem, (intervalo que definiremos por A[2]), um valor constante e igual a 1, bem como ter um comportamento "suave" nessa zona, ou seja $g'(0) = 0$.

Este facto costuma ser designado, na literatura em língua inglesa, por *shape criterion.*

2.6.3 Eficiência

É desejável seleccionarmos um modelo que, para além de ser robusto e coerente, forneça estimações precisas (com variância reduzida). Para tal, é recomendável a utilização do método da máxima verosimilhança, dado possuir boas propriedades estatísticas, nomeadamente, uma variância assimptótica mínima. Mesmo que o modelo não permita resultados exactos, do ponto de vista matemático, por este método, com a utilização de meios computacionais obtêm-se valores aproximados com margens de erro aceitáveis.

2.6.4 Ajustamento do modelo

O ajustamento é avaliado, usualmente, com o teste do qui-quadrado, baseado nos dados observados, agrupados por classes.

Embora este teste tenha grandes limitações relativamente à escolha do modelo (e.g. Buckland et al, 2001, pág 42-44), deve ser tomado em conta, quando é rejeitada a hipótese nula, do modelo ser apropriado.

2.6.5 *Akaike's information criterion* (AIC)

Geralmente, quando o número de parâmetros do modelo aumenta, o enviesamento diminui mas aumenta a variância (*e.g.* Burnham & Anderson, 1992). Pretende-se, então, um modelo parcimonioso, ou seja, que contenha um número de parâmetros que evite o enviesamento mas que não provoque uma diminuição da precisão.

O ajustamento relativo dos diferentes modelos pode ser avaliado utlilizando o AIC, *Akaike's information criterion,* que se baseia numa relação entre o

[2] $A =]-\delta, +\delta[\cap \mathbb{R}^+ =]0, \delta[\,,\ \delta > 0$

logaritmo da função de verosimilhança (l) e a informação de Kullback-Leibler (Sakamoto *et al.*,1986), uma quantidade que mede a informação perdida quando se usa um determinado modelo para aproximar a realidade, e que permite avaliar as suas vantagens e desvantagens relativas a aumentar o número de parâmetros do modelo, procurando, desta forma um que seja parcimonioso.

O AIC é calculado através da fórmula

$$AIC = -2\ln(l) + 2c, \tag{2.34}$$

onde c é o número de parâmetros do modelo.

Considerando exclusivamente o AIC, é seleccionado o modelo que apresentar um valor mais baixo.

Capítulo 3

Estimação da função densidade por *logsplines*

Como vimos no capítulo anterior, um dos aspectos fundamentais da metodologia dos transectos lineares prende-se com a escolha adequada de uma função de detecção, que permita obter estimadores com propriedades desejáveis e ao mesmo tempo, consiga "captar" de uma forma realista a informação recolhida através das distâncias observadas. Os exemplos apresentados por Buckland *et al* (2001: pág. 324 e seguintes) são reveladores da preocupação que se deve ter, após a recolha dos dados, em escolher um modelo que se adapte bem ao histograma que os representa. Por exemplo, no caso em que os modelos apresentam valores de A.I.C. idênticos poderá ser essa escolha que permitirá ao investigador escolher o modelo mais adequado à realidade.

No entanto, a utilização, neste contexto, de uma função paramétrica ajustada pela adição de termos de uma série, não é consensual na literatura. Diversos autores têm proposto métodos alternativos. Destacam-se pelos resultados obtidos a estimação por Kernel aplicada aos transectos pontuais (Quang, 1991) e aos transectos lineares (Chen, 1995), a estimação local por máxima verosimilhança proposta por Barabesi (2000), mínimos quadrados por Barabesi *et al.* (2002) e a estimação por métodos bayesianos (Karunamuni & Quinn II, 1995).

A utilização de *splines* polinomiais é actualmente uma das técnicas mais populares na estimação de funções não paramétricas (Hansen & Kooperbeg, 2002). Será apresentada, neste capítulo uma dessas metodologias para estimar funções densidade de probabilidade (f.d.p.) denominada na literatura anglo-saxónica por *logspline density estimation* a que chamaremos estimação de densidades por *logsplines*, ou simplesmente, *logsplines*. Esta, permite, através de um conjunto de dados, obter um modelo do logaritmo de uma f.d.p. desconhecida através de *splines* cúbicos, que passaremos a descrever na secção seguinte, utilizando o método da máxima verosimilhança para estimar os respectivos coeficientes. O seu desenvolvimento deve-se a Stone & Koo (1986a e 1986b) e Stone (1990), observando-se

sucessivos melhoramentos a partir desta última data, destacando-se os trabalhos de Kooperberg & Stone (1991, 1992 e 2001) e Stone *et al* (1997). Estes melhoramentos tiveram, essencialmente, haver com a maneira de posicionar os nós, a forma de lidar com as caudas das distribuições e o tipo de dados que se poderiam utilizar nesta metodologia.

A escolha das *logsplines* deveu-se, essencialmente, ao facto de revelar, segundo Stone *et al.* (1997), uma excelente adaptação "espacial", no sentido de possuir uma grande facilidade de adaptação às diferentes distribuições, permitindo estimações precisas mesmo nos casos de funções com descontinuidades ou "picos" muito acentuados. São prova deste facto as aplicações, com sucesso, desta metodologia em diferentes áreas científicas, como por exemplo, na econometria (Dias, 2002), aplicações financeiras (Takada, 2001) e medicina (Kooperberg *et al.*, 1992).

Começaremos este capítulo por fazer uma pequena abordagem à teoria das funções *splines*, suporte fundamental da estimação de densidades por *logsplines* apresentada na secção seguinte.

3.1 O que é um *spline*

O conceito de *spline* significa, na sua forma mais simples, uma curva construída com recurso a outras curvas mais pequenas e o seu nome teve origem num instrumento que se utilizava na indústria naval e que permitia o desenho de curvas suaves passando por um determinado número de pontos, conseguindo-se assim apurar a forma dos cascos de navios (e. g. Pina , 1995). A análise desses instrumentos permitiu constatar que produziam curvas, cuja equação era um polinómio de terceiro grau, com segundas derivadas contínuas, mas descontínuas nos nós a partir da terceira ordem, facto que estabelece a principal diferença relativamente à interpolação clássica utilizando polinómios. Segundo Pina (1995), Schoenberg estabeleceu em 1946 a primeira definição rigorosa deste tipo de funções.

Seguidamente, apresentaremos algumas definições que nos levarão até ao conceito de B-*spline* (de Boor, 1978), forma de definir *splines* que facilita a sua utilização na análise numérica, sendo utilizadas na estimação por *logsplines*.

Definição 3.1 *Seja k um inteiro positivo. Um polinómio de ordem k é uma função definida por*

$$p(x) = a_0 + a_1 x + a_2 x^2 + ... + a_{k-1} x^{k-1},\ x \in \mathbb{R} \tag{3.1}$$

e $a_0, a_1, a_2, .., a_{k-1}$ são números reais. Os polinómios de ordem um, dois, três e quatro serão denominados por polinómios constantes, lineares, quadráticos e cúbicos, respectivamente.

Consideremos $m+2$ números distintos, tais que

$$t_0 < t_1 < t_2 < ... < t_m < t_{m+1} \tag{3.2}$$

Designaremos por **polinómio definido por ramos de ordem k** *com nós* t_1, t_2, ..., t_m, *num intervalo* $[t_0, t_{m+1}]$ *uma função cuja restrição a cada um dos* $m+1$ *intervalos* $[t_0, t_1[$, $[t_1, t_2[$,..., $[t_{m-1}, t_m[$, $[t_m, t_{m+1}[$ *é um polinómio de ordem menor ou igual a k.*

Definição 3.2 *Uma* ***função spline de ordem*** *k com* t_1, t_2, ..., t_m *nós distintos é um polinómio definido por ramos, de ordem k com nós* t_1, t_2, ..., t_m, $(k-2)$ *vezes contínuo e diferenciável, e será definida por:*

$$s^{k-1}(x) = a_0 + a_1 x + a_2 x^2 + ... + a_{k-1} x^{k-1} + \sum_{i=1}^{m} b_i \, (x - t_i)_+^{k-1}, \quad k \geq 2 \tag{3.3}$$

em que $(x-t_i)_+^k = \max\left(0, (x-t_i)^k\right)$ *e* $a_0, a_1, a_2, .., a_{k-1}, b_1, b_2, .., b_m$ *são constantes reais.*

Por exemplo os *splines* lineares, quadráticos e cúbicos, nos nós indicados, escrevem-se, respectivamente, por

$$\begin{aligned} s^1(x) &= a_0 + a_1 x + \sum_{i=1}^{m} b_i \, (x - t_i)_+^1 \\ s^2(x) &= a_0 + a_1 x + a_2 x^2 + \sum_{i=1}^{m} b_i \, (x - t_i)_+^2 \\ s^3(x) &= a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \sum_{i=1}^{m} b_i \, (x - t_i)_+^3 . \end{aligned} \tag{3.4}$$

Reparemos que, para descrever uma função *spline*, necessitamos de conhecer $m+k$ coeficientes. Como a família de funções *spline* de ordem k é um espaço vectorial de dimensão $m+k$, uma das bases possíveis para esse espaço é

$$\left\{1, x, x^2, ..., x^{k-1}, (x-t_1)^{k-1}, (x-t_2)^{k-1}, ..., (x-t_m)^{k-1}\right\} \tag{3.5}$$

usualmente chamada a **base de potências truncadas.**

A função *spline* e a base definidas em (3.3) e (3.5) possuem propriedades pobres, no que respeita a métodos numéricos iterativos. Uma alternativa para representar a base de um espaço de funções *spline* são as denominadas B-*splines* que definiremos seguidamente. Uma descrição detalhada destas bases e respectivas propriedades podem ser encontradas em Schumaker (1993).

Definição 3.3 *Consideremos uma sequência infinita de nós tais que,*

$$... < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < ...$$

Uma ***B-spline constante*** *é definida por*

$$B_i^1(x) = \begin{cases} 1, & t_i \leq x \leq t_{i+1} \\ 0, & \text{caso contrário} \end{cases}, \; i = 0, \; \pm 1, \; \pm 2, ... \tag{3.6}$$

Podem-se obter, recursivamente, as splines de ordem superior:

uma ***B-spline linear*** *define-se por*

$$B_i^2(x) = \frac{x - t_i}{t_{i+1} - t_i} B_i^1(x) + \frac{t_{i+2} - x}{t_{i+2} - t_{i+1}} B_{i+1}^1(x), \; i = 0, \; \pm 1, \; \pm 2, ...; \tag{3.7}$$

no geral define-se uma **B-*spline* de ordem k** *por:*

$$\begin{aligned} B_i^k(x) &= \frac{x - t_i}{t_{i+k-1} - t_i} B_i^{k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1}^{k-1}(x), \\ i &= 0, \; \pm 1, \; \pm 2, ... \\ k &= 2, 3, ... \end{aligned} \tag{3.8}$$

No intervalo $[t_0, t_{m+1}]$ *o conjunto*

$$\left\{ B_{-k+1}^k(x), \; B_{-k+2}^k(x), \; ..., \; B_m^k(x) \right\} \tag{3.9}$$

de B-splines constitui uma base do espaço vectorial de qualquer função spline de

ordem k, nesse intervalo. Logo, poderemos dizer que essa função se pode definir por

$$s(x) = \sum_{i=1}^{k+m} c_i B^k_{-k+i}(x), \quad c_i \text{ constantes reais.} \tag{3.10}$$

Apresentamos como exemplo, a estimação de uma função densidade por *logsplines*. Foram gerados, aleatoriamente, 300 dados provenientes de uma distribuiçao normal reduzida. A função estimada é

$$\begin{aligned} \widehat{y} \;=\; & \exp\left[\widehat{s}\,(x)\right] = \exp(0.765947419643013 + 1.77730509931026 * x - \\ & -1.59817412895946 * \max(0, (x + 1.66375238281057)^3) + \\ & +1.64795170558326 * \max(0, (x + 1.50900489652431)^3) - \\ & -0.0497775766238032 * \max(0, (x - 3.45936531911114)^3) \end{aligned} \tag{3.11}$$

cuja representação gráfica se encontra na figura 3.1. Reparemos que a utilização da função exponencial deve-se ao facto da estimação por *logsplines* logaritmizar a função densidade de probabilidade. Chamamos também a atenção para a "aproximação" obtida com a amostra considerada.

3.2 Estimação da função densidade por *logsplines*

Começaremos por apresentar o modelo *logspline*, bem como alguns aspectos teóricos mais relevantes relativos a estimação pelo método da máxima verosimilhança. A abordagem seguida baseia-se nos trabalhos de Kooperberg & Stone (1991) e Stone *et al* (1997).

Não serão abordados os casos da estimação quando os dados são censurados, truncados ou agrupados os quais poderão ser consultados, respectivamente, em Kooperberg & Stone (1991), Koo *et al.* (1999) e Koo & Kooperberg (2000).

Seguidamente, será apresentada a forma como são seleccionados os nós e os critérios subjacentes à escolha final do modelo. Para finalizar apresentam-se os resultados que permitem determinar a variância e os intervalos de confiança no contexto dos *logsplines*. Não vamos considerar o caso em que os extremos do intervalo em que estão contidos os nós são infinitos dado não terem interesse para o nosso estudo.

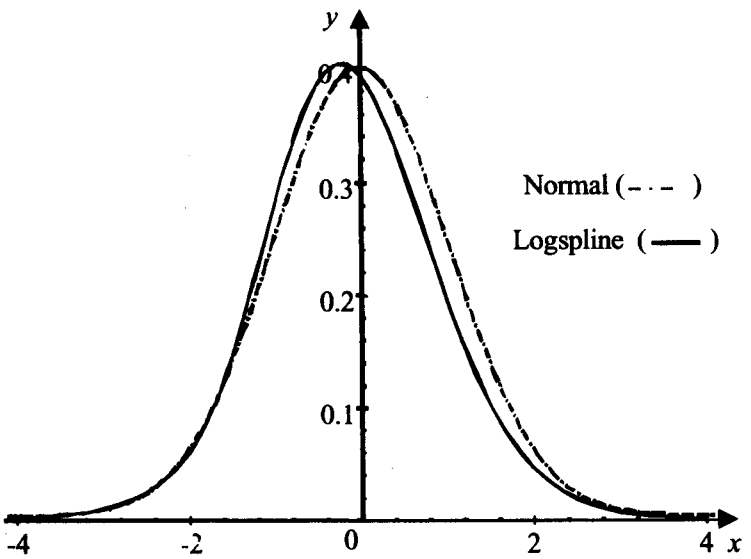

Figura 3.1: Gráficos da distribuição normal reduzida (a tracejado) e da função logspline estimada, (a cheio) $\widehat{y} = \exp(\widehat{s}(x))$, obtida através de 300 dados gerados aleatoriamente de uma distribuiçao normal reduzida.

3.2.1 O modelo *logspline*

Sejam K um inteiro maior ou igual três, L e U números reais tais que $-\infty < L < U < +\infty$ e uma sequência de nós t_1, t_2, ..., t_k que verificam

$$L < t_1 < t_2 < ... < t_k < U \tag{3.12}$$

Seja G o espaço das funções s de classe C^2, definidas no intervalo $]L, U[$ em que as restrições aos intervalos $[t_1, t_2]$,..., $[t_{k-1}, t_k]$ são funções polinomiais cúbicas e aos intervalos $]L, t_1]$ e $[t_k, U[$ são funções lineares. Repare-se que G é um espaço vectorial de dimensão K. Logo, tem uma base da forma $1, B_1, B_2, ..., B_p$, com $p = K - 1$. A escolha desta base é feita de modo que:

B_1 seja linear com declive negativo em $]L, t_1]$ e $B_2, ..., B_j$ sejam constantes nesse intervalo,

B_j seja linear com declive positivo em $[t_k, U[$ e $B_1, ..., B_{j-1}$ sejam constantes nesse intervalo.

Consideremos o vector-coluna $\theta = (\theta_1, \theta_2, ..., \theta_p)^t \in R^p$, tal que

$$\int_L^U \exp\left[\sum_{j=1}^{k-1} \theta_j B_j(x)\right] dy < \infty. \tag{3.13}$$

Seja Θ a família dos vectores-coluna anteriormente definidos. Dado $\theta \in \Theta$, defina-se:

$$C(\theta) = \log\left\{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx\right\} \tag{3.14}$$

e

$$\phi(x;\theta) = \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x) - C(\theta)\right], \; L < x < U \tag{3.15}$$

Reparemos que

$$\int_L^U \phi(x,\vartheta)\, dx = \int_L^U \frac{\exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right]}{\exp\left(\log\left\{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx\right\}\right)} dx \tag{3.16}$$

logo

$$\int_L^U \phi(x,\theta)\, dx = \frac{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx}{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx} = 1 \tag{3.17}$$

O que permite concluir que $\phi(.,\theta)$, $\theta \in \Theta$ é uma função densidade em $]L,U[$.

Esta família de funções é denominada por **família *logspline***.

Seja Y uma variável aleatória tendo uma função densidade contínua e positiva e $Y_1, Y_2, ..., Y_n$ variáveis aleatórias independentes com a mesma distribuição de Y.

A função de máxima verosimilhança, relativa à família *logspline* é dada por

$$l(\theta) = \sum_{i=1}^{n} \log\left[\phi(x_i,\theta)\right]; \; \theta \in \Theta \tag{3.18}$$

ou, de outro modo,

$$l(\theta) = \sum_{i=1}^{n}\left(\sum_{j=1}^{k-1}\theta_j B_j(x_i) - C(\theta)\right) = \sum_{i=1}^{n}\sum_{j=1}^{k-1}\theta_j B_j(x_i) - nC(\theta). \tag{3.19}$$

O estimador de máxima verosimilhança $\widehat{\theta}$ obtém-se maximizando o logaritmo da função de verosimilhança.

Seja E a matriz $(n \times k)$ com entradas $B_{j-1}(X_i)$ na linha i e coluna j para $1 \leq i \leq n$ e $2 \leq j \leq K$ e entrada 1 em todas as linhas da coluna 1:

$$E = \begin{bmatrix} 1 & B_1X_1) & B_2(X_1) & \ldots & \ldots & B_K(X_1) \\ 1 & B_1(X_2) & B_2(X_2) & \ldots & \ldots & B_K(X_2) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & B_1(X_n) & B_2(X_n) & \ldots & .. & B_K(X_n) \end{bmatrix} \tag{3.20}$$

Se tiver característica k (número máximo de linhas linearmente independentes), o estimador de máxima verosimilhança $\widehat{\theta}$ existe. Nestas condições denominaremos

$$\hat{f}_s = \hat{f}(.,\theta) \tag{3.21}$$

por **estimador *logspline* da função densidade.**

A partir de (3.21) podemos determinar o estimador da densidade populacional utilizando *logsplines* e que denotaremos por $\widehat{D}_s$. Utilizando (2.14), obtemos,

$$\widehat{D}_s = \frac{n\widehat{f}_s(0)}{2L}. \tag{3.22}$$

Seja $H(\theta), \theta \in \Theta$ a matriz hessiana de $C(\theta)$. Determinemos as expressões genéricas dos seus elementos.

$$\begin{aligned} \frac{\partial C(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j}\left(\log\left\{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right]dx\right\}\right) = \\ &= \frac{\frac{\partial}{\partial \theta_j}\left\{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right]dx\right\}}{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right]dx} = \\ &= \frac{\int_L^U B_j(x)\exp\left[\sum_{j=1}^{k-1}\theta_j B_j(y)\right]dx}{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right]dx}. \end{aligned} \tag{3.23}$$

Por outro lado,

$$\phi(x,\theta) = \frac{\exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right]}{\exp\left[C(\theta)\right]} \Leftrightarrow \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] = \phi(x,\theta)\exp\left[C(\theta)\right] \Leftrightarrow$$

$$\Leftrightarrow \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] = \phi(x,\theta)\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx \qquad (3.24)$$

Utilizando a expressão anterior para substituir em (3.23), obtém-se,

$$\frac{\partial C(\theta)}{\partial \theta_j} = \frac{\int_L^U B_j(x)\left\{\phi(x,\theta)\int_R \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx\right\} dx}{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx} \Leftrightarrow$$

$$\Leftrightarrow \frac{\partial C(\theta)}{\partial \theta_j} = \int_L^U B_j(x)\phi(x,\theta)\,dx. \qquad (3.25)$$

Relativamente às derivadas parciais de segunda ordem, temos,

$$\frac{\partial^2 C(\theta)}{\partial\theta_j\partial\theta_k} = \frac{\partial}{\partial\theta_k}\left(\frac{\partial C(\theta)}{\partial\theta_j}\right) =$$

$$= \frac{\partial}{\partial\theta_k}\left[\int_L^U B_j(x)\phi(x,\theta)\,dx\right] =$$

$$= \int_L^U B_j(x)\left\{B_k(y) - \frac{\int_L^U B_j(x)\exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx}{\int_L^U \exp\left[\sum_{j=1}^{k-1}\theta_j B_j(x)\right] dx}\right\}$$

$$\phi(x,\theta)\,dx. \qquad (3.26)$$

Utilizando novamente (3.24) obtemos, finalmente, os elementos genéricos da matriz hessiana:

$$\frac{\partial^2 C(\theta)}{\partial\theta_j\partial\theta_k} = \int_L^U B_j(y)B_k(y)f(y,\theta)\,dy - \int_L^U B_j(y)f(y,\theta)\,dy \int_L^U B_k(y)\phi(y,\theta)\,dy. \tag{3.27}$$

Sejam $X_1,...,X_n$ uma amostra aleatória de dimensão n proveniente de f e $S(\theta)$ a função *Score*, constituida por um vector de dimensão p, com elementos dados por

$$S_j(\theta) = \frac{\partial l(\theta)}{\partial\theta_j} = b_j - n\frac{\partial C(\theta)}{\partial\theta_j} \tag{3.28}$$

onde as estatísticas suficientes $b_1, b_2, ..., b_j$ são definidas por

$$b_j = \sum_{i=1}^{n} B_j(x_i), \tag{3.29}$$

sendo a equação de máxima verosimilhança representada por

$$S(\theta) = 0.$$

Seja

$$I(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta_j\partial\theta_k} = \frac{\partial}{\partial\theta_k}\left[\sum_{i=1}^{n} B_j(x_i) - n\frac{\partial C(\theta)}{\partial\theta_j}\right] = = -n\frac{\partial^2 C(\theta)}{\partial\theta_j\partial\theta_k} = -nH(\theta) \tag{3.30}$$

a informação matricial correspondente a amostra aleatória considerada.

Então, utilizando o método de Newton Raphson, (ver por exemplo Stoer, J. (1996)), começa-se inicialmente com um valor $\widehat{\theta}^{(0)}$ (a forma de o calcular encontra-se descrita no apêndice sobre a implemetação numérica do método em Kooperberg & Stone, 1992) e obtém-se, iterativamente,

$$\hat{\theta}^{(m+1)} = \hat{\theta}^{(m)} - \left[H\left(\hat{\theta}^{(m)}\right)\right]^{-1} S\left(\hat{\theta}^{(m)}\right). \tag{3.31}$$

Uma informação mais detalhada sobre a implementação numérica destes procedimentos pode ser consultada em Kooperberg & Stone (1992).

3.2.2 Selecção de nós e escolha do modelo

A selecção de nós envolve a sua colocação inicial e a adição e delecção passo a passo, metodologia seguida por Kooperberg & Stone (1992). Inicia-se o processo com K nós, em que

$$K = \min\left(2, 5n^{\frac{1}{5}}; \frac{n}{4}; N; 25\right) \tag{3.32}$$

onde N é o número de valores distintos dos x_i's.

Para a localização dos nós são utilizadas estatísticas de ordem, método descrito em Kooperberg & Stone (1991). Um processo mais complexo, englobando casos em que os dados poderão ser censurados, é desenvolvido em Kooperberg & Stone (1992).

O procedimento de adição/delecção de nós é feito da seguinte forma: para cada passo do algoritmo adicionam-se nós, procurando-se em cada intervalo a sua melhor localização determinada pela existência dos nós t_1, t_2, ..., t_k. Este processo é realizado maximizando a estatística de Rao para os nós potenciais localizados nos quartis dos dados, em cada intervalo. O valor máximo de nós permitido é, por defeito,

$$K_{\max} = \min\left(4n^{\frac{1}{5}}; \frac{n}{4}; N; 30\right). \tag{3.33}$$

Durante o processo delecção de nós são retirados os que apresentam valores menos significativos. Para se obter essa medida de significância utiliza-se a estatística de Wald. Este processo encontra-se descrito com mais detalhe em Stone *et al* (1997).

De entre todos os modelos ajustados é escolhido o que apresentar o valor de AIC mais reduzido, em que

$$AIC_\alpha = -2\widehat{l} + \alpha J \tag{3.34}$$

sendo $\alpha = \log n$, $\widehat{l}_v$ o logaritmo da função de verosimilhança avaliado nas estimativas de máxima verosimilhança e J a dimensão do espaço das *splines* utilizado.

3.2.3 Variância e intervalos de confiança

"Obter intervalos de confiança (I. C.) na estimação de funções através de *splines* polinomiais é uma tarefa bastante complicada" (Kooperberg & Stone , 2002). Uma das formas de o fazer é tratar o modelo final como se fosse um modelo

paramétrico fixo, e portanto, empregar a metodologia clássica, só que ignorando-se todo o processo anterior que levou à escolha desse modelo, obtêm-se I.C. com uma pequena taxa de cobertura.

Só recentemente, Kooperberg & Stone , (2002 e 2003) analisaram com detalhe esta problemática, motivados por uma nova metodologia chamada "estimação de densidades por *logsplines* com nós livres" que se caracteriza, no essencial pelo facto dos nós, em vez de serem seleccionados por um algoritmo de adição/delecção passo a passo, serem incorporados na estimação por máxima verosimilhança funcionando como parâmetros adicionais do modelo. Para mais informação sobre esta metodologia consultar Kooperberg & Stone, (2001).

São comparados os dois métodos, nós livres e adição/delecção passo a passo, utilizando vários processos para obter os intervalos de confiança:

- utilização do vector gradiente e da matriz hessiana obtidos na estimação por máxima verosimilhança;
- técnica de Bootstrap não paramétrica;
- métodos bayesianos;
- uma técnica mista combinando bootstrap não paramétrico para obter o desvio padrão de $\widehat{f}_s(y)$,que designaremos por $\sigma_{[\widehat{f}_s(y)]}$, e o método clássico, ou seja, obtém-se um I.C. a 95% da forma

$$\left]\widehat{f}_s(y) - 1,96 \times \sigma_{[\widehat{f}_s(y)]} \;;\; \widehat{f}_s(y) + 1,96 \times \sigma_{[\widehat{f}_s(y)]}\right[. \qquad (3.35)$$

Curiosamente, foi neste último caso que se obtiveram os melhores resultados, chegando-se também à conclusão que a metodologia de selecção de nós passo a passo fornecia, computacionalmente, resultados bastante aproximados ao método dos nós livres e de uma forma bastante mais rápida, sendo portanto, aconselhada no tratamento de casos práticos.

Nos capitulos seguintes utilizaremos

Capítulo 4

Estimação de densidades populacionais por *logsplines* em amostragem por transectos lineares. Exemplos.

A primeira abordagem que fizemos no sentido de aplicar a metodologia *logspline* ao cálculo de $\widehat{f}(0)$, no contexto dos transectos lineares, revelou resultados pouco satisfatórios. Constatamos que o problema era causado pelo facto de se impôr um comportamento linear nas caudas da distribuição (ver secção 3.2.1.). Como, no nosso caso, as amostras são constituídas por valores não negativos, o valor que se pretende estimar, $f(0)$, encontra-se, precisamente, numa dessas caudas, o que fazia com que esse estimador fosse bastante enviesado.

A forma de contornar o problema, já utilizada em Chen, (1996), mas no contexto da estimação de densidades por *kernel* aplicada aos transectos lineares, foi, a partir da amostra $x_1, x_2, ..., x_n$, construir-se outra amostra

$$-x_n, ..., -x_2, -x_1, x_1, x_2, ..., x_n.$$

A partir desta obtém-se um novo estimador $\widehat{h}(x)$ é uma função par e

$$\widehat{f}_s(x) = \widehat{h}(-x) + \widehat{h}(x), \quad x \geq 0;$$

logo

$$\widehat{f}_s(0) = 2\widehat{h}(0).$$

Vamos ilustrar este procedimento com uma amostra aleatória de dimensão 1000 proveniente de uma distribuição *Uniforme (0,1)*. Utilizando o *package*

POLSPLINE (ver secção 4.1), estimamos $f_s(0)$, sem duplicar os dados por simetria, sendo o resultado obtido

$$\widehat{fs}(0) = 0,3691891. \tag{4.1}$$

Quando se considerou os valores da amostra e os seus simétricos obtivémos

$$\widehat{f_s}(0) = 2\widehat{h}(0) = 1,00178.$$

Reparemos que, no segundo caso, o resultado é muito satisfatório já que $f(0) = 1$. A figura 4.1 ilustra este procedimento, sendo claramente visível o comportamento instável em torno de zero no caso em que se estimou directamente a partir dos dados, sem fazermos a sua duplicação por simetria.

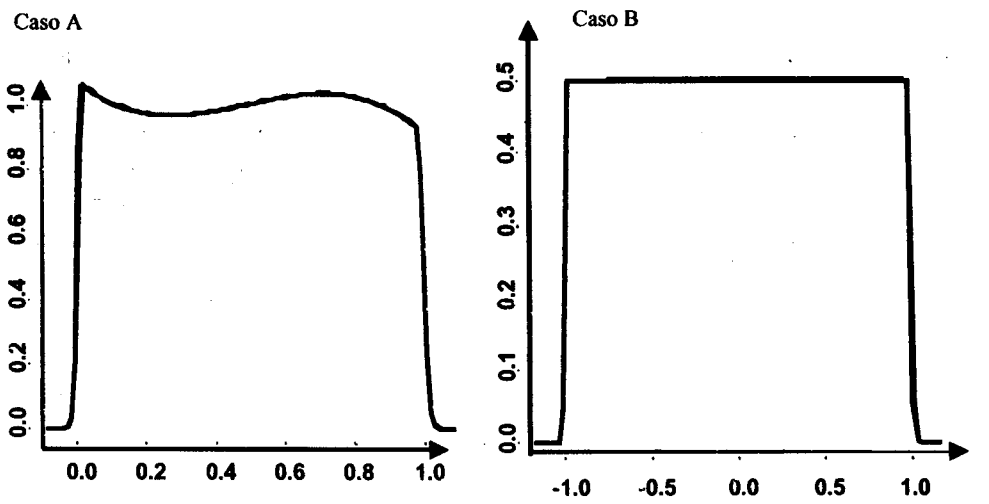

Figura 4.1: Caso A: representação gráfica de $\widehat{f}(x)$ no caso da amostra $x_1, ..., x_{1000}$ proveniente de uma distribuição uniforme $(0, 1)$.Caso B: representação gráfica de $\widehat{h}(x)$ no caso da amostra $-x_{1000}, ..., -x_1, x_1, ..., x_{1000}$.

4.1 Software

Utilizámos diversos tipos de software. A maior parte dos gráficos e do cálculo integral realizaram-se com o SCIENTIFIC WORKPLACE (versão 3.0), no qual também foi escrito o trabalho. Toda a progamação necessária para a realização das simulações foi escrita na linguagem utilizada pelo ambiente **R** (versão 1.6.2) desenvolvido por Ihaca & Gentleman (1996), um software de utilização gratuita

com largas potencialidades no domínio da programação relacionada com estatística e que é, relativamente à programação, bastante similar às linguagens C e FORTRAN. Encontra-se informação completa sobre o mesmo em **http://www.r-project.org/.**

Relativamente à estimação por *logsplines*, utilizamos o programa LOGSPLINE, que faz parte do *package* POLSPLINE (versão 1.0.1), sendo a parte referente às *logsplines* escrita por Charles Kooperberg , de acordo com a metodologia utilizada em Stone *et al.* (1997). Este *package* foi executado em ambiente **R** através dos programas atrás referidos. A escolha final do modelo é feita de acordo com o menor valor de A.I.C. dado pela equação (3.34).

Os estimadores relativos à teoria descrita no capítulo 2 foram obtidos com o programa DISTANCE, versão 3.5, no que diz respeito a análise dos casos práticos, e com a versão 2.2, nas simulações. A utilização destas duas versões têm a ver com o facto da versão mais recente não ser viável, em termos práticos, para se efectuarem as simulações, como fui amavelmente informado por um dos autores do programa. No entanto, os resultados obtidos, relativamente às duas versões, apresentam sensivelmente, a mesma qualidade. Encontra-se informação completa sobre este software em **http://www.ruwpa.st-and.ac.uk/distance.** Como é sugerido em Buckland *et al.*(2001), utilizámos neste programa, as funções:

$$\begin{array}{l} \text{Uniforme + Polinómio;} \\ \text{Uniforme + Série de cosenos;} \\ \text{Normal + Polinómio hermítico;} \\ \text{Normal + Série de cosenos;} \\ \text{Hazard-Rate + Série de cosenos.} \end{array} \tag{4.2}$$

O modelo escolhido é o que apresentar o valor de A.I.C. mais reduzido, de acordo com a equação (2.34).

4.2 Exemplos práticos

4.2.1 Estacas de madeira

Laake (1978) conduziu uma investigação sobre amostragem por transectos lineares numa zona semi-desértica a este de Logan, Utah, nos Estados Unidos da América.

Foram colocadas, numa área rectangular, 150 estacas de madeira de forma aleatória e de modo que a sua distribuição fosse uniforme. Os dados que vamos

analisar, retirados de Burnham *et al* (1980), foram recolhidos por um dos observadores que percorreu um transecto com um comprimento de 1000 metros, tendo avistado 68 estacas. Não foi considerada a truncatura de dados, sendo o valor de w correspondente à maior distância observada.

Existem duas razões para a escolha deste estudo: o facto de ser conhecida a verdadeira densidade (37,5 estacas / hectare), o que facilita a avaliação dos resultados, e também por estes dados já terem sido analisados por vários autores (*e.g.* Laake, 1978, Burnham *et al*, 1980, Buckland *et al*, 2001, Barabesi 2000, Quang & Lanctot, 1991 e Karunamuni & Quinn, 1995), o que permite uma melhor comparação com outras metodologias.

Apresentam-se na **tabela 4.1** os resultados obtidos. Um resumo dos *outputs* apresentados pelos programas e os dados originais encontram-se no apêndice A.

Tabela 4.1- Valores dos estimadores $\widehat{f}(0)$ e $\widehat{D}$ e *I.C.* a 95% relativos aos dados das estacas de madeira.O *I.C* para *logspline* foi calculado utilizando-se (3.35) e o de DISTANCE por (2.24)

$n = 68$ $L = 1000$	$f(0)$			D		
Verdadeiro	0,11			37,5		
	$\widehat{f}(0)$	*inf I.C.*	sup *I.C*	$\widehat{D}$	*inf I.C.*	sup *I.C.*
DISTANCE	0,1063	0,0837	0,1350	36,16	25,785	50,726
logspline	0,1122	0,0749	0,1495	38,17	11,673	64,656

A figura 4.2 apresenta as curvas obtidas das funções densidade de probabilidade estimadas por DISTANCE e POLSPLINE que representam o ajustamento ao histograma representativo das distâncias observadas.

As duas metodologias apresentam resultados satisfatórios, muito aproximados dos valores reais, apresentando intervalos de confiança contendo o verdadeiro valor que se pretende estimar.

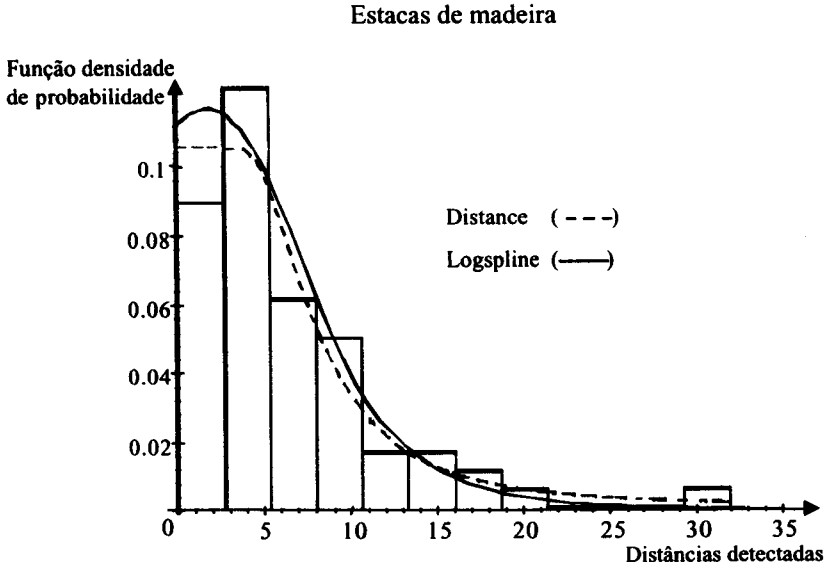

Figura 4.2: Representação gráfica das f.d.p. estimadas por DISTANCE (a tracejado) e por LOGSPLINE (a cheio).

Embora o I.C. relativo à metodologia *logspline* apresente uma amplitude superior à de Buckland, os valores obtidos para $\widehat{f}(0)$ e $\widehat{D}$, tem uma maior precisão.

A análise do gráfico da figura 4.2 revela que, nos dois casos analisados, as curvas se ajustam ao histograma representativo das distâncias observadas. Nota-se um comportamento diferente em A. A função estimada pela metodologia *logspline* não é monótona decrescente nesse intervalo. Uma das principais vantagens desta metodologia tem haver com a sua adaptabilidade aos dados observados sendo portanto, normal esse seu comportamento, se tomarmos em conta o histograma apresentado.

4.2.2 Ungulados africanos

Paul Hemingway estudou a densidade populacional de vários ungulados no Continente Africano, utilizando a metodologia dos transectos lineares. Num dos seus conjuntos de observações, recolheu 73 distâncias ao longo de um transecto com 60 Km. Os dados que vamos analisar, também foram retirados de Burnham *et al* (1980). Apresentam-se na tabela 4.2 os resultados obtidos. Um resumo dos *outputs* apresentados pelos programas e os dados originais encontram-se no apêndice C.

Tabela 4.2- Valores dos estimadores $\widehat{f}(0)$ e $\widehat{D}$ e *I.C.* a 95% relativos aos dados dos ungulados africanos. O *I.C* para *logspline* foi calculado utilizando-se (3.35).

$n = 73$ $L = 60km$						
	$\widehat{f}(0)$	*inf I.C.*	sup *I.C*	$\widehat{D}$	*inf I.C.*	sup *I.C.*
DISTANCE	0,0063	0,0053	0,0075	38,8	29,05	51,83
logspline	0,0065	0,0058	0,0073	40	35,553	44,62

A figura 4.3 apresenta as curvas das funções densidade de probabilidade estimadas por DISTANCE e POLSPLINE e o ajustamento ao histograma representativo das distâncias observadas.

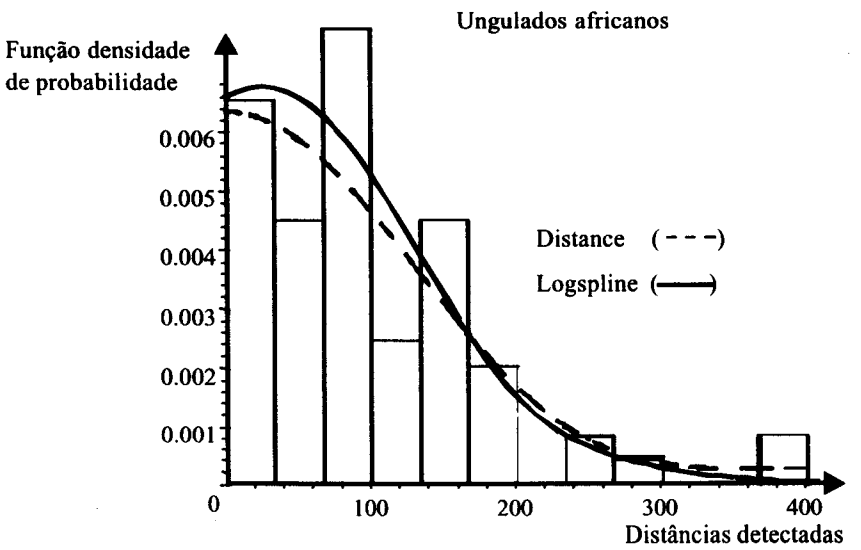

Figura 4.3: Representação gráfica das f.d.p. estimadas por DISTANCE (a tracejado) e por LOGSSPLINE (a cheio).

Analisando a tabela 4.2 constata-se que a amplitude dos intervalos de confiança, para D e $f(0)$ são menores para a metodologia *logspline*, o que não tinha acontecido no outro exemplo apresentado, apresentando logspline, neste caso uma menor variabilidade.

A análise do gráfico da figura 4.3 revela que, nos dois casos analisados, as curvas se ajustam razoalvelmente ao histograma representativo das distâncias observadas. A função estimada pela metodologia *logspline* para este exemplo também não é monótona.

Da análise destes dois exemplos, podemos afirmar que os resultados obtidos pela metodologia *logspline* parecem-nos adequados. Os valores obtidos no primeiro exemplo são bastante satisfatórios relativamente aos valores conhecidos. No segundo exemplo, embora não se conheçam os verdadeiros valores, a comparação com uma das metodologias mais utilizadas actualmente permite dizer que, pelo menos, são tão satisfatórios quanto esta.

Capítulo 5

Simulações. Avaliação das metodologias

Para testarmos a adequação da teoria de *logsplines* à metodologia dos transectos lineares, realizámos um conjunto alargado de simulações, no sentido de cobrir uma grande diversidade de situações que possam eventulamente, surgir nos casos práticos. Para tal, utilizaram-se diferentes tipos de funções de detecção e de dimensões das amostras. Utilizámos também várias medidas para avaliarmos a perfomance dos estimadores obtidos.

Começamos por gerar os valores $y_1, y_2, ..., y_N$, representando as distâncias de N objectos distribuidos uniformemente ao longo de uma faixa rectangular em que $w = 1$ e $L = 100$, ou seja,

$$D = \frac{N}{2wL} = \frac{N}{200}. \tag{5.1}$$

Seguidamente, simulou-se a detecção desses objectos de acordo com uma dada função de detecção $g(x)$, resultando n valores de distâncias "detectadas" $x_1, x_2, ..., x_n$. O pressuposto 1 (secção 2.1) será sempre verificado.

5.1 *logspline* versus distance

Numa primeira abordagem optamos por comparar as metodologias decritas nas secções 2.4 e 3.2. O processo de simulação foi repetido 200 vezes para cada uma das populações de dimensão $N = \{40; 80; 120\}$, sendo os valores das densidades $D = \{0,2; 0,4; 0,6\}$, repectivamente. Seguidamente utilizamos para cada uma delas seis funções de detecção para obter as amostras das populações, tendo-se, portanto realizado um total de $3 \times 200 \times 6$ simulações.

O processo de escolha das amostras foi feito calculando-se a probabilidade de detecção para cada um dos objectos através da função de detecção. Gerando-se um número aleatório entre zero e um, selecciona-se objecto se a sua probabilidade detecção for superior a esse número.

Para cada uma das amostras foram calculados os estimadores de D, utilizando os programas DISTANCE e LOGSPLINE. Escrevemos várias rotinas em ambiente **R** (as mais importantes encontram-se no apêndice C), que permitiram realizar o processo de simulação bem como formatar os dados para serem executados no DISTANCE. Obtiveram-se, para cada conjunto de 200 simulações, várias medidas, calculadas para os dados provenientes das duas metodologias, que passamos a descriminar:

Valor esperado

Representa o valor médio dos $\widehat{D}_i$'s calculados nas 200 simulações

$$\widehat{E}\left(\widehat{D}_i\right) = \frac{\sum_{i=1}^{200} \widehat{D}_i}{200}. \tag{5.2}$$

Viés

Mede o enviesamento através da diferença entre o valor esperado do estimador e o verdadeiro valor do parâmetro que se quer estimar.

$$\widehat{v}\left(\widehat{D}_i\right) = \widehat{E}\left(\widehat{D}_i\right) - D. \tag{5.3}$$

Viés relativo (%)

Dá, em percentagem, o peso relativo do enviesamento.

$$\widehat{v}(\%) = \frac{\widehat{E}\left(\widehat{D}_i\right) - D}{D} \times 100. \tag{5.4}$$

Variância

Mede a precisão do estimador.

$$\widehat{Var}\left(\widehat{D}_i\right) = \frac{\sum_{i=1}^{200} \left[\widehat{D}_i - \widehat{E}\left(\widehat{D}_i\right)\right]^2}{200 - 1}. \tag{5.5}$$

Erro padrão

$$\widehat{Ep}\left(\widehat{D}_i\right) = \sqrt{\frac{\widehat{Var}\left(\widehat{D}\right)}{200}} \tag{5.6}$$

Erro padrão relativo

$$\widehat{ES}\,(\%) = \frac{\widehat{ES}\left(\widehat{D}_i\right)}{D} \times 100 \tag{5.7}$$

Raiz quadrada do erro quadrático médio

Mede a perfomance do estimador.

$$RE\widehat{Q}M\left(\widehat{D}_i\right) = \sqrt{E\widehat{Q}M\left(\widehat{D}_i\right)} = \sqrt{\widehat{Var}\left(\widehat{D}_i\right) + \left[\widehat{v}\left(\widehat{D}_i\right)\right]^2} \tag{5.8}$$

Passamos a descrever, sucintamente, as funções de detecção utilizadas bem como os resultados obtidos.

Função Normal

$$g_1(x) = \exp\left(-\frac{x^2}{2 \times 0,6^2}\right), \quad 0 \leq x \leq 1. \tag{5.9}$$

Trata-se de uma das funções de detecção mais utilizadas. O parâmetro $\sigma = 0,6$ assegura que pelo menos metade da população é observada. Repare-se que

$$\mu = \int_0^1 \left[\exp\left(-\frac{x^2}{2 \times 0,6^2}\right)\right] dx \approx 0,68011. \tag{5.10}$$

A figura 5.1 ilustra a sua representação gráfica. O facto de $g'(0) = 0$, permite que o gráfico da função tenha um comportamento "suave" em A.

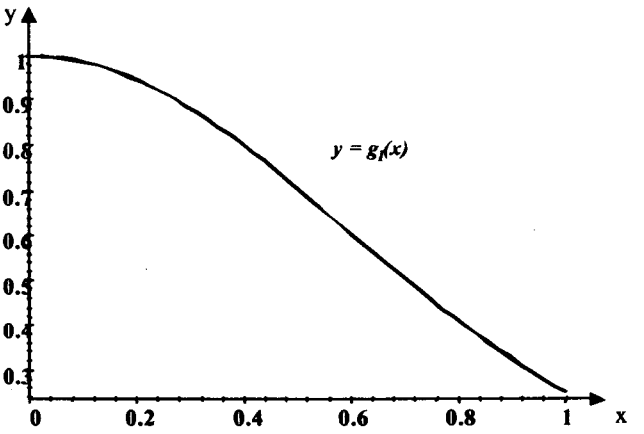

Figura 5.1: Representação gráfica da função de detecção, $g_1(x) = \exp\left(-\frac{x^2}{2\times 0.6^2}\right)$, $0 \leq x \leq 1$.

Tabela 5.1 - Comparação dos resultados obtidos por LOGSPLINE e DISTANCE relativos à função de detecção $g_1(x)$.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$\widehat{Var}\left(\widehat{D}_i\right)$	$\widehat{v}\,(\%)$	$\widehat{Ep}\,(\%)$	$RE\widehat{Q}M\left(\widehat{D}_i\right)$
LOGSPLINE	40	0,217	0,007	8,742	3,134	0,089
DISTANCE	40	0,192	0,004	−3,581	2,314	0,063
LOGSPLINE	80	0,417	0,018	4,447	2,394	0,134
DISTANCE	80	0,404	0.006	1,056	1,427	0,077
LOGSPLINE	120	0,611	0,026	1,863	1,913	0,161
DISTANCE	120	0,605	0,009	0,981	1,124	0,077

A análise da tabela 5.1 permite concluir que os resultados obtidos com DISTANCE são substancialmente melhores que os obtidos por LOGSPLINE. Estes resultados já seriam previsíveis uma vez que um dos modelos utilizados por DISTANCE é precisamente a distribuição normal.

Função Exponencial

$$g_2(x) = \exp\left(-\frac{x}{0,6}\right), \quad 0 \leq x \leq 1. \tag{5.11}$$

Trata-se de uma função utilizada em situações muito específicas, por exemplo, em florestas muito densas (Mack *et al*,1999), quando o número de detecções decai rapidamente (ver figura 5.2) à medida que aumentam as distâncias ao transecto. Neste caso não se verifica a *shoulder condiction* já que $g_2'(0) \neq 0$, como se poderá também observar na figura 5.2. A faixa efectiva de amostragem é dada por

$$\mu = \int_0^1 \exp\left(-\frac{x}{0,6}\right) dx \approx 0,48667. \tag{5.12}$$

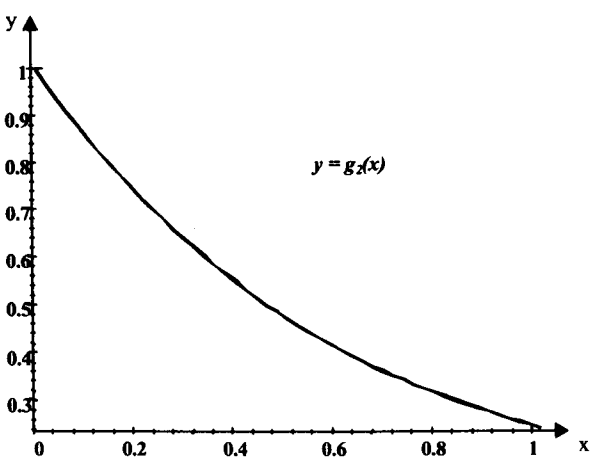

Figura 5.2: Representação gráfica da função de detecção, $g_2(x) = \exp\left(-\frac{x}{0.6}\right)$, $0 \leq x \leq 1$.

Tabela 5.2 - Comparação dos resultados obtidos por LOGSPLINE e DISTANCE relativos à função de detecção $g_2(x)$.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$\widehat{Var}\left(\widehat{D}_i\right)$	$\widehat{v}\,(\%)$	$\widehat{Ep}(\%)$	$\widehat{REQM}\left(\widehat{D}_i\right)$
LOGSPLINE	40	0,203	0,007	1,959	3,057	0,083
DISTANCE	40	0,222	0,674	11,491	29,036	0,821
LOGSPLINE	80	0,383	0,012	-4,028	2,014	0,114
DISTANCE	80	0,351	0,021	-12,149	2,607	0,154
LOGSPLINE	120	0,569	0,021	-5,119	1,743	0,148
DISTANCE	120	0,539	0,040	-10,089	2,371	0,209

A análise da tabela 5.2 revela que os resultados obtidos por LOGSPLINE são substancialmente melhores do que os de DISTANCE. Este facto estará directamente relacionado com a função de detecção utilizada já que apresenta duas características diferentes em relação às outras: decresce rapidamente e tem derivada diferente de zero na origem.

Exponencial generalizada

$$g_3(x) = e^{-\left(\frac{x}{0,6}\right)^5}, \quad 0 \leq x \leq 1 \tag{5.13}$$

O gráfico desta função caracteriza-se por apresentar um decrescimento pouco acentuado perto do transecto (figura 5.3). Temos, neste caso, $g_3'(0) = 0$, sendo

$$\mu = \int_0^1 \left(e^{-\left(\frac{x}{0,6}\right)^5}\right) dx = 0,5509 \tag{5.14}$$

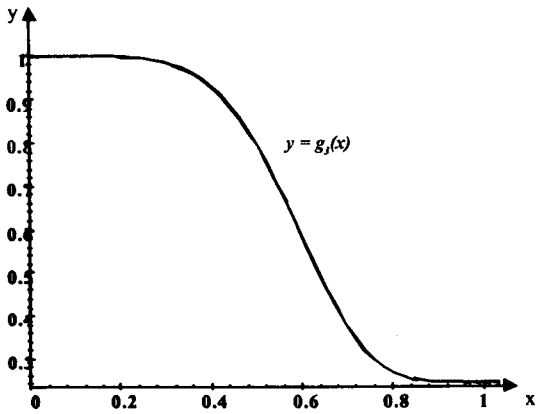

Figura 5.3: Representação gráfica da função de detecção, $g_3(x) = \exp\left[-\left(\frac{x}{0.6}\right)^5\right]$, $0 \leq x \leq 1$.

Tabela 5.3 - Comparação dos resultados das estimações entre LOGSPLINE e DISTANCE relativos à função de detecção $g_3(x)$.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$\widehat{Var}\left(\widehat{D}_i\right)$	$\widehat{v}\,(\%)$	$\widehat{Ep}(\%)$	$\widehat{REQM}\left(\widehat{D}_i\right)$
LOGSPLINE	40	0,220	0,012	10,177	3,932	0,109
DISTANCE	40	0,202	0,003	1,149	1,945	0,054
LOGSPLINE	80	0,408	0,022	2,060	2,655	0,148
DISTANCE	80	0,425	0,007	6,334	1,482	0,083
LOGSPLINE	120	0,612	0,037	2,067	2,277	0,192
DISTANCE	120	0,656	0,011	9,423	1,278	0,118

Os resultados apresentados na tabela 5.3 não são inconclusivos quanto à comparação dos dois métodos. Enquanto que LOGSPLINE apresenta valores médios superiores a DISTANCE para $N = 80$ e $N = 120$, a $\widehat{REQM}$ é superior para qualquer uma das dimensões consideradas o que é motivado pela maior variabilidade apresentada por *logspline*.

Normal+termos de uma série de cosenos

$$g_4(x) = 2e^{-\frac{x^2}{2\times 0,45^2}} \times \left(1 - 0,5\cos(\frac{\pi x}{10}) + 0,4\cos(\frac{\pi x}{5}) - 0,4\cos(\frac{\pi x}{15})\right) \quad (5.15)$$

Temos, neste caso, um dos modelos utilizados pelo programa DISTANCE. Pretende-se testar até que ponto o programa "detecta" uma das suas funções-chave, bem como o comportamento da metodologia *logspline* relativamente a esta função. A figura 5.4 ilustra o comportamento do gráfico da função. Também temos, neste caso, $g_4'(0) = 0$. A faixa efectiva de amostragem é dada por:

$$\mu = \int_0^1 g_4(x)dx \approx 0,540 \quad (5.16)$$

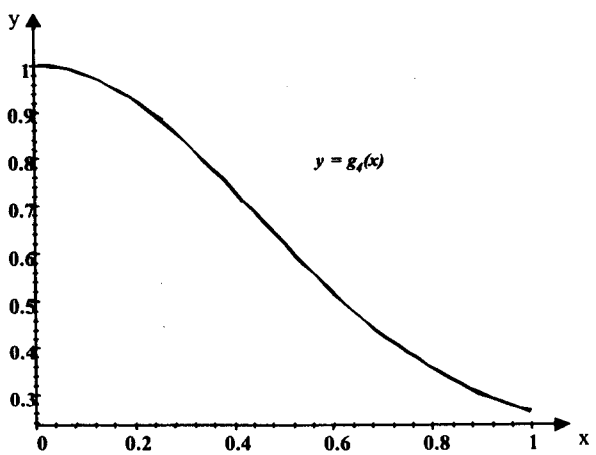

Figura 5.4: Representação gráfica da função de detecção, $g_4(x) = 2e^{-\frac{x^2}{2\times 0.45^2}} \times \left(1 - 0.5\cos(\frac{\pi x}{10}) + 0.4\cos(\frac{\pi x}{5}) - 0.4\cos(\frac{\pi x}{15})\right)$, $0 \leq x \leq 1$.

Tabela 5.4 - Comparação dos resultados das estimações entre LOGSPLINE e DISTANCE relativos à função de detecção $g_4(x)$.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$V\widehat{a}r\left(\widehat{D}_i\right)$	$\widehat{v}(\%)$	$\widehat{Ep}(\%)$	$RE\widehat{Q}M\left(\widehat{D}_i\right)$
LOGSPLINE	40	0,215	0,007	7,947	3,078	0,083
DISTANCE	40	0,184	0,003	-7,560	2,215	0,063
LOGSPLINE	80	0,429	0,015	5,729	2,210	0,126
DISTANCE	80	0,393	0,008	-1,56	1,642	0,089
LOGSPLINE	120	0,607	0,024	1,247	1,832	0,154
DISTANCE	120	0,595	0,009	-0,750	1,122	0,077

A análise da tabela 5.4 indica-nos um melhor comportamento de DISTANCE relativamente a esta função de detecção o que não contraria as nossas previsões já que se trata de um modelo incorporado neste programa e, portanto, permitindo uma maior qualidade nas estimativas.

Função não paramétrica 1

$$g_5(x) = \exp\left(-\frac{x^3}{0,2}\right) + 0,5x^3 \exp\left(-\frac{x^2}{0,1}\right) \tag{5.17}$$

Apresentamos uma função não paramétrica, embora com características idênticas à anterior, na sua representação gráfica (ver figura 5.5) e também com $g_5'(0) = 0$. Neste caso tem-se,

$$\mu = \int_0^1 g_5(x)dx = 0,52431 \tag{5.18}$$

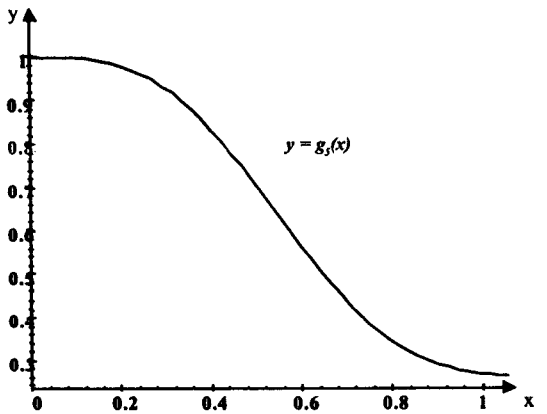

Figura 5.5: Representação gráfica da função de detecção, $g_5(x) = \exp\left(-\frac{x^3}{0,2}\right) + 0,5x^3 \exp\left(-\frac{x^2}{0,1}\right)$, $0 \leq x \leq 1$.

Tabela 5.5 - Comparação dos resultados das estimações entre LOGSPLINE e DISTANCE relativos à função de detecção $g_5(x)$.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$\widehat{Var}\left(\widehat{D}_i\right)$	$\widehat{v}\,(\%)$	$\widehat{Ep}(\%)$	$\widehat{REQM}\left(\widehat{D}_i\right)$
LOGSPLINE	40	0,220	0,010	10,428	3,650	0,104
DISTANCE	40	0,196	0,003	-1,791	2,064	0,054
LOGSPLINE	80	0,421	0,020	5,349	2,507	0,141
DISTANCE	80	0,421	0,007	5,330	1,495	0,083
LOGSPLINE	120	0,612	0,025	2,035	1,896	0,161
DISTANCE	120	0,626	0,008	4,350	1,085	0,094

Os resultados obtidos por DISTANCE são melhores, embora as duas metodologias apresentem bons resultados.

Função não paramétrica 2

$$g_6(x) = \frac{1}{5x^2 + 1} \tag{5.19}$$

Optamos por mais uma função com características idênticas à anterior (ver gráfico 5.6). O objectivo é testar em que medida, os resultados obtidos na simulação são idênticos nas duas funções. Reparemos que $g_6'(0) = 0$, sendo

$$\mu = \int_0^1 g_6(x)dx = 0,51441. \tag{5.20}$$

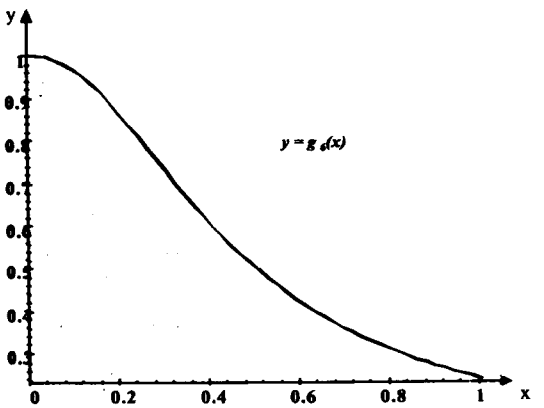

Figura 5.6: Representação gráfica da função de detecção, $g_6(x) = \frac{1}{5x^2+1}$, $0 \leq x \leq 1$.

A análise da tabela 5.6 revela um melhor comportamento de DISTANCE já que apresenta valores de erro quadrático médio inferiores. A comparação entre as duas metodologias nestes dois últimos casos apresentados, com funções de detecção apresentando propriedades idênticas, revelou uma maior sensibilidade de DISTANCE, o que permite concluir que neste caso, o modelo *logspline* apresentou maior robustez. Reparemos que os valores dos erros quadráticos médios mantiveram-se em LOGSPLINE, havendo diferenças maiores quando se aplicou o programa DISTANCE.

Na tabela 5.7 apresentamos os resultados obtidos nas várias simulações relativamente ao número de objectos detectados.Observam-se resultados bastante aceitáveis para as dimensões das amostras obtidas pelo processo de simulação sendo muito idênticos aos valores obtidos para μ.

Tabela 5.6 - Comparação dos resultados das estimações entre LOGSPLINE e DISTANCE relativos à função de detecção $g_6(x)$.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$\widehat{Var}\left(\widehat{D}_i\right)$	$\widehat{cv}\,(\%)$	$\widehat{Ep}(\%)$	$\widehat{REQM}\left(\widehat{D}_i\right)$
LOGSPLINE	40	0,220	0,007	10,386	2,961	0,083
DISTANCE	40	0,194	0,011	-2,791	3,797	0,104
LOGSPLINE	80	0,423	0,015	5,855	2,223	0,126
DISTANCE	80	0,393	0,019	-1,58	2,444	0,137
LOGSPLINE	120	0,624	0,023	4,055	1,793	0,151
DISTANCE	120	0,600	0,034	0,122	2,203	0,184

Tabela 5.7 - Comparação entre o número de objectos detectados e o e o nº dos que se esperariam detectar (μ).

	N		40	80	120
		μ			
$g_1(x)$	$\widehat{E}(\widehat{n})$		27,02	54,27	80,93
	$\widehat{\sigma}(\widehat{n})$		3,18	4,04	5,57
	$\frac{\widehat{E}(\widehat{n})}{N}$		0,67	0,67	0,67
		0,68			
$g_2(x)$	$\widehat{E}(\widehat{n})$		19,38	38,57	58,28
	$\widehat{\sigma}(\widehat{n})$		3,43	4,33	5,4
	$\frac{\widehat{E}(\widehat{n})}{N}$		0,48	0,482	0,48
		0,48			
$g_3(x)$	$\widehat{E}(\widehat{n})$		22,06	44,31	65,97
	$\widehat{\sigma}(\widehat{n})$		3,43	4,19	5,99
	$\frac{\widehat{E}(\widehat{n})}{N}$		0,55	0,553	0,54
		0,55			
$g_4(x)$	$\widehat{E}(\widehat{n})$		21,54	43,2	64,4
	$\widehat{\sigma}(\widehat{n})$		3,12	4,27	5,19
	$\frac{\widehat{E}(\widehat{n})}{N}$		0,53	0,54	0,53
		0,54			
$g_5(x)$	$\widehat{E}(\widehat{n})$		20,82	41,98	62,14
	$\widehat{\sigma}(\widehat{n})$		3,15	4,12	5,3
	$\frac{\widehat{E}(\widehat{n})}{N}$		0,52	0,524	0,517
		0,42			
$g_6(x)$	$\widehat{E}(\widehat{n})$		20,58	40,89	61,57
	$\widehat{\sigma}(\widehat{n})$		3,25	4,23	5,38
	$\frac{\widehat{E}(\widehat{n})}{N}$		0,52	0,511	0,513
		0,51			

5.2 Avaliação da metodologia *logspline*

Com o objectivo de testarmos a *perfomance* da metodologia *logspline* aplicada a amostragem por transectos lineares decidimos realizar um conjunto de simu-

lações num contexto mais alargado. Por um lado decidimos testar o método para dimensões populacionais maiores, neste caso

$$N = \{250, 500, 1000\},$$

por outro realizamos 1000 simulações em vez das 200 realizadas anteriormente. Por esse facto decidimos também manter as dimensões utilizadas anteriormente, ou seja,

$$N = \{40, 80, 120\}.$$

Relativamente às funções de detecção, utilizamos as mesmas que na secção anterior para as dimensões mais reduzidas. Para

$$N = \{250, 500, 1000\},$$

resolvemos modificar algumas das constantes dessas funções para garantir que fossem observados, em média, aproximadamente 20% das populações (as funções utilizadas encontra-se na tabela 5.8.). Mantivémos valores aproximados de 50% nas populações pequenas porque era a única forma de aferirmos a qualidade das estimações para essas populações.

Não utilizamos o programa DISTANCE para estas simulações porque apresentou problemas ao nível do cálculo numérico (conclusão retirada das mensagens de erro apresentadas pelo programa). Pensamos que não funcionará de um modo adequado pelo facto da percentagem de detecções ser pequena, o que motivará problemas nos cálculos de estimação por máxima verosimilhança, por incorporar modelos paramétricos.

Tabela 5.8 - Funções de detecção utilizadas para a realização das 1000 simulações.

$g_1^*(x) = \exp\left(-\frac{x^2}{2*0,16^2}\right)$
$g_2^*(x) = \exp\left(-\frac{x}{0,2}\right)$
$g_3^*(x) = e^{-\left(\frac{x}{0,25}\right)^5}$
$g_4^*(x) = 2e^{-\frac{x^2}{2\times 0,16^2}} \times \left(1 - 0,5\cos(\frac{\pi x}{10}) + 0,4\cos(\frac{\pi x}{5}) - 0,4\cos(\frac{\pi x}{15})\right)$
$g_5^*(x) = \exp(-\frac{x^3}{0,01}) + 0,4x^3\exp(-\frac{x^2}{0,2})$
$g_6^*(x) = \frac{1}{50x^2+1}$

Os resultados obtidos nestas simulações (tabela 5.9) são satisfatórios, observando-se no entanto, valores elevados do erro quadrático médio para as dimensões $N = 500$ e $N = 1000$, devidos à variabilidade, já que os valores do enviesamento se mantêm em níveis aceitáveis. Constatamos também que, para dimensões iguais, observamos valores de erro quadrático médio semelhantes, o que permite concluir que a metodologia não é sensivel às diferentes funções de detecção utilizadas.

Tabela 5.9- Resultados obtidos nas 1000 simulações realizadas para cada uma das funções de detecção consideradas.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$V\widehat{a}r\left(\widehat{D}_i\right)$	$\widehat{v}(\%)$	$\widehat{E}p(\%)$	$RE\widehat{Q}M\left(\widehat{D}_i\right)$
	40	0,222	0,007	11,205	1,397	0,091
g_1	80	0,422	0,014	5,657	0,944	0,121
	120	0,613	0,023	2,185	0,812	0,154
	250	1,126	0,097	-9,86	0,788	0,335
g_1^*	500	2,276	0,242	-8,938	0,622	0,540
	1000	4,64	0,551	-7,18	0,469	0,824
	40	0,203	0,007	1,836	1,369	0,086
g_2	80	0,388	0,014	-2,899	0,951	0,120
	120	0,564	0,025	-5,907	0.834	0,162
	250	1,051	0,072	-15,888	0,680	0,334
g_2^*	500	2,092	0,148	-16,293	0,487	0,56
	1000	4,197	0,402	-16,055	0,401	1,023
	40	0,218	0,012	9,263	1,765	0,113
g_3	80	0,413	0,022	3,309	1,195	0,151
	120	0,602	0,035	0,346	1,993	0,188
	250	1,265	0,19	1,278	1,104	0,436
g_3^*	500	2,509	0,316	0,398	0,711	0,562
	1000	5,027	0,638	0,541	0,505	0,799

Tabela 5.9 - Continuação.

	N	$\widehat{E}\left(\widehat{D}_i\right)$	$\widehat{Var}\left(\widehat{D}_i\right)$	$\widehat{v}(\%)$	$\widehat{Ep}(\%)$	$E\widehat{Q}M\left(\widehat{D}_i\right)$
	40	0,224	0,01	12,143	1,593	0,103
g_4	80	0,428	0,017	7,147	1,036	0,134
	120	0,623	0,026	3,927	0,859	0,164
	250	1,332	0,152	6,577	0,988	0,399
g_4^*	500	2,607	0,307	4,28	0,701	0,564
	1000	5,124	0,739	2,493	0,543	0,868
	40	0,227	0,013	13,666	1,827	0,118
g_5	80	0,429	0,02	7,404	1,145	0,147
	120	0,613	0,023	2,185	0,812	0,154
	250	1,364	0,157	9,131	1,005	0,413
g_5^*	500	2,678	0,332	7,153	0,729	0,603
	1000	5,182	0,666	3,651	0,516	0,836
	40	0,222	0,007	11,205	1,397	0,091
g_6	80	0,422	0,014	5,657	0,944	0,121
	120	0,613	0,023	2,185	0,812	0,154
	250	1,126	0,097	-9,86	0,788	0,335
g_6^*	500	2,276	0,242	-8,938	0,622	0,540
	1000	4,64	0,551	-7,18	0,469	0,824

Tabela 5.10 - Comparação entre o número de objectos detectados e o n^{o}dos que se esperariam detectar (μ).

N			40	80	120			250	500	1000
		μ					μ			
$\widehat{E}(\widehat{n})$			27,15	54,3	81,5			50,18	100,94	200,93
$\widehat{\sigma}(\widehat{n})$	$g_1(x)$		3,05	4,02	5,12	$g^*(x)$		6,43	9,1	13,54
$\frac{\widehat{E}(\widehat{n})}{N}$			0,67	0,67	0,67			0,19	0,19	0,19
		0,68					0,2			
$\widehat{E}(\widehat{n})$			19,47	38,84	58,25			49,64	99,84	198,86
$\widehat{\sigma}(\widehat{n})$	$g_2(x)$		3,07	4,43	5,42	$g_2^*(x)$		6,42	9,15	13,23
$\frac{\widehat{E}(\widehat{n})}{N}$			0,48	0,48	0,48			0,2	0,20	0,2
		0,48					0,19			
$\widehat{E}(\widehat{n})$			21,98	43,92	65,93			57,51	115,18	230,07
$\widehat{\sigma}(\widehat{n})$	$g_3(x)$		3,3	4,38	5,61	$g_3^*(x)$		6,54	9,29	13,94
$\frac{\widehat{E}(\widehat{n})}{N}$			0,54	0,54	0,54			0,23	0,23	0,23
		0,55					0,23			
$\widehat{E}(\widehat{n})$			21,56	43,17	64,79			50,07	100,7	200,48
$\widehat{\sigma}(\widehat{n})$	$g_4(x)$		3,15	4,37	5,31	$g_4^*(x)$		6,43	9,1	13,54
$\frac{\widehat{E}(\widehat{n})}{N}$			0,53	0,53	0,53			0,2	0,2	0,2
		0,54					0,2			
$\widehat{E}(\widehat{n})$			20,87	42,02	61,56			50,17	100,6	200,64
$\widehat{\sigma}(\widehat{n})$	$g_5(x)$		3,11	4,44	5,42	$g_5^*(x)$		6,43	9	13,44
$\frac{\widehat{E}(\widehat{n})}{N}$			0,52	0,52	0,51			0,2	0,2	0,2
		0,42					0,2			
$\widehat{E}(\widehat{n})$			20,63	41,09	61,56			50,64	101,75	202,58
$\widehat{\sigma}(\widehat{n})$	$g_6(x)$		3,12	4,42	5,42	$g_6^*(x)$		6,49	9,04	13,42
$\frac{\widehat{E}(\widehat{n})}{N}$			0,51	0,51	0,51			0,2	0,2	0,2
		0,51					0,2			

Na tabela 5.10 observa-se novamente resultados bastante aceitáveis para as dimensões das amostras obtidas pelo processo de simulação sendo muito idênticos aos valores obtidos para μ.

Capítulo 6

Conclusão

Após uma pequena introdução no capítulo 1, apresentamos no capítulo 2 as ideias fundamentais sobre a metodologia de amostragem por transectos lineares. Nessa apresentação tornou-se evidente a importância da função de detecção no cálculo do estimador da densidade populacional dependendo, em grande parte desta escolha, a qualidade desse estimador. É de referir que o principal foco das investigações realizadas nesta área nas últimas décadas (e.g. Buckland *et al.*, 2000), foi a procura de estimadores de $f(0)$ que apresentassem boas propriedades estatísticas. Os trabalhos de Buckland (1992a) com a utilização de uma função paramétrica, a que são adicionados alguns termos de uma série não paramétrica, os resultados obtidos com estimação por Kernel (Quang, 1991 e Chen, 1995), a estimação local por máxima verosimilhança proposta por Barabesi (2000) e Barabesi *et al.* (2002) e a estimação por métodos bayesianos (Karunamuni & Quinn II, 1995) são elucidativos do trabalho que tem sido feito nesta área.

No capítulo 3 apresentaram-se e desenvolveram-se os aspectos mais importantes de uma metodologia recente denominada *logspline density estimation* (Stone & Koo, 1986a e 1986b, e Stone,1990). No início desse capítulo encontra-se um conjunto de definições relacionadas com splines polinomiais, suporte fundamental desta teoria, seguindo-se a definição de família *logspline* e ainda uma explicação detalhada da estimação por máxima verosimilhança que é utilizada neste modelo. Por fim, explicámos a forma como são seleccionados e posicionados os nós, como é escolhido o modelo e a forma de deteminar intervalos de confiança.

Começámos o capítulo 4 apresentando as modificações necessárias para que se pudesse aplicar a metodologia logspline ao cálculo da estimação de $f(0)$ no contexto dos transectos lineares. Seguiu-se uma descrição do *software* utilizado neste trabalho para a análise da metodologia e para a compararação com a proposta por Buckland *et al.* (2001). Fizémos uma descrição dos resultados obtidos com dados provenientes de duas investigações de campo, as **estacas de madeira** e os **ungulados africanos,** ambas descritas detalhadamente em Burnham *et al.*, (1980). Concluiu-se que, nestes exemplos, os resultados apresentados pelas

duas metodologias são satisfatórios, não apresentando diferenças significativas ao nível dos estimadores de $f(0)$ e de D, assim como dos respectivos intervalos de confiança. A análise dos gráficos das funções de detecção seleccionadas pelos programas através do A.I.C., revelou um bom ajustamento ao histograma representativo das distâncias observadas. A leitura dos resultados obtidos por alguns autores, relativos aos mesmos dados (e.g. Barabesi, 2000; Karunamuni & Quinn II, 1995 e Burnham *et al.*,1980) não revelaram diferenças significativas relativamente às metodologias supracitadas.

Para avaliarmos a qualidade dos resultados obtidos pela estimação de f.d.p, através da metodologia *logspline* quando aplicada aos trasectos lineares, efectuámos um conjunto alargado de simulações. No capítulo 5 descrevemos e analisámos todo esse processo. Começámos por comparar os resultados obtidos por DISTANCE e LOGSPLINE, simulando-se populações com dimensões de 40, 80 e 120 indivíduos distribuídos uniformemente numa determinada área. Através da utilização de 6 funções de detecção efectuámos as amostragens estimando-se seguidamente D. Este processo foi repetido num total de $200 \times 6 \times 3$ simulações.

Concluímos que as metodologias apresentaram, em geral, bons estimadores. Comparando os dois métodos, verificámos que utilizando-se funções de detecção já incorporadas no programa DISTANCE, este se revelou superior, como seria de esperar. Quando utilizámos como função de detecção a função exponencial, obtivémos resultados melhores com LOGSPLINE. Isto dever-se-á ao facto da metodologia *logspline* revelar uma boa adaptabilidade, mesmo na presença de pontos angulosos (repare-se que neste caso tinhamos $g_2'(0) \neq 0$). A utilização de duas funções de detecção não paramétricas, com propriedades semelhantes, revelou uma maior robustez do modelo logspline, nestes casos, já que apresentou erros quadráticos médios idênticos havendo diferenças maiores aquando da aplicação do modelo de Buckland.

Para uma melhor avaliação da metodologia *logspline* procedemos a um conjunto mais alargado de simulações, utilizando-se 12 funções de detecção e populações com dimensões de 40, 80, 120, 250, 500 e 1000 indivíduos, repetindo-se o processo 1000 vezes para cada um dos casos num total de $1000 \times 12 \times 6$ realizações. Os resultados obtidos permitem concluir que os estimadores apresentam boas propriedades. Para além disso, podemos constatar que o erro quadrático médio não difere significativamente nas diferentes funções de detecção utilizadas, o que parece indicar uma boa robustez da metodologia.

No decorrer deste trabalho deparámo-nos com algumas questões de difícil resolução e interpretação no que diz respeito à metodologia *logspline*. Tal deveu-se, essencialmente, à pouca informação disponível sobre o assunto, sendo todo o trabalho apresentado nesta área baseado na leitura dos artigos referenciados já que não existe nenhuma obra de referência nesta área (está previsto o lançamento de

um livro sobre esta metodologia ainda no ano de 2003). Pensámos ter conseguido, no entanto, mostrar que se trata de um método bastante interessante e com potencial na estimação de densidades populacionais por transectos lineares.

Não conseguimos mostrar até que ponto o facto da metodologia logspline não impôr que a função de detecção seja estritamente decrescente afecta a qualidade das estimativas, embora nos dois exemplos práticos analisados esse facto não tenha, aparentemente, afectado a qualidade das estimativas. A imposição de que $f'(x) < 0$, para $x > 0$, pareceu-nos pouco tratável em termos teóricos, já traria profundas alterações a um método que se caracteriza pela a sua boa adaptabilidade aos mais diversos tipos de dados. De qualquer forma, pensamos que será um assunto interessante de abordar, futuramente.

A forma de calcular os intervalos de confiança na metodologia *logspline* pareceu-nos estar ainda numa forma bastante embrionária como é referido por Kooperberg e Stone (2003). Portanto, optámos por não incluir esses resultados nas simulações que realizámos. Esperamos poder fazê-lo numa próxima oportunidade.

Para finalizar, pensámos que seria conveniente que, futuramente, se desenvolvesse um software amigável, sob o ponto de vista do utilizador, que englobasse as diversas metodologias que apresentaram resultados competitivos na amostragem por transectos lineares, por forma a garantir uma melhor qualidade nas estimativas apresentadas para as densidades populacionais, uma medida cada vez mais importante para o estudo, conservação e preservação das espécies animais e vegetais.

Apêndice A

Estacas de madeira

Dados utilizados

[1] 2,02 0,45 10,40 3,61 0,92 1,00 3,40 2,90 8,16 6,47 5.66 2,95
[13] 3,96 0,09 11,82 14,23 2,44 1,61 31,31 6,50 8,27 4,85 1,47 18,60
[25] 0,41 0,20 0,40 11,59 3,17 7,10 10,71 3,86 6,05 6,42 3,79 15,24
[37] 3,47 3,05 7,93 18,15 10,05 4,41 1,27 13,72 6,25 3,59 9,04 7,68
[49] 4,89 9,10 3,25 8,49 6,08 0,40 9,33 0,53 1,23 1,67 4,63 3,12
[61] 3,05 6,60 4,40 4,97 3,17 7,67 18,16 4,08.

Output de LOGSPLINE

knots	A(1)/D(2)	loglik	AIC	minimum penalty	maximum penalty
3	2	-475.45	960.72	3.08	Inf
4	2	-473.91	962.56	0.24	3.08
5	2	-473.83	967.31	NA	NA
6	2	-473.74	972.04	NA	NA
7	2	-473.55	976.58	0.02	0.24
8	2	-473.54	981.47	0.02	0.02
9	2	-473.54	986.37	0.01	0.02
10	1	-473.53	991.27	0.00	0.01

The present optimal number of knots is 3

Penalty(AIC) was the default: BIC=log(samplesize): log(136)= 4.91

[1] 0.1122500

Output de DISTANCE

Hazard/Cosine

	Estimate	%CV	df	95% Confidence Interval

m	2.0000				
AIC	382.31				
Chi-p	0.79968				
f(0)	0.10637	12.01	66	0.83762E-01	0.13507
p	0.30027	12.01	66	0.23645	0.38130
ESW	9.4013	12.01	66	7.4033	11.939
D	36.165	17.07	66	25.785	50.725

Apêndice B

Ungulados africanos

Dados utilizados

[1] 0,00 0,00 0,00 0,00 0,00 0,00 0,00 0,00 8,72 10,50
[11] 22,30 26,00 26,00 30,50 30,50 31,70 34,20 35,10 38,00 41,00
[21] 42,10 50,80 55,10 58,50 63,60 64,30 65,00 68,80 71,10 71,80
[31] 71,90 72,10 73,10 76,60 77,60 78,10 84,50 84,50 86,00 86,00
[41] 87,00 90,00 92,30 94,00 96,40 96,40 106,00 115,00 123,00 123,00
[51] 129,00 129,00 143,00 150,00 151,00 157,00 153,00 161,00 164,00 272,00
[61] 378,00 400,00 164,00 164,00 166,00 175,00 188,00 193,00 200,00 200,00
[71] 246,00 260,00 143,00.

Output de LOGSPLINE

knots	A(1)/D(2)	loglik	AIC	minimum penalty	maximum penalty
3	2	-916.20	1842.37	3.03	Inf
4	2	-914.69	1844.33	0.45	3.03
5	2	-914.66	1849.26	NA	NA
6	2	-914.24	1853.39	0.41	0.45
7	2	-914.04	1857.98	NA	NA
8	2	-913.94	1862.76	NA	NA
9	2	-913.63	1867.12	0.09	0.41
10	1	-913.58	1872.02	0.00	0.09

the present optimal number of knots is 3

penalty(AIC) was the default: BIC=log(samplesize): log(146)= 4.98

[1] 0.006590021
Output de DISTANCE

Uniform/Cosine

	Estimate	%CV	df	95% Confidence Interval	
m	2.0000				
AIC	813.94				
Chi-p	0.98633E-01				
f(0)	0.63796E-02	8.72	71	0.53635E-02	0.75882E-02
p	0.39188	8.72	71	0.32946	0.46612
ESW	156.75	8.72	71	131.78	186.45
D	0.38809E-01	14.59	71	0.29055E-01	0.51838E-01

Apêndice C

Programas

C.1 Rotina para a gerar as populações, simular a amostragem, apresentar os resultados de LOGSPLINE e formatar as amostras de modo a serem utilizadas por DISTANCE.

```
logfunct <- function (n,m) {
    cat("Options;",file="quangexp.inp",append=TRUE,fill=TRUE)
    cat("Title='Fill in your title';",file="quangexp.inp",append=TRUE
    ,fill=TRUE)
    cat("Distance=Perp/Exact;",file="quangexp.inp",
    append=TRUE,fill=TRUE)
    cat("Object=Single;",file="quangexp.inp",append=TRUE,fill=TRUE)

        cat("Distance/Units = Meters;", file = "quangexp.inp",

    append=TRUE,fill=TRUE)
    cat(" Length/Units='Meters';",file="quangexp.inp",
    append=TRUE,fill=TRUE)
    cat("Area/Units='Sq. Meters';",file="quangexp.inp",
    append=TRUE,fill=TRUE)
    cat("End;",file="quangexp.inp",append=TRUE,fill=TRUE)
    cat("Data;",file="quangexp.inp",append=TRUE,fill=TRUE)
    set.seed(1)
    D<-(n/200)
    d<-c(1:m)
    v<-c(1:m)
    for (k in 1:m)
```

```
{
j<-1
t<- runif(n)
f<-c(1:1)
{for (i in 1:n)
{
h<-(exp(-t[i]*t[i]/(0.16*0.16*2)))
e<-runif(1)
if (h > e)
{f[j]<-t[i]
j<-j+1}
else
j<-j
g<-c(-f,f)
}
cat("Stratum /Label=",k,";",file="quangexp.inp",append=TRUE,
fill=TRUE)
cat("Sample /Effort=100;",file="quangexp.inp",append=TRUE,
fill=TRUE)
write.table(f,file="quangexp.inp",append=TRUE,quote=FALSE,
sep=" ",row.names = FALSE,col.names = FALSE)
cat(";",file="quangexp.inp",append=TRUE,fill=TRUE)
write.table(f,file="dados.txt",append=TRUE,quote=FALSE,

                sep =, row.names = FALSE, col.names = FALSE)

}
fit<-logspline(g)
s<-2*dlogspline(0,fit)
w<-length(f)
d[k]<-(length(f)*s/(2*100))
v[k]<-w
}
write.table(v,file="dim.txt",append=TRUE,quote=FALSE,sep=" ",

                row.names = FALSE, col.names = FALSE)

cat("End;",file="quangexp.inp",append=TRUE,fill=TRUE)
cat("Estimate;",file="quangexp.inp",append=TRUE,fill=TRUE)
cat("PRINT / NO=Fxplot;",file="quangexp.inp",
append=TRUE,fill=TRUE)
cat("PRINT / NO=Fxfit;",file="quangexp.inp",
```

```
append=TRUE,fill=TRUE)
cat("PRINT / NO=Fxest;",file="quangexp.inp",
append=TRUE,fill=TRUE)
cat("PRINT / NO=Fxtest;",file="quangexp.inp",
append=TRUE,fill=TRUE)
cat("PRINT / NO=Sbarest;",file="quangexp.inp",
append=TRUE,fill=TRUE)
cat("PRINT / NO=Sbarplot;",file="quangexp.inp",
append=TRUE,fill=TRUE)
cat("Estimator /Key=Uniform /Adjust=Polynomial;",file="quangexp.inp",
append=TRUE,
fill=TRUE)
cat("Estimator /Key=HNormal
```

$/Adjust = Hermite;", file = "quangexp.inp", append = TRUE, fill = TRUE)$

```
cat("Estimator /Key=Hazard
```

$/Adjust = Cosine;", file = "quangexp.inp", append = TRUE, fill = TRUE)$

```
cat("Estimator /Key=Uniform
```

$/Adjust = Cosine;", file = "quangexp.inp", append = TRUE, fill = TRUE)$

```
cat("Estimator /Key=HNormal
```

$/Adjust = Cosine;", file = "quangexp.inp", append = TRUE, fill = TRUE)$

```
cat("DENSITY by stratum;",file="quangexp.inp",
append=TRUE,fill=TRUE)
cat("End;",file="quangexp.inp",append=TRUE,fill=TRUE)
write.table(d,file="densidade.txt",append=TRUE,quote=FALSE,sep=" ",
row.names = FALSE,col.names = FALSE)
meand<-mean(d)
biasd<-meand-D
rbiasd<-(biasd/D)*100
vard<-((1/(m-1))*sum((d-meand)^2))
sed<-sqrt(vard/m)
rse<-(sed/D)*100
msed<-vard+((biasd)^2)
rmsed<-sqrt(msed)
```

```
    lrrmse<-(sqrt((1/m)*sum(((d-meand)/D)^2)))
    lnmed<-mean(v)
    lndp<-(sqrt((sum((v-lnmed)^2))/m))
    return(meand,biasd,rbiasd,vard,sed,rse,msed,rmsed,lrrmse,
    lnmed,lndp)
    }
```

C.2 Rotina para apresentar os resultados relativos a DISTANCE.

```
luis <- function (n,m) {
    d<-scan("densdistance.txt")
    D<-(n/200)
    meand<-mean(d)
    biasd<-meand-D
    rbiasd<-(biasd/D)*100
    vard<-((1/(m-1))*sum((d-meand)^2))
    sed<-sqrt(vard/m)
    rse<-(sed/D)*100
    msed<-vard+((biasd)^2)
    rmsed<-sqrt(msed)
    lrrmse<-(sqrt((1/m)*sum(((d-meand)/D)^2)))
    return(meand,biasd,rbiasd,vard,sed,rse,msed,rmsed,lrrmse)
    }
```

C.3 Rotina para apresentar os intervalos de confiança para a metodologia logspline.

```
bootstrap <- function (n) {
    y<-scan("hemingway.txt")
    x<-c(1:73)
    t<-c(1:15)
    v<-c(0,0)
    a<-c(-y,y)
    fut<- logspline(a)
    b<- dlogspline(0,fut)
    D<-73*2*b/(2*60)
```

```
set.seed(0)
for (j in 1:15) {
for (i in 1:73) {
k<-round(runif(1,1,73))
x[i]<-y[k]}
u<-c(-x,x)
fit<- logspline(u)
z<- dlogspline(0,fit)
r[j]<-73*2*z/(2*60000)
t[j]<-2*z
}
med<- mean(t)
dp<- sqrt(sum((t-med)^2))/length(t)
v[1]<-2*b-1.96*dp
v[2]<-2*b+1.96*dp
amp<-v[2]-v[1]
return (2*b,v,amp,med,dp,t)
}
```

Bibliografia

[1] Alpizar-Jara, R. & Pollock, K. H. (1996). A combination line transect and capture recapture sampling model for multiple observers in aerial surveys. *Environmental and Ecological Statistics* **3**, 311-327.

[2] Alpizar-Jara, R. (1997). *Assessing assumption violation in line transect sampling.* Phd Thesis. North Carolina State University, Raleigh NC.

[3] Barabesi, L. (2000). Local likelihood density estimation in line transect sampling. *Environmetrics* **11**, 413-422.

[4] Barabesi, L., Greco, L. & Naddeo, S (2002). Density estimation in line transect sampling with grouped data by least squares. *Environmetrics* **13**, 167-176.

[5] De Boor, C. (1978). *A practical guide to splines.* Springer Verlag, New York.

[6] Borchers, D., Buckland, S. T., Goedhart, P., Clarke E. & Hedley S. (1998). Horvitz Thompson estimators for double-platform line transects surveys. *Biometrics* **54**, 1221-1237.

[7] Buckland, S.T. (1985). Perpendicular distance models for line transect sampling. *Biometrics* **41**, 177-195.

[8] Buckland, S. T. (1992a). Fitting density functions with polinomials. *Applied Statistics* **41**, 63-76.

[9] Buckland, (1992b). Maximum likelihood fitting of the Hermite and simple polynomial densities. *Applied Statistics.***41**, 241–266.

[10] Buckland, S. T.& Turnock, B. J. (1992). A robust line transect method. *Biometrics* **48**, 901–909

[11] Buckland, S. T., Anderson, D., Burnham, K. P. & Laake J. L. (1993). *Distance sampling: estimating animal abundance of biological populations.* Chapman and Hall, London.

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

[13] Buckland, S. T., Goodie, I. & Borchers, D. (2000). wildlife population assessment: past developments and future directions. *Biometrics* **56**, 1-12.

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

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

[16] Burnham, K. & Anderson, D. (1992). Data-based selection of an appropriate biological model: the key to modern data analyses. Pages 16-30 in D. R. Mc Cullough, & R. H. Barrett (eds) wildlife 2001: *populations.* Elsevier Sci. Publ.. Ltd., London

[17] Casella, G. & Berger, R. (2002). *Statistical inference.* Duxbury, Pacific Grove, California.

[18] Chen, S. X. (1995). Studying school size effects in line transect sampling using the kernel method. *Biometrics* **52**, 1283-1294.

[19] Chen, S. X. (1996). A Kernel estimate for the density of a biological population by using line transect sampling. *Appl. Statitics* **45**, 135-150.

[20] Chen, S. X. (1998). Measurement errors in line transect sampling. *Biometrics* **54**, 899-908.

[21] Chen, S. X. (2000). Animal abundance estimation in independent line transect surveys. *Environmental and ecological statistics* **7**, 285-289.

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

[23] Davison, A. C. & Hinkley, D. V. (1997). *Bootstarp methods and their aplication.* Cambridge University press, Cambridge.

[24] Dias, R. (2002). A review of non-parametric curve estimation with aplication to econometrics. *Economia.* Vol. **3**, n°1.

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

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

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

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

[29] Hansen, M.H. & Kooperberg, C. (2002). *spline* adaptation in extended linear models. Statistical Science **17**, 2-51.

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

[31] Ihaca, R. & Gentleman, R. (1996). A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, vol 3 n°5, 299-314.

[32] Karunamuni, R. J. & Quinn II, T.J. (1995). Baysian estimation of animal abundance for line transect sampling. *Biometrics* **51**, 1325-1337.

[33] Koo, J. Y., Kooperberg, C. (2000). *logspline* density estimation for binned data. *Statistics & Probability Letters.* **46**, 133-147.

[34] Koo, J.Y., Kooperberg, C & Park, J.(1999). *logspline* density estimation under Censoring and Truncation. The Scandinavian J. of Statistics **26**, 87-105.

[35] Kooperberg, C. & Stone, C. J. (1991). A study of *logspline* density estimation. *Computational statistics & data analysis* **12**, 327-347.

[36] Kooperberg, C. & Stone C. J. (1992). *logspline* density estimation for censored data. *J. Comput. Graphical statistics* **1**, 321-347.

[37] Kooperberg, C. & Stone, C. J. (2001). *logspline* density estimation with free knots. *manuscript.*

[38] Kooperberg, C. & Stone, C. J. (2002). Confidence intervals for *logspline* density estimation. *Proceedings of the MSRI workshop on Nonlinear Estimation and Classification.* Springer: New York, 285-296.

[39] Kooperberg, C. & Stone, C. J. (2003). Comparison of parametric and bootstraps, and bayesian approaches to obtaining confidence intervals for *logspline* density estimation. *Journal of graphical and computational statistics,* to appear.

[40] Laake, J. (1978). *Line transect sampling estimators robust to animal movement.* Msc Thesis. Utah State University, Logan.

[41] Mack I. P., Quang P. X. & Zhang S. (1998). Kernel estimation in transect sampling without the shoulder condition. *Communications in Statistics: Theory and Methods* 1999. Vol **28**(10).

[42] Marques, T. (2002). *Amostragem por distâncias em populaçõ̃es naturais.* Tese de Mestrado. D.E.I.O. Fac. de Ciências da Universidade de Lisboa. Lisboa.

[43] Melville, G. J. & Welsh, A. H. (2001). Line transect sampling in small regions. *Biometrics* **57**, 1130-1137.

[44] Nice, M. M. & Nice L. B. (1921). The roadside census. *Wilson bulletin* **33**, 113-123.

[45] Pina, H (1995). *Métodos numéricos.* Mc Graw-Hill, Portugal.

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

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

[48] Quang, P. X.(1991). A nonparametric approach to size-biased line transect sampling. *Biometrics* **47**, 269-279.

[49] Quang, P. X. & Lanctot, R. B.(1991). A line transect model for aerial surveys. *Biometrics* **47**, 1089-1102.

[50] Quinn II, T. (1985). Line transect estimators for school populations. *Fisheries Research* **3**, 183-189.

[51] Ramsey, F. L. (1979). Parametric models for line transect surveys. *Biometrika* **66**, 505-512.

[52] Ramsey, F.L., Wildman, V. & Engbring, J. (1987). Covariate adjustments to effective area in variable-area wildlife surveys. *Biometrics* **43**, 1-11.

[53] Sakamoto, Y., Ishiguro, M. & Kitagawa, G.(1986). *Akaike information criterion statistics.* KTK Scientific Publishers, Japan.

[54] Schumaker, L. L. (1993). *spline functions: basic theory.* Wiley, New York.

[55] Seber, G. A. F. (1973). *The estimation of animal abundance and related parameters.* New York: Hafner.

[56] Stoer, J. & Bulirsch, R. (1996). *Introduction to Numerical Analysis.* Springer-Verlag, New York.

[57] Stone, C. J. (1990). Large-sample inference for log-*spline* models. *The Annals of Statistics* **18**, 717-741.

[58] Stone, C. J. & Koo, C. Y. (1986a). Additive *splines* in statistics. *1985 Statistical computing section proc. amer. statist. assoc.* 45-48. Academic, Boston.

[59] Stone, C. J. & Koo, C. Y. (1986b). *logspline* density estimation. AMS *Contemporary Mathematics Ser.***29**, 1-15. Amer. Math. Soc., Providence.

[60] Stone, C. J, Hansen, M. H., Kooperberg, C. & Truong,Y. K. (1997). Polynomial *splines* and their tensors products in extended linear models. *The Annals of Statistics* **25,** n°4, 1371-1470.

[61] Takada, T. (2001). Nonparametric density estimation: a comparative study. *Economics Bulletin*, Vol 3, No 3, 1-10.