INCORPORATING MEASUREMENT ERROR AND DENSITY GRADIENTS IN DISTANCE SAMPLING SURVEYS

Tiago André Marques

Thesis submitted for the degree of
DOCTOR OF PHILOSOPHY
in the School of Mathematics and Statistics
UNIVERSITY OF ST ANDREWS
ST ANDREWS
APRIL 2007

Abstract

Distance sampling is one of the most commonly used methods for estimating density and abundance. Conventional methods are based on the distances of detected animals from the center of point transects or the center line of line transects. These distances are used to model a detection function: the probability of detecting an animal, given its distance from the line or point. The probability of detecting an animal in the covered area is given by the mean value of the detection function with respect to the available distances to be detected. Given this probability, a Horvitz-Thompson-like estimator of abundance for the covered area follows, hence using a model-based framework. Inferences for the wider survey region are justified using the survey design.

Conventional distance sampling methods are based on a set of assumptions. In this thesis I present results that extend distance sampling on two fronts.

Firstly, estimators are derived for situations in which there is measurement error in the distances. These estimators use information about the measurement error in two ways: (1) a biased estimator based on the contaminated distances is multiplied by an appropriate correction factor, which is a function of the errors (PDF approach), and (2) cast into a likelihood framework that allows parameter estimation in the presence of measurement error (likelihood approach).

Secondly, methods are developed that relax the conventional assumption that the distribution of animals is independent of distance from the lines or points (usually guaranteed by appropriate survey design). In particular, the new methods deal with the case where animal density gradients are caused by the use of non-random sampler allocation, for example transects placed along linear features such as roads or streams. This is dealt with separately for line and point transects, and at a later stage an approach for combining the two is presented.

A considerable number of simulations and example analysis illustrate the performance of the proposed methods.

Declarations

I, Tiago André Lamas Oliveira Marques, hereby certify that this thesis, which is approximately 40000 words in length, has been written by me, that it is the record of work carried out by me and that it has not been submitted in any previous application for a higher degree.

date:__________ signature of candidate:_________________

I was admitted as a research student in February 2003 and as a candidate for the degree of Doctor of Philosophy in Statistics in February 2003; the higher study for which this is a record was carried out in the University of St Andrews between 2003 and 2007.

date:__________ signature of candidate:_________________

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of Doctor of Philosophy in Statistics in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree.

date:__________ signature of supervisor:_________________

In submitting this thesis to the University of St Andrews I understand that I am giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. I also understand that the title and abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that my thesis will be electronically accessible for personal or research use, and that the library has the right to migrate my thesis into new electronic forms as required to ensure continued access to the thesis. I have obtained any third-party copyright permissions that may be required in order to allow such access and migration.

date:__________ signature of candidate:__________________

Acknowledgements

I have no doubts that very few people have been as lucky as I have with respect to research environment. This work is the result of my interaction with many people since I first arrived to St Andrews in August 2001. It would not be fair if I did not acknowledged them here.

- Steve Buckland, my supervisor, provided all the support I could need during this PhD. He was always available and ready to explore new ideas, and his suggestions were always good (even if often I could only grasp the full extent of these a while later). The response to my e-mails in periods I was away from St Andrews was always so fast that it was like having real time replies, which made all the difference. I also thank Steve for the opportunity to teach in workshops and participate in several other projects, which contributed decisively to my formation as a scientist. When I doubted about my own work, Steve was always there to defend it.

- Dinis Pestana was my supervisor in Portugal, to whom I owe the incentive to actually come to Scotland for the PhD. He has been a supportive friend since the beginning of my explorations of statistics, and I hope he finds some reward in this work, for the endless times he wasted explaining me simple things like Taylor series expansions, random variable transformations and Jacobians.

- David Borchers had a large input in the development of the chapter about measurement error, and many of the ideas are his or matured after discussions with him. Later along the thesis I have participated in the analysis of some hares data sets, supervised by David, which resulted in the development of the extensions to the methods of chapters 5 and 6. I am grateful for the constant support when I felt that I would not be able to finish.

- Len Thomas, my second office mate, had a strong influence in my work, helping out constantly, discussing ideas, helping to fix problems in Distance or my own code, and providing interesting subjects for additional work. Almost at the end, Len provided a very thorough review of the thesis, with endless useful comments.

- Eric Rexstad, a true provider of food for thought. The first person I shared an office in CREEM, while he was in a Sabbatical from Fairbanks. He helped me to understand better what is a PhD, how people should look at it, and what are the important things. During many long discussions about everything and nothing he contributed to my formation as a scientist - never having sit on one of his classes, one of the best teachers I had. Additional thanks for the proofreading of this work.

- Jeff Laake provided comments and discussion over several issues related to density gradients and their integration with *MRDS* methods. I am only sorry I was not able to actually implement these for the thesis.

- Russel Alpizar-Jara's PhD thesis was my first contact with distance sampling, and he was the external examiner on my MSc thesis. He encouraged me to come to St Andrews, at the very early stages of this work. For that, and the discussions we had about distance sampling whenever we met in congresses/conferences, I am thankful.

- To Rhona Rodgers and Phil Le Feuvre I thank the support given with all the little details related with bureaucracy and computer issues. When others take care of these for you it becomes much easier to concentrate on your own work. And when those "others" treat you as a friend it is even better.

- I thank Jon Aars and Mads Peter Heide-Jørgensen, who through CREEM and RUWPA, allowed me to participate respectively on a polar bear and a whale/seal distance sampling survey, which contributed decisively to my understanding of what distance sampling is. I believe the perspective gained from field experience cannot be overstated when trying to develop new methods which in the end are to be used in the field.

- To Jon Bishop and Ciara Brewer, my last office mates at CREEM, for the company and distraction while at work, and to Jon for the road trips in particular.

- To Charles Paxton, house mate and partner in crime, with whom I had many discussions about work, and also about life in general. To the true scientist-gentleman, thanks for all the lessons and the thorough proof reading of this thesis.

- Nelson Silva, Filipa Samarra, Bruno Caneco and Liliana Martins were the Portuguese friends that helped to keep Portugal close while in Scotland.

- All the people in Distance Workshops, an amazing source of inspiration about potential issues to work on through all their questions and data sets.

- "Cathy" Brown provided, in the early days of the PhD, the right support for being away from home. In the same way, I thank Camila Dixon, Sharon Hedley and Fiona Underwood for their friendship and for realizing that the hardest part could be being away from home.

- To all the above and the remaining "creeminals", especially Arliss Winship, Laura Marshall, Louise Burt, Carl Donovan, Monique Mackenzie, Olivier Gimenez, Ruth King, Chris Lynam, Ken Newman, Joanne Potts, Filipa Samarra, Catriona Stephenson, Martin Cox (a special thanks for the use of the room at The Rummond!), Leslie New, Lindesay Scott-Hayward and Rodrigo Wiff for all the amazing coffee (and non-coffee!) times.

- A special thank you to my "UK foster family", Carl Donovan and Monique Mackenzie, without whom I doubt this would have ever been a reality. Their importance to this work is directly proportional to how unimportant they think they were. The home, the car, the food, the fast rides to trains and planes, the discussions about work and the opportunities to get away from work were fundamental to me.

- Silvio Velosa was a good friend which helped me continuously during my MSc degree and without whom I would most likely not have finished the MSc thesis nor even begun this one.

- Nicky Haxell, from the Student Support Services, was very helpful and provided useful advice when things seemed to get out of control.

- To all those friends which have been around these last few years, either in Portugal or scattered around the world, providing the "environment-friendly" support for this work: Filipe, Luis e Ana Luisa, Patricia e Tiago, Marta e Tiago, Luis e Teresa, Luzia, Nuno e Sónia, Camacho e Gonçalo, Branquinho, Filipa, Sofia e Tiago, Sofia Morais, Luis e Maria.

- To all my family, most especially to my grandmother Gabriela, who has always been there for me, helping without conditions, and without whom's efforts this most likely would not have happened. Thanks for the lessons of life.
- A special thank you to Luisa, grandmother of my grandchildren, probably the person which had more to loose from this PhD, for allowing me to be away for 2 out of the 4 preceding years, and for providing both personally and through phone and internet all the support I could need, as well as for all the suggestions and corrections to this work. I hope in the long term this work might be considered a good investment. Also to the still unborn Filipe, which provided the final incentive to get this thesis out of the way and get on with the rest of my life.

Goes without saying, but I say it nonetheless. Despite the direct contribution of everyone else, the remaining mistakes are my sole responsibility.

Data acknowledgements

- I thank Miguel Bernal for the use of the golf tee data set, used in the measurement error chapter to illustrate the PDF approach, which Miguel collected as part of his MSc.
- I thank Thorvaldur Gunnlaughsson and Gisli Vikingsson for allowing the use of the data from the 1987 aerial minke whale survey, used in the measurement error chapter to illustrate the likelihood approach. The data was collected by the Marine Research Institute, Iceland.
- I would like to thank Quercus, Queen's University Belfast, for the opportunity to use the hares data sets, mostly in the density gradient related work. Robbie McDonald facilitated access to the data set. The key people involved in data collection were David Tosh for the Northern Ireland data, in 2004 and 2005, and Neil Reid for the Republic of Ireland data, in 2006. The data would have not been collected without the financial support of the Environment & Heritage

Service for the Northern Ireland data and National Parks and Wildlife Service for the Republic of Ireland data.

Institutional funding acknowledgements

- Preliminary work, to prepare the submission of an application for the PhD, was made possible through a 2002 grant from the educational services of Fundação Calouste Gulbenkian.

- During the PhD I had the financial support of Fundação para a Ciência e Tecnologia and Fundo Social Europeu, III quadro comunitário de apoio, through their grant number SFRH/BD/10812/2002.

- The Centro de Estatística e Aplicações da Universidade de Lisboa provided some additional financial support to participate in conferences.

Table of Contents

List of Figures

List of Tables

Notation

The following notation is listed here for an easy reference, being used throughout this work.

Known constants or data

- N - total number of animals in a survey region (also referred as abundance);
- A - the area of the survey region, which represents the total area we are interested in drawing inferences about;
- D - the density of animals, given by $\frac{N}{A}$;
- a - the area covered by samplers (usually $a \ll A$);
- N_c - total number of animals in the covered area;
- n - number of detected animals in the covered area; if multiple independent samples are used, a subscript is used to distinguish them (e.g. n_p and n_s respectively for data coming from primary and secondary transects and n_l and n_p for data from lines and points);
- L - total length of transect surveyed;
- k - total number of points surveyed;
- x - perpendicular distance (line transects);
- r - radial distance (point transects);
- v - a generic distance, either radial or perpendicular;

- w - truncation distance (distances larger than w are not used at the analysis stage).

Parameters and functions

- $g(v)$ - the detection function, representing the probability of detecting an animal, given the animal is at distance v from the transect (perpendicular distance x for lines and radial distance r for points). Unless otherwise stated it is assumed that $g(0) = 1$;

- $g(0)$ - the detection function evaluated at 0 distance;

- $\underline{\phi_1}$ - parameter vector associated with the detection function;

- $f(v)$ - the probability density function (*pdf*) of the detected distances;

- $f(0)$ - the *pdf* of the detected distances, evaluated at distance 0;

- $h(0)$ - the slope of the *pdf* of the detected distances, evaluated at distance 0;

- P_c - probability of an animal being in the covered area, given it is in the survey region (usually known by design);

- P - probability of detection (given that the animal is in the covered area);

- $Pr(S)$ - probability of event S;

- $\pi(v)$ - the *pdf* of distances v available for detection; note that v can either be a perpendicular (x) or a radial (r) distance;

- $\underline{\gamma}$ - parameter vector associated with a measurement error model;

- $D(x)$ - the density at a distance x from the transect, referred to as absolute density gradient;

- $d(x)$ - a relative density gradient function, proportional to $D(x)$, but defined as a being a *pdf*;

- α - the proportionality constant that relates the relative and absolute density gradient, $D(x) = \alpha d(x)$;

- $\underline{\phi_2}$ - parameter vector associated with the density gradient;
- σ - the scale parameter of a model (either normal, half-normal, gamma or hazard-rate);
- b - the shape parameter of a model (gamma or hazard rate);
- μ - effective strip (half-)width;
- ρ - effective radius;
- ν - effective area of detection ($= \pi\rho^2$);
- K - correction factor in the PDF approach. Where appropriate, K_l is used for the case of lines and K_p for the case of points;
- U - a beta or uniform random variable;
- $\mathcal{L}$ - represents a likelihood. l represents a log-likelihood;
- $\mathbb{E}(W)$ - the mean value of the random variable W.

General abbreviations

- AIC - Akaike Information Criterion. ΔAIC is the difference between the AIC of a given model and the lowest AIC in the set of models being compared;
- $APTA$ - the availability proportional to area condition;
- CI - confidence interval;
- CDS - conventional distance sampling;
- CV - coefficient of variation (usually in percentage);
- GOF - goodness-of-fit;
- GPS - global positioning system (always used referring to a particular device);
- HN - half-normal;
- HR - hazard rate;

- HTL - Horvitz-Thompson-like (estimators);
- $MCDS$ - multiple covariate distance sampling;
- MLE - maximum likelihood estimator;
- $MRDS$ - mark recapture distance sampling;
- PDF - the approach to deal with measurement error using a correction factor;
- *pdf* - probability density function.

Special cases

An estimator for a given parameter of interest α is represented by $\hat{\alpha}$.

For the measurement error chapter only, a true distance is X, a contaminated or observed distance is Y, and the error is R (i.e., $Y = XR$ for a multiplicative error model).

Other notation, with restricted rather than general use, is not listed here, and only defined when necessary.

Chapter 1

Introduction

1.1 Overview

The raising impacts of human activities on natural resources, with corresponding dramatic losses of biodiversity all over the world, has made clear that is fundamental to obtain a better understanding of the way that wildlife, habitat and humans interact, to model and predict distribution and abundance of animal populations and hence to be able provide sound advice on ways to prevent and mitigate the effects of human impacts.

With that purpose in mind, arguably the most important question one can ask about any given population of interest is “How many are there?”. Despite being the most important question, it is nonetheless a very difficult one, and over the years an impressive quantity of literature has been devoted to this question. Seber (1986, 1992) and Schwarz and Seber (1999) present comprehensive literature reviews on this subject.

For most real life surveys, it is not possible to count every animal in the population of interest. Therefore, to make inferences about population size, one must rely on the information contained in a sample, in which only a fraction of the population is observed. Then, using either design or model based approaches, or a combination of these, one can draw inferences with respect to the entire population of interest. This inference is usually conditional on a set of assumptions, which need to be clearly stated and evaluated for the exercise to be valid.

Over the years, a large number of different methods have been developed with this goal in mind. The choice of an appropriate method in a given situation should be a function of the species and habitat characteristics, resources available, desired precision and objectives of the study. Background knowledge on the environment and the population becomes therefore invaluable in the choice of the method to use.

Distance sampling is one among several other possible alternatives, such as plot sampling, mark-recapture studies or removal, catch-effort and change-in-ratio approaches (e.g. Borchers *et al.*, 2002), and arguably it is the most frequently used when the main goal is abundance or density estimation. It is especially relevant for animal populations distributed over large areas with low to medium density, and it has not only been used to estimate density and abundance of species belonging to all major animal groups, but also for many plant populations and fungi, in all types of habitats and environmental conditions. Hence, although for simplicity I usually refer to animals, distance sampling methods are not in any way restricted to the analysis of animal populations. A large number of examples are cited throughout this thesis, and an extensive (although not complete) reference list of distance sampling methods and applications is available online at http://www.creem.st-and.ac.uk/tiago/webpages/distancesamplingreferences.html.

The fundamental idea behind the methods is that the distances to detected animals can be used to estimate the probability of detecting an animal. The way the probability of detection is obtained is through the modelling of a detection function, which represents the probability of detecting an animal, conditional on it being at a given distance from the transect. The number of animals detected divided by the unconditional probability of detecting an animal in the covered area is an estimator of the number of animals in the covered area.

Compared to other contending methods for animal abundance estimation, distance sampling is a relatively new technique, with the first ideas appearing around the 1960's and the first key reference, Burnham *et al.* (1980), being less than 30 years old. It is therefore not surprising that the last few years have witnessed many new developments in this area, like the use of covariates, in addition to distance, to model the detection function (e.g. Beavers and Ramsey, 1998; Marques, 2001) or the incorporation of spatial models in the process (e.g. Hedley, 2000).

In the remaining of this introductory chapter I present an intuitive overview of distance sampling methods, followed by some insights on the personal motivation behind this thesis, and finishing with a brief description of the remaining chapters of this thesis.

1.2 Intuitive principles of distance sampling

If an estimate of abundance is needed, then the methods used must, except for the rare cases where one can be sure of detecting all animals, account for detectability. Some authors argue that if focus is on trend, then it might be enough to collect an index of abundance (e.g. Hutto and Young, 2002, 2003). However, the assumption of constant detectability, or at least no long-term trend in detection ratios (see Bart *et al.*, 2004), needed to interpret a relative index as a true index of abundance, is rarely likely to hold. Most commonly this is at best a leap of faith based on untested assumptions. Hence the use of index methods, like raw counts, has been criticized in the literature (e.g. Anderson, 2001, 2003; Pollock *et al.*, 2002; Rosenstock *et al.*, 2002; Thompson, 2002; Ellingson and Lukacs, 2003; Norvell *et al.*, 2003). Even raw counts might contain useful information, and given a sufficiently large effect is present, a simple index might be enough to detect it. However it is also clear that in any such scenario true abundances should be able to detect even smaller effects, and hence should, in my opinion, always be recommended. On the other hand, although a comparison across time or space might be based on an index, sometimes one is required to obtain absolute estimates. Managing a population of an endangered species or setting quotas for harvesting usually requires more than an index of abundance. The often expressed view that methods that estimate detectability rely on assumptions that do not hold in practice is not constructive, because it fails to recognize that index methods rely in even further demanding, and often unstated, assumptions.

Distance sampling is one of several ways to account for detectability, and although the modelling involved might be complicated, the estimators are intuitive in nature. In this section I describe how these estimators can be derived in an intuitive, non-mathematical way.

1.2.1 From total counts to distance sampling

Usually there is a clear distinction between the survey region A, the entire area over which we are interested in making inferences, and the covered area a, the area that we actually sample, with $a < A$. Estimation procedures are usually separated into two steps: (1) estimating abundance in the covered area (N_c) and (2) the scaling up of that estimate to the wider survey region (N). Under the conventional setting, this second stage relies on the covered area being a representative sample of the wider survey region, hence validating the inferences by design. Because we rely on a (detection) model for the first stage and on the properties of the design for the second, Fewster and Buckland (2004, p. 286) refer to distance sampling neither as model or design based, defining it as a composite approach.

Provided an abundance estimate is available, valid for a given area, a natural estimator of density is obtained by dividing the abundance estimate by the corresponding area. Hence, in the following I usually refer to abundance estimators, although the corresponding density estimators follow easily.

Under very restricted circumstances, it might be possible to count all the animals in a given study area. In such case, there is no sampling involved, and no estimation takes place. The abundance is by definition the total number of counted animals.

In most situations, it is not possible to cover the entire survey region, and a set of sampling units covering the fraction $\frac{a}{A}$ of the survey region are surveyed. The allocation of these samplers should follow a random design. The reader is referred to Strindberg *et al.* (2004) for further details about survey design in the distance sampling context. Assuming that all animals in the covered area are detected, an estimator of animal abundance in the covered area is simply the number of detected animals, n, and hence an estimator of total number of animals in the survey region is given by

$$\widehat{N} = \frac{n}{P_c} \tag{1.1}$$

where, given a randomized design, P_c=a/A is the proportion of the survey region covered, which needs not to be estimated as it is known by design. Note that this estimator is unbiased only if the randomized design is one with equal coverage probability.

However, this assumes that all animals in the covered areas are detected. If, as in most applications, we are likely to miss some animals in the covered area, then an estimator of animal abundance in the covered area is given by

$$\hat{N}_c = \frac{n}{\hat{P}} \tag{1.2}$$

where $\hat{P}$ represents the estimated probability of detecting an animal, given that it is in the covered area. From this to an estimator of abundance in the survey region the reasoning is analogous to the certain detection case, and hence

$$\hat{N} = \frac{\hat{N}_c}{P_c} = \frac{n}{\hat{P}P_c}. \tag{1.3}$$

Note that $\hat{N}$ is an intuitive estimator. Consider the product PP_c to represent the inclusion probability, i.e. the probability of an animal being included in the sample. Given that we detect 20 animals in an area, and we know that the inclusion probability of an animal is 0.25, the intuitive estimate of 80 animals for the area follows. This estimator can be seen as a Horvitz-Thompson-like (HTL) estimator, in the sense that the inclusion probabilities are not solely given by design but also estimated based on the data (e.g. Borchers *et al.*, 1998a). Hence, unlike the traditional Horvitz-Thompson estimator (e.g. Thompson, 1992, p. 49), these abundance estimators are not necessarily unbiased, since even unbiased estimates of P do not warrant unbiased abundance estimates, because $\mathbb{E}[1/\hat{P}]$ is not the same as $1/\mathbb{E}(\hat{P})$. In general, distance sampling estimators can be shown to be HTL estimators (see Borchers *et al.*, 2002).

Distance sampling gives us a well established framework to estimate the P in equation 1.2 and hence abundance or density. Note that the innovation with respect to traditional sampling is the estimation of abundance in the covered area, as the scaling up to the entire survey region is usually done based on the sampling design properties, just as in traditional plot sampling.

1.2.2 Estimating probability of detection

The samplers used in most distance sampling applications are strips (line transects) or circles (point transects). The distances from the lines or points to the detected animals are used to model a detection function, $g(v)$. This function represents the probability of detecting an animal, given that it is at a distance v from the line (perpendicular distance) or point (radial distance). Note that v represents either a perpendicular or radial distance; when the distinction is useful, x is used for the former and r for the latter. A related function is $f(v)$, the *pdf* of the detected distances, which as is shown below is directly related to P.

Considering for the moment just line transects, if all the animals up to a truncation distance w were seen, then a histogram of the detected distances should be approximately uniform (Figure 1.1a), because the area available at a given distance from the line is constant and independent from the distance itself. However, since the probability of detection decreases with increasing distance, the detected distances histogram usually tends to look more like figure 1.1b. Graphically, the area above the curve, inside the rectangle of figure 1.1(c), corresponds to the probability of not detecting an animal. Hence, the probability of detecting an animal is the area under the *pdf* divided by the area of the rectangle, $f(0) \times w$, but since, by the *pdf*'s own definition, the area under it is 1, P is given by

$$P = \frac{1}{f(0)w}. \tag{1.4}$$

Define μ as the effective strip (half-)width, in the sense that had all the animals at distances shorter than μ been seen, we would see (on average) the same number of animals as were seen in the actual survey. Then areas A and B (see Figure 1.1c) would be the same if the vertical line that divided them is at μ, which leads to the result $P = \frac{\mu}{w}$, since the area to the left of the vertical line is the same as the area under the function.

In the case where samplers are point transects, a similar argument is valid. If all the animals up to a truncation distance w were seen, then a histogram of the detected distances would increase linearly (Figure 1.1d), because the area available increases

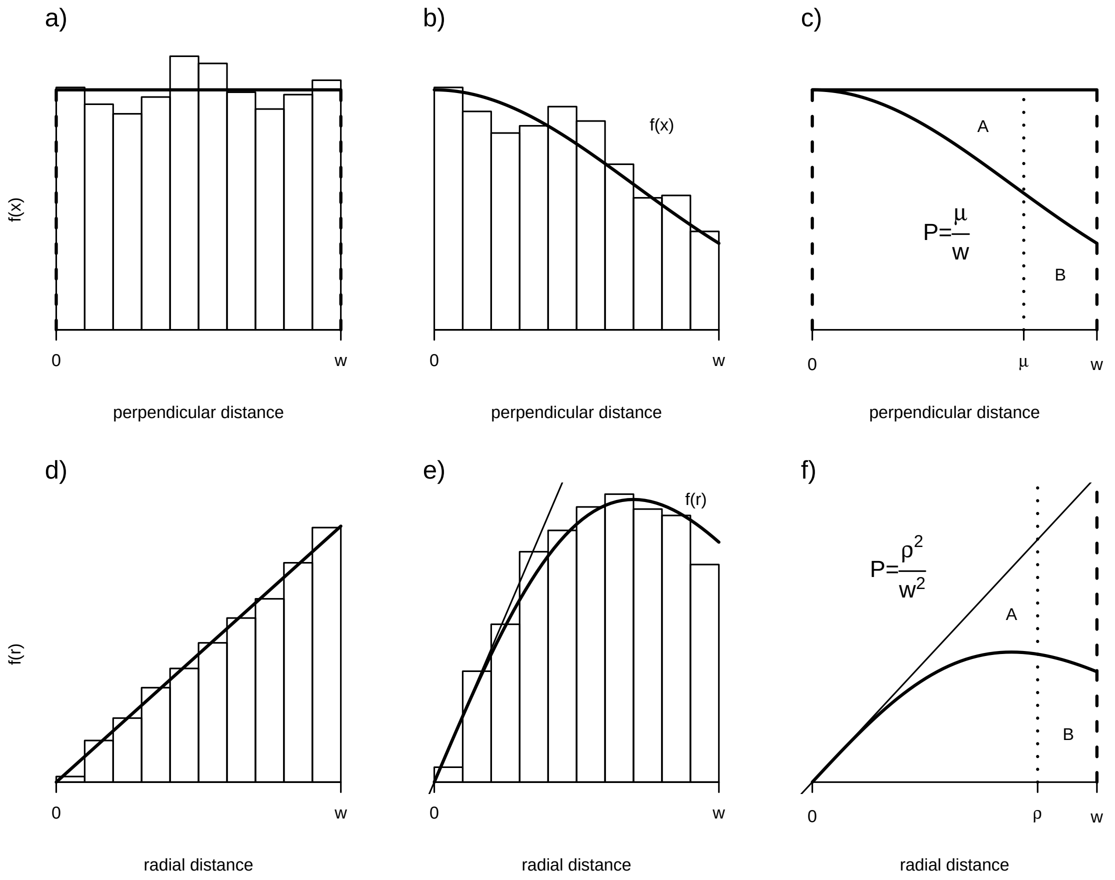

Figure 1.1: Intuitive rationale behind the derivation of probability of detection P (top row refers to line transects and bottom row to point transects). a) and d) Distances available for detection; b) and e) Detection function and detected distances; c) Relation between μ and P; f) Relation between ρ and P.

linearly as a function of the distance to the point. However, since the probability of detection decreases with increasing distance, the detected distances histogram usually looks more like figure 1.1e. The diagonal line shown is the tangent to the function at $r = 0$, with slope $h(0)$. As before, we can think of the animals missed as those above the *pdf*. The probability of detecting an animal is then the area under the *pdf* divided by the area of the triangle, and hence

$$P = \frac{1}{\frac{wwh(0)}{2}} = \frac{2}{w^2 h(0)} \tag{1.5}$$

where $h(0)$ represents $\left.\frac{d\ f(r)}{dr}\right|_{r=0}$, i.e. the value of the derivative of $f(r)$ evaluated at distance 0.

Define ρ as the effective radius, in the sense that had all the animals until ρ been seen, (on average) we would see the same number of animals as were seen in the actual survey. Then areas A and B (see Figure 1.1f) would be the same if the vertical line that divided them is at ρ, which (given the fact that the area of a circle with radius η is $\pi\eta^2$) leads to the result $P = \frac{\rho^2}{w^2}$, since the area to the left of the vertical line is the same as the area under the function.

This shows that, both for points and lines, the probability of detection, and hence the corresponding abundance estimators involved, can be related to the *pdf* of detected distances, and hence what is at stake for conventional methods is estimating a *pdf*, and evaluate it or its first derivative at distance 0. This means that in this context we can rely on the large statistical toolbox available to estimate *pdf*'s to obtain the desired abundance estimators.

In chapter 2 these estimators are derived in a mathematical way, and other useful forms for the same estimators are presented.

1.3 Personal motivation

Any given work is the result of some personal motivation. In my case, a biological background and a number of suspicions about the proper way to use statistics in biology led me to the field of ecological statistics.

I believe one of the main tasks for people that work with distance sampling, as well as with any statistical method with direct application in biology, is not only to extend existing methods and derive new ones, but to contribute to the spread of adequate use of such methods by the actual practitioners. Teaching in distance sampling workshops, biologists and ecologists oriented talks, tutorial-like papers as Marques *et al.* (in press), my participation in current software Distance project (Thomas *et al.*, 2005) or maintenance of bibliography lists and tutorial-like material available through my personal web page are the reflection of this attitude. I further hope that the work documented in this thesis can to some extent help others to acknowledge that the use of statistical methods should be done only after their assumptions, and implications of assumption failure, are fully understood.

The issues addressed in this thesis are the result of interacting with people that use the methods in practice, and the development of each of the new methods is driven by the wish that they might become useful to the people that inspired them.

1.4 Thesis outline

As any other method for estimating animal abundance, distance sampling estimators are derived under a number of assumptions that ensure the methods are asymptotically unbiased. Under certain circumstances, these assumptions do not hold, and the conventional methods may become severely biased.

Conventional distance sampling assumes $g(0) = 1$, no undetected responsive movement or random movement of animals[1] and no measurement error in the distances. Because a specific form for the distribution of available distances with respect to samplers is used in the derivation of conventional estimators, we can also view this known distribution as an assumption. Conventional methods can be extended by (1) estimating $g(0)$ - hence allowing it to be less than 1, by (2) incorporating a model of animal movement, by (3) incorporating a model for errors in measured distances or by (4) incorporating a model for the distribution of animals with respect to distance from the line or point. The 1st case has been studied extensively (e.g. Borchers *et al.*, 1998a,b; Laake and Borchers, 2004) and the 2nd is hard to tackle due to the difficulty in collecting information on the animal's movement (but see Smith, 1979; Turnock and Quinn, 1991; Buckland and Turnock, 1992). The 3rd and the 4th cases are the subject of this thesis.

After this introductory chapter, chapter 2 is dedicated to the theory and assumptions of conventional distance sampling, laying down material needed for subsequent chapters and setting the scene for the reminder of the thesis.

It is followed by a chapter about the effects of measurement error in the detected distances used for estimating the detection function, and ways to correct for the bias induced by such errors (Chapter 3). The material in this chapter is largely taken from Marques (2004) and Borchers *et al.* (in prep a), with some ideas also taken from Burnham *et al.* (2004) and Marques *et al.* (2006).

Chapter 4 introduces the issues arising due to non random allocation of samplers. Special emphasis is given to the case where sampling takes place along linear features, like roads or rivers, a situation discouraged but often used due to logistic constraints (or just bad practice). With standard plot sampling, the main issue is whether these

[1] In practice no responsive movement and slow movement relative to observer speed is usually enough.

areas can be a representative sample that allows inference for a wider survey region. I show that in addition to that, in the case of distance sampling, a further complication arises due to the potential presence of a density gradient[2] with respect to the linear features involved, and hence to the samplers themselves. Therefore, there is the need to account for that density gradient at the analysis stage.

Chapter 5 and 6 deal with approaches to correct for the bias induced by the presence of a density gradient with respect to samplers. In chapter 5 the use of line transects is considered, while in chapter 6 the focus is on the use of points transects. In the proposed setting these need to be considered separately because, while the information collected from point transects can be used to draw inferences about both the detection and availability (for detection) processes, we need to collect independent information to do so in the case of line transects.

In chapter 7 the material of chapters 5 and 6 is extended and combined to deal with particularities of a given data set, showing how once cast in a likelihood framework the methods can be easily modified to deal with data particularities.

Each chapter with methods development contains its own discussion section, but a short final chapter serves as wrap up and integrating discussion of the key ideas in the thesis, pointing to some loose ends and potential directions for further research in this area.

[2] The density gradient is a function which describes density as a function of the distance from the linear feature. It should not be confounded with the strict mathematical term of gradient (as a rate or slope), although the function does describe the rate of change in density as a function of distance from the linear feature, and hence the choice of wording.

Chapter 2

Background information - theory, practice and assumptions

2.1 Introduction

The first thorough reference to distance sampling, presenting the methods in an integrated perspective, was the monograph by Burnham *et al.* (1980). After that, the book by Buckland *et al.* (1993a) laid out what is now referred to as conventional distance sampling (CDS). This 1993 book has been updated, with currently conventional methods covered by Buckland *et al.* (2001) and more advanced methods by Buckland *et al.* (2004). The above references are the key sources for the material in this chapter.

Most of the statistical research related to distance sampling in the last decade addresses situations in which conventional methods are not an option, focussing on assumption violations and ways to deal with them.

In this chapter the general theory of CDS, needed for later chapters, is presented. I start by presenting the derivation of conventional estimators (section 2.2) and the available strategies to obtain the respective variances (section 2.3). After some comments about the industry standard distance sampling software Distance (section 2.4) I present CDS key assumptions, and main developments, dealing with overcoming their failure, are reviewed (section 2.5). The final section 2.6 describes what as been called the underlying uniformity assumption of distance sampling (e.g. by Melville and Welsh, 2001) and its implications for the analysis of distance sampling data.

2.2 Deriving conventional distance sampling estimators

In this section I derive the CDS density estimators (for the covered area), first for line and then for point transects, finishing with a subsection that shows that results for points and lines are analogous, the difference being due to the geometry of the problem.

2.2.1 Line transects

Consider that the covered area is composed of a number of line transects, with total length L and truncation distance w, conveniently allocated in a larger survey region according to some random sampling design. Note that usually, for analysis purposes, distances are folded along the line transect, i.e. distances in $(-w, w)$ are mapped into $(0, w)$. Note however that exceptions exist under which it is simpler to use the signed distances (e.g. Mack and Quang, 1998).

The probability that a detected animal is at a perpendicular distance x from the line is given by

$$f(x)dx = Pr(\text{animal in } x, x + dx|\text{animal detected}) \tag{2.1}$$

$$= \frac{Pr(\text{animal detected}|\text{animal in } x,\ x\text{+ } dx)Pr(\text{animal in } x,\ x\text{+ } dx)}{P} \tag{2.2}$$

$$= \frac{g(x)\frac{2L\ dx}{2Lw}}{P} \tag{2.3}$$

where $g(x)$ is the detection function, indexed by an unknown parameter vector $\underline{\phi_1}$, and P represents the probability of detecting an animal. Therefore,

$$f(x) = \frac{g(x)}{wP}, \quad 0 < x < w \tag{2.4}$$

and because $f(x)$ is a *pdf*, integration of this expression leads to

$$P = \mu/w \tag{2.5}$$

where $\mu = \int_0^w g(x)dx$. Using this result in equation 1.2 leads to the following density estimator for the covered area a

$$\begin{aligned}
\hat{D} &= \frac{\hat{N}}{a} && (2.6)\\
&= \frac{\frac{n}{\hat{P}}}{a} && (2.7)\\
&= \frac{\frac{n}{\frac{\hat{\mu}}{w}}}{2Lw} && (2.8)\\
&= \frac{n}{2L\hat{\mu}}. && (2.9)
\end{aligned}$$

This last equation justifies why μ is referred to as the effective strip (half-)width: during the survey, you see on average as many animals as you would have seen if all the animals to a distance μ were seen. This estimator has another useful form. From equation 2.4 it follows that

$$f(x) = \frac{g(x)}{\mu}. \tag{2.10}$$

Considering distance sampling first assumption, $g(0) = 1$, leads to $f(0) = 1/\mu$, and hence another expression for the above estimator is

$$\hat{D} = \frac{n\hat{f}(0)}{2L}. \tag{2.11}$$

Therefore, given an estimate of the *pdf* of the detected distances, evaluated at 0, we can get an estimate of density.

2.2.2 Point transects

Assume that the covered area is now composed of k point transects with a radius w, conveniently allocated in a larger survey region accordingly to some random sampling design. Consider the probability that an observed animal is at a radial distance r from the center of a point transect,

$$f(r)dr = Pr(\text{animal in } r, r + dr|\text{animal detected}) \tag{2.12}$$

$$= \frac{Pr(\text{animal detected}|\text{animal in } r,\ r\text{+ } dr)Pr(\text{animal in } r,\ r\text{+ } dr)}{P} \tag{2.13}$$

$$= \frac{g(r)\frac{2\pi r dr}{\pi w^2}}{P} \tag{2.14}$$

where $g(r)$ is the detection function, indexed by an unknown parameter vector $\underline{\phi_1}$. Therefore,

$$f(r) = \frac{g(r)\frac{2\pi r}{\pi w^2}}{P}, \quad 0 < r < w \tag{2.15}$$

and $f(r)$ being a *pdf*, integration of both sides of this equation leads to

$$\pi w^2 P = \int\limits_0^w 2\pi r g(r)dr = \nu \tag{2.16}$$

where ν is sometimes referred as the effective area of detection: on average you see as many animals as you would see if all the animals in the area ν were seen. Analogous to line transect's effective strip (half-)width (μ), the effective radius of detection is $\rho = \sqrt{\frac{\nu}{\pi}}$. The use of this terminology is due to the fact that, from equations 1.2 and 2.16, a density estimator in the covered area, $a = k\pi w^2$, is given by

$$\begin{aligned} \hat{D} &= \frac{n}{k\pi w^2 \frac{\hat{\nu}}{\pi w^2}} && (2.17) \\ &= \frac{n}{k\hat{\nu}} && (2.18) \\ &= \frac{n}{k\pi\hat{\rho}^2}. && (2.19) \end{aligned}$$

Hence an estimator for D can be readily obtained if we can derive an estimator for ν. Rearranging equations 2.15 and 2.16 leads to

$$\nu = \frac{2\pi r g(r)}{f(r)} \tag{2.20}$$

or the alternative

$$\frac{2\pi}{\nu} = \frac{f(r)}{rg(r)}. \tag{2.21}$$

Using distance sampling first assumption, $g(0) = 1$, leads to

$$\lim_{r\to 0} \frac{f(r)}{rg(r)} = f'(0) = h(0) \tag{2.22}$$

where $h(0)$ represents $\frac{d\ f(r)}{dr}\Big|_{r=0}$. The most common form of the estimator for density is finally obtained by using equation 2.22 in equation 2.18, leading to

$$\hat{D} = \frac{n\widehat{h}(0)}{2\pi k}. \tag{2.23}$$

Therefore, given an estimate of the slope at 0 of the *pdf* of the detected distances, we can get an estimate of density.

2.2.3 Bringing points and lines together

From the two previous subsections, it is clear that the derivation of density estimators for line and point transects follows exactly the same rationale. The only difference lies in the way $g(v)$ and $f(v)$ relate, since for lines they are proportional, while for points they are not. The reason for this stems from the geometry of the problem. In the case of lines, the area available at a given distance from the line is constant, while for points it increases linearly with distance, which leads to differences in the *pdf*'s of distances available for detection. While for lines the cumulative distribution function of perpendicular distances, detected or not, is given by

$$F(x) = Pr(X \leq x) = \frac{2Lx}{2Lw}, \quad 0 \leq x \leq w \tag{2.24}$$

which leads to the *pdf*

$$\pi(x) = \frac{dF(x)}{dx} = \frac{1}{w}, \quad 0 \leq x \leq w \tag{2.25}$$

for the case of points it is given by

$$F(r) = Pr(R \leq r) = \frac{2\pi r^2}{2\pi w^2} = \frac{r^2}{w^2}, \quad 0 \leq r \leq w \tag{2.26}$$

which leads to the *pdf*

$$\pi(r) = \frac{dF(r)}{dr} = \frac{2r}{w^2}, \quad 0 \leq x \leq w. \tag{2.27}$$

In both cases, the estimator of abundance is obtained by using an appropriate estimator for P in expression 1.2. Note that P, being the probability of detecting an animal (unconditional on its actual position in the covered area), can be seen as the mean value of the detection function, where the mean value is evaluated with respect to the distribution of v, the distances available for detection, irrespective of whether they are detected or not. So, in general,

$$P = \int_0^w g(v)\pi(v)dv = \mathbb{E}[g(v)]. \tag{2.28}$$

Substituting $\pi(v)$ by the appropriate distribution for the case of points and lines (respectively equations 2.25 and 2.27, which we assume known by design, given a sufficient number of samplers randomly allocated in the study area) leads to an alternative way to derive the density estimators in equations 2.11 and 2.23. So, for the case of lines

$$P = \int_0^w g(x)\frac{1}{w}dx = \frac{\mu}{w} \tag{2.29}$$

while for the case of points

$$P = \int\limits_0^w g(r)\frac{2r}{w^2}dr = \frac{\nu}{\pi w^2} = \frac{\rho^2}{w^2} \tag{2.30}$$

therefore leading to the same estimators obtained before.

This is important as it shows that under a given setting, provided we can derive an estimate of P, we can derive a HTL estimator of abundance of the general form $\hat{N} = \frac{n}{\hat{P}}$. This will be useful in later chapters in situations were the conventional methods do not apply, and equations 2.11 and 2.23 are not valid, because the distribution of distances available for detection cannot be assumed known by design.

2.3 Variance estimation

An estimator is only useful if one can also assess its precision. For the simpler settings, distance sampling estimators have analytical variance estimators, and these along with corresponding confidence intervals can be easily obtained in software Distance (Thomas *et al.*, 2005). For more complicated scenarios, like the ones presented in subsequent chapters, the use of resampling strategies might be a more straightforward approach.

2.3.1 Analytical variance estimators

From equations 2.11 and 2.23 it is clear that the CDS estimator has two random components, namely one due to detection function estimation, and one due to encounter rate (i.e., the expected number of animals per unit effort) estimation.

In the standard analytical variance estimators, these two components are estimated separately, and then combined using the delta method to obtain an approximation for the variance. Using the delta method (e.g. Seber, 1982, p. 7-9), the variance of T, the quotient of two independent random variables, W and Z, is approximated by

$$var(T) = var\left(\frac{W}{Z}\right) \simeq T^2\{CV^2(W) + CV^2(Z)\}. \tag{2.31}$$

As an example, this expression applied to line transects leads to variance for the density estimator in equation 2.11 to be estimated by

$$var(\hat{D}) \simeq \hat{D}^2\{\hat{CV}^2(n) + \hat{CV}^2(f(0))\}. \tag{2.32}$$

The variance in n is typically estimated using empirical estimators on the counts from replicate samplers (lines or points), while the $f(0)$ variance is obtained via maximum likelihood theory. Details on obtaining these variances can be found in (Buckland *et al.*, 2001, p. 62-64, 78-80). A recent paper by Fewster *et al.* (in review) presents an improvement to the way the encounter rate variance is estimated, useful in situations in which a density gradient throughout the area is expected.

This variance estimator is easily extended under more complex scenarios when the estimator includes further random components, e.g. if some multiplier M is used, such as an independently obtained estimate of $g(0)$, an estimate of mean cluster size, or any other bias correction factor (see section 3.3 for an example). The variance of the corresponding estimator ($\hat{D}_M = \hat{D}\hat{M}$) is just given by

$$var(\hat{D}_M) = var(\hat{D}\hat{M}) \simeq \hat{D}_M^2\{\hat{CV}^2(n) + \hat{CV}^2(f(0)) + \hat{CV}^2(M)\}. \tag{2.33}$$

2.3.2 Resampling variance estimators

As an alternative to analytical variance estimators, resampling strategies are useful and are the only option for some of the more complicated methods. Such resampling strategies will be the recommended way to obtain variance estimates and confidence intervals for all the methods in this thesis.

Typically the nonparametric bootstrap is the preferred option. The bootstrap relies on obtaining samples, with replacement, from the set of original sampling units, and then calculating the statistic of interest (say a density estimate) for each of these resamples. It can be shown that, under certain mild regularity conditions, (e.g. Davison and Hinkley, 1997, p. 37-44), the variance properties of such statistic over a large number of resamples approximates well the variance that would be observed across a large number of samples from the population of interest, hence providing reliable variance estimators for the quantity we are interested.

For distance sampling estimators, rather than resampling observations, we usually resample samplers (either line or point transects), as it is more realistic to expect these to be independent sampling units. Hence, a bootstrap procedure for a CDS estimator consists in repeating B times (where B should be large, say $B = 999$) the following algorithm: (1) obtain a resample, with replacement, of independent sampling units, either lines or points and (2) for said resample, calculate and store the statistic of interest. The variance of the B estimated values will be a good approximation for the variance of the estimator used.

A simple method to obtain bootstrap confidence intervals is the percentile method. Given the B bootstrap estimates of the quantity of interest, the limits of a $(1 - \alpha)\%$ confidence interval are given by the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ percentiles. Note the use of $B = 999$ is convenient because the corresponding 95% confidence interval limits correspond to the lowest 25th and largest 975th observations.

Under some settings, a parametric bootstrap might be preferred. However I did not find the need for such approach in this work (but see e.g. Hedley, 2000, p. 42-47, for such an example), and hence the reader is referred to e.g. Efron and Tibshirani (1993) for further details on the subject.

2.4 Using Distance software for analysis

As shown before, P can be related to the *pdf* of detected distances for both lines and points. Since there is a large body of theory devoted to the estimation of *pdf*'s, that is the way estimation is carried out for conventional methods. The estimation procedure in practice is done by using appropriate software to model the detected distances *pdf*, even if usually we refer to it as modelling the detection function. The current version of the only widely used software to analyze distance sampling data is Distance 5 (Thomas *et al.* 2005).

The modelling strategy available in this software is based on the key+series adjustments approach suggested by Buckland (1992), where the fit of a key model belonging to a set of parametric families is improved using series adjustment terms. In that sense, it can be seen as a semi-parametric approach. The key functions available in the software are the hazard-rate, half-normal, uniform and negative exponential, which for brevity I refer to as HN, HR, UNI and NE throughout this thesis, and the adjustment terms are the cosine series, simple and hermite polynomials, referred to as *cos*, *sp* and *hp*. A half normal with 1 and 2 cosine adjustments would be respectively referred to as $HNcos$ or $HNcos2$. The reader is referred to Buckland *et al.* (2001, p. 45-48) for further details about different models and series adjustments.

For a given analysis, a set of plausible candidate models are assumed for the detection function and the data are used to derive estimates of the model parameter vector $\underline{\phi_1}$. This is done by numerical maximization of the following likelihood:

$$\mathcal{L}(\underline{\phi_1}|\underline{v}) = \prod_{i=1}^{n} f(v_i) = \prod_{i=1}^{n} \frac{g(v_i)\pi(v_i)}{\int_0^w g(v)\pi(v)\ dv}. \tag{2.34}$$

The choice of w, the truncation distance, is necessary because the integrals involved are evaluated numerically, and hence the above integral's upper limit needs to be defined. If the user does not choose a specific distance, the software will use the maximum observed distance for w. Note that this procedure results in a small bias, especially for low sample sizes (T. Marques, unpublished data). Nonetheless, the choice of w has usually little impact in the final estimates.

The software includes a number of goodness-of-fit (GOF) statistics and model selection criteria to assist in the modelling exercise, as well as the ability to include different level stratification and dealing with size bias due to detection probability being a function of cluster size. Several different ways to estimate variance for these estimators are available, both based on empirical and resampling strategies. Advanced methods available in Distance 5 include Multiple Covariate Distance Sampling (*MCDS*, e.g. Marques and Buckland 2004) and Mark-Recapture Distance Sampling (*MRDS*, e.g. Laake and Borchers 2004), as well as a design engine that can be used to investigate the properties of different design strategies or lay down realizations of a given design (e.g. Strindberg *et al.* 2004), provided a digital map of the study area is available. Spatial Modelling methods (e.g. Hedley *et al.* 1999, Hedley 2000) are available in a beta version of Distance 6.

Any of the software's analysis engines[1] can be run directly from the Windows command line or from inside other software, like R (R Development Core Team 2006). This was helpful at several points during the preparation of this thesis as it allows for the fast analysis of a large number of data sets, generated under different scenarios, therefore making it easy to test the methods through simulation exercises.

2.5 Distance sampling assumptions

The validity and reliability of estimates derived from *CDS* methods requires that 3 key assumptions are respected. Along with these key assumptions there are 2 other, which are somewhat less important but mentioned here for completeness.

Provided that *a sufficient number of samplers are placed over the study area independently of animal distribution, following some previously specified random design*, the conventional estimators are asymptotically unbiased given that:

1. Animals on the transect line or at the point center are detected with probability 1 ($g(0) = 1$).

This is considered the fundamental assumption of *CDS*, and its failure will lead

[1] There are separate engines for different types of analysis; the reader is referred to the software's manual for further details.

directly to an underestimation of density. If $g(0) = g$, with $g < 1$, then density estimates will be biased down by a factor of g, i.e., $\mathbb{E}[\hat{D}] = gD$.

Note that two different phenomena can lead to $g(0) < 1$. A detectability or perception bias occurs when the animals are available to be detected, but the observer misses them anyway. An availability[2] bias occurs when the animals are not available to detection. If only the latter case is present, then estimates are unbiased for the portion of animals available for detection.

It is important to realize that there is no information on the distances collected in conventional methods (either the perpendicular or the radial distances) to allow for the detection of the failure of this assumption when analyzing the data. There is however some evidence that $g(0)$ is more likely to be 1 if the detection function appears to have a shoulder[3], while a spiked detection function might be an indication that $g(0) < 1$ (Burnham *et al.*, 2004, p. 350); note this is only true for perception bias.

Availability bias is more difficult to deal than perception bias, essentially because there are no data you can collect in the distance sampling survey itself that can provide information on its occurrence. If you never see a fraction of the animals (say, desert tortoise in their burrows), the only way to correct for such a bias is by using independent information on the proportion of the population available for detection, e.g. using radio-telemetry data.

[2] Note that throughout the thesis I also use the term availability when referring to the distances available to be detected, in terms of their distribution with respect to the samplers. However the purpose of the word should always be clear from the context.

[3] It is said that a detection function has a shoulder if detection is certain at the line or point and remains (at least almost) certain for some non-negligible distance. Note this non-negligible distance is not precisely defined here, but it is a function of the scale of the distances being measured. Conversely, a spiked detection function is one for which detection probability shows a steep drop even for small distances.

Violation of this assumption appears to be quite widespread. Even in field simulation studies, where one might imagine it should be possible to detect all objects on the line, $g(0)$ has been found to be less than 1 (e.g. Laake, 1978; Otto and Pollock, 1990; Anderson *et al.*, 2001). Where perception bias is the cause, it may be that alternative field methods can reduce or eliminate the problem. However, in some cases this will not be possible, and then field and analysis methods must be employed that enable the estimation of $g(0)$.

Most of the ways to deal with $g(0) < 1$ involve more than one observer searching for animals, and using some kind of mark-recapture approach to correct for the fraction of animals missed on the line by both, hence the general name of mark-recapture distance sampling ($MRDS$). The reader is referred to Laake and Borchers (2004, p. 108-111) for further references on the development of these methods.

2. There is no animal movement.

Although strictly the assumption is that there is no animal movement, or in other words the survey is considered to be a snapshot, the methods are applicable provided that there is no undetected movement in response to the observer, and that movement independent of the observer is slow compared to observer velocity.

Undetected responsive movement can lead to upward or downward bias. If the animals tend to be detected after movement towards the observer occurs, density is overestimated. If the undetected movement is away from the observer, then density is underestimated.

Random movement (with respect to the observer) leads to overestimation of density, because the distance to the animals tends to be underestimated. This effect arises from the fact that as an animal moves around, it is more likely to be detected when close to the observer. A line transect example is useful for illustration: consider a situation with an effective strip (half-)width of around 40 m. A fast moving animal with a hypothetical circular home range with center at 50 m from the transect, with a radius of 30 m, will most of the times be detected closer to 20 m, and rarely (if ever) closer to 80 m. So, although the distance we would like to record (on average) is 50 m, it will most likely be less than that due to the combined effects of movement and detection processes.

There is no simple way to address the failure of this assumption, and usually one hopes that the survey procedures are such that it is likely to hold to a reasonable extent. Buckland and Turnock (1992), based on work by Turnock and Quinn (1991), developed an alternative approach in the presence of responsive movement, but this approach requires multiple observers, in which at least some of the animals must be seen by one of the observation platforms before responsive movement. Palka and Hammond (2001) extended these methods for situations in which most of the animals started to move before being detected by any of the platforms.

3. Distance measurements are recorded without errors.

Given that the basic information used to derive the probability of detecting an animal are distance measurements, it is not surprising that the recordings must be accurate to derive unbiased estimates.

The effect of measurement error can be thought to be equivalent to that of undetected movement. The final position of an animal after movement might be interpreted as a distance measured with error, where the error results from the movement and not from the measuring process itself.

If a systematic error is present in the recording of distances, then underestimation of distances leads to overestimation of density, and vice versa. The same amount of error will be more influential in point transects than in line transects (e.g. Buckland *et al.*, 2001, p. 265).

Perhaps more surprising is the fact that bias in density/abundance estimates might arise even if the errors are unbiased. Chen (1998), using an additive model, showed that unbiased errors lead to underestimation of abundance, while Marques (2004), using a multiplicative model, showed that unbiased errors lead to overestimation. Therefore, the effect of errors is dependent on the characteristics of the errors.

Chapter 3 deals exclusively with the failure of this assumption, and presents ways to deal with that failure.

Although usually thought somewhat less important, two other assumptions can be considered:

4. Detections of animals are independent events.

Strictly speaking, this is also an assumption as the maximum likelihood methods used for fitting the detection function are based on an underlying model where the observations are assumed independent.

However, Buckland *et al.* (2001, p. 36) point out that the methods are extraordinarily robust to the failure of this assumption. Point estimates should be the same irrespective of whether this assumption holds or not, but variances might be underestimated if we wrongly assume independent detections. Nonetheless, considering robust variance estimators as usually recommended, based on empirical variances of encounter rate across transects, the practical impact of this assumption failure is usually minimal (Buckland *et al.*, 2001, p. 36).

If detections are not independent, additional care must be given to the interpretation of GOF measures, as they will be much more sensitive to this assumption failure than the density estimates themselves (Buckland *et al.*, 2001, p. 171). Because model selection tools also assume independent observations, there might be a tendency to over fit if observations are not independent.

5. The detection function has a shoulder.

The estimation procedure is unreliable if the detection function has no shoulder. Since we are primarily interested in estimating the value of the *pdf* (or its derivative, in the case of point transects) of observed distances, evaluated at 0, it is important that this function behaves smoothly near 0, and this is achieved by ensuring a shoulder on the detection function.

Note that this applies both to the true detection function as well as to the assumed model for the detection function. One does not expect steep functions at $x = 0$ to be likely to arise because the detection process is intuitively smooth near the transect. Only in extremely closed habitats would one expect to see all objects on the line and yet miss a large proportion close to the line. For this reason the negative exponential detection function, available in software Distance for historical reasons, should not be used in practice. If the best fit to the data results in a spiked detection function, it is usually a reflection of failures of one or more of the main assumptions of the methods, rather than a reflection of the true underlying detection process.

Given appropriate field methods, namely adequate search procedures, a shoulder should be present in the detection function. The wider the shoulder the more reliable the estimation process will be, and the smaller the influence of choosing different models in resulting the density estimates.

2.6 The availability proportional to area condition

There has been some confusion regarding the assumptions made about the spatial distribution of the population of interest. For example Seber (1982) states that the population should be distributed randomly in the study area, i.e., according to a Poisson process. This is not true for the methods outlined above, as the only requirement is that the population distribution in the study area is a stochastic process with rate D, with no need to assume an equality between the rate and variance of the process.

Given such a stochastic process, the random allocation of samplers, independently of the animal population distribution, ensures that the distances to animals in the covered area, with respect to samplers, are proportional to the area available for the animals to be in. Ignoring edge effects, this leads, on average, to a uniform distribution of distances in the case samplers are lines and a triangular distribution in the case of points (cf. Figure 1.1). I will refer to this as the availability proportional to area ($APTA$) condition.

The $APTA$ condition is only valid on average, and for a given realization of a sampling design (the actual location of the samplers) deviations from it might be observed. The larger the number of transects used, the more likely it is that the condition holds to a reasonable extent. From this it should be clear that a given realization of the design, coupled with the animals locations, might lead to an estimate that is far from the true value of the quantity being estimated. This is however true for all estimators: they can only be considered biased or unbiased on average, and no statements can be made about the bias from a single realization of the estimation process.

A good density estimator should be asymptotically unbiased, provided its underlying assumptions hold, and that is the case for CDS estimators, provided they are used under the appropriate settings. Some authors have raised the question of an unstated *uniformity assumption* of distance sampling, in the sense that the objects of interest are uniformly distributed in the two-dimensional space of interest (e.g. Melville and Welsh, 2001). These authors further noted that such an assumption can not be tested based on the conventional data alone (e.g. Welsh, 2002; Melville and Welsh, 2001) and presented simulations to show how badly things can go when it is violated. I note that this is not the case, and stems from a confusion between requirements on the population of interest and requirements on the survey design, because the $APTA$ condition holds on average, given adequate survey design. The uniformity is desired for the animals present in the covered area, after transect placement. No such requirement is made for the animal distribution in the areas not sampled and, furthermore, provided survey design is adequate, that uniformity is achieved (on average), especially so once the distances from the sample of transects are pooled together. Their argument was misleading, based on examples conditioning on extreme animal configurations, coupled with a design that was clearly not independent of the animal population itself, and hence was flawed to begin with. For further details see Fewster *et al.* (2005).

Thus, the real emphasis should lie on the previous statement that "*...a sufficient number of samplers are placed over the study area independently of animal distribution, following some previously specified random design...*". This could potentially be itself seen as an assumption, which provided it holds, leads to the $APTA$ condition.

It is nonetheless important to acknowledge that this is a somewhat vague statement, because what constitutes a sufficient number of samplers is not clearly stated, and it might be dependent on the animal population. A minimum number of 10-20 is usually suggested (e.g. Buckland *et al.*, 2001, p. 232). Note that, even if only one single transect is used, on average (over many realizations of the design), the design should be unbiased. Using more transects just means that we can be more confident that a single realization is adequate (because in a survey, there is only a realization of the process), in the sense that the underlying distribution of distances will be close to what would be expected under $APTA$.

Since often sampling is restricted to a small fraction of the entire region of interest, the above statement ensures not only the $APTA$ condition, but also that inferences made for the covered area are representative of the wider survey region, as well as more robust encounter rate variance estimates.

The methods in this thesis are intended to be useful when, due to either inadequate survey design (e.g. samplers along roads) or assumption violation (e.g. measurement error), the $APTA$ condition does not hold and in fact a density gradient is observed with respect to the samplers.

2.6.1 Edge effects

Note that even when a "sufficient" number of randomly located samplers are used, the $APTA$ condition might not be achieved due to edge effects. This is where some parts of the transects actually lie outside the study area, and hence even if animals exist outside the study area they are not considered[4].

The consequences of such edge effects are less important than one might expect, because modelling the joint effect of availability and detectability should still lead to unbiased density estimation provided models are used that are flexible enough (e.g. Buckland *et al.*, 2001, p. 216). The estimate of the probability of detecting each of the detected animals in the covered area will nonetheless be biased. The following line transect example should clarify these points. As seen in section 2.2.3 (cf. equation 2.28), in the absence of an edge effect, the probability of detecting each of the n detected animals is given by

$$P = \int_0^w g(x)\pi(x)dx = \int_0^w g(x)\frac{1}{w}dx = \frac{\mu}{w}, \tag{2.35}$$

[4] Note that if animals are available outside the study area, and densities are not expected to change drastically, we could just include animals detected outside the study area, hence avoiding the edge effect.

and hence as usual $N = n/P$.

Now imagine that an edge effect is present, hence, in the covered area, there are only $N^* < N$ animals. Let $e(x)$ represent the proportion of area available with respect to what would be available in the absence of edge effects. Further let $e(0) = 1$, reflecting the fact that the transect lines themselves are, by design, actually inside the study area. Consider that if there was no edge effect, n animals would be detected. In the presence of an edge effect, we detect $n^* < n$ animals, because some are not there to be detected to begin with.

In this case, the true probability of detecting each of the animals is again given by 2.28, and hence

$$P_{eT} = \int_0^w g(x)\pi(x)dx = \int_0^w g(x)\frac{e(x)}{\int_0^w e(x)dx}dx = \frac{\int_0^w g(x)e(x)dx}{h}. \tag{2.36}$$

However, since we are not taking account of the edge effects, we estimate that probability to be

$$P_{eE} = \frac{\int_0^w g^*(x)dx}{w} = \frac{\int_0^w g(x)e(x)dx}{w}, \tag{2.37}$$

where $g^*(x)$ represents a compound function of availability and detectability. Ignoring the edge effect we assume the detected distances shape to be due exclusively to the detection process, but it actually is also due to an availability process. So, we know that

$$\frac{n^*}{P_{eT}} = N^* = N\frac{\int_0^w e(x)dx}{w} = N\frac{h}{w}. \tag{2.38}$$

Note $\frac{h}{w}$ is the fraction of the area that you actually cover, with respect to the total area if there were no edge effects. This expression can be rearranged, solving for N, leading to

$$N = \frac{wn^*}{P_{eT}h} \tag{2.39}$$

and hence

$$N = \frac{n^*}{P_{eE}} \tag{2.40}$$

which shows that abundance is estimated with bias (true abundance in the area covered is N^*), but density is unbiased. Note that we really want to estimate N rather than N^* because the density we are trying to estimate is

$$D = \frac{N}{A} = \frac{N}{2Lw} = \frac{N^*}{2Lw\frac{h}{w}} = \frac{N^*}{2Lh}. \tag{2.41}$$

This holds as long as the models used are flexible enough so that the compound function of detectability and availability, and hence P_{eE}, can be estimated with no (or negligible) bias.

The bias in encounter rate (compared to what would be seen in the absence of an edge effect) is compensated with the bias in estimating the probability of detection. The edge effect leads to an underestimation of that probability (we think we miss more animals away from the line, while in fact they are not there to be seen to begin with), but to a proportional decrease on the encounter rate compared to the absence of edge effects, and these two balance each other out.

The key difference between this case and the one where a true density gradient exists is that in the case of edge effects, the quantity we wish to estimate is the density at the transects itself, i.e. at $x = 0$, while if a density gradient is present, the density on the transect is likely a biased estimate of density on the area surrounding it and hence of no interest on its own.

Chapter 3

Distance sampling with measurement error

3.1 Introduction

As pointed out in the previous chapter, a key assumption of distance sampling is the absence of measurement error[1] in the distances used to model the detection function. Until recently, the effect of measurement error was largely ignored, and field experiments to access the presence and magnitude of measurement errors were rarely carried out (but see Butterworth *et al.*, 1984; Øien and Schweder, 1992, for exceptions). In many studies distances are obtained through eyeball estimation (e.g. Kulbicki and Sarramegna, 1999; Heydon *et al.*, 2000; Baldi *et al.*, 2001; Biswas and Sankar, 2002; Skaug *et al.*, 2004; Bårdsen and Fox, 2006), which is prone to several types of measurement errors. On the other hand, some studies do not describe in adequate detail the methods used for distance measurement, which makes it impossible to determine to what extent measurements, and hence results, are reliable (e.g. Becker *et al.*, 1997; Brown and Boyce, 1998; Ashenafi *et al.*, 2005).

Compared with other key assumptions, the effect of violation of the assumption

[1] Due to the continuous nature of distances, any recorded distance has a rounding error associated with. Hence in the following I ignore the errors due to the discrete recording of an otherwise accurately measured continuous distance.

of no measurement errors has received considerably less attention. The International Whaling Commission promoted the first attempts to look at these issues, with experiments being carried to access accuracy of distance (and angle) estimates for minke whale surveys (e.g Butterworth *et al.*, 1984; Thompson and Hiby, 1985) using different measurement methods. The poor accuracy reported in those trials would represent potential for problems in the analysis of the data. DeJong and Emlen (1985) present an early assessment of the effect of measurement error in distance sampling, with field trials to estimate its characteristics, and simulations to assess their influence in the estimation of the detection function. Unfortunately, this work was almost completely ignored by subsequent authors, possibly due to some of the awkward conclusions presented. Hiby *et al.* (1989) presented the first attempt to incorporate a measurement error model in distance sampling methods. They developed a method to account for error measurement in cue-counting methods using grouped data, by incorporating a measurement error function, with unbiased multiplicative Gaussian errors, in the estimation of the detection function. Schweder (1996, 1997) also dealt with measurement errors in radial distances, in relation to cue based methods for estimating minke whale abundance. The simulated likelihood method of Schweder *et al.* (1999), considering hazard models (as described by Skaug and Schweder, 1999), has the flexibility to remove the bias due to measurement errors (as well as potentially other factors not included in the likelihood). Chen (1998) described the effect of additive errors in line transect surveys, and assuming some knowledge about the distribution of the errors, derived a corrected estimator based on the method of moments. This work was the first to describe that even unbiased errors might lead to biased estimators of density. Chen and Cowling (2001) generalized this approach to the case where errors are present both in distances and other covariates that affect the detection function, namely cluster size. The model proposed by Chen can be seen as a particular case of the model proposed by Alpizar-Jara (1997). This author, using an approach based on the SIMEX algorithm (Cook and Stefanski, 1994), also presented a way of correcting density estimates in the presence of errors.

The errors in distance estimation are dependent on the process generating the detections. Most commonly detections will be made visually or aurally by humans, with other methods like radar, acoustic or thermal devices still seldom used (although an increase in the use of such devices is expected in the near future). The magnitude of the errors could be a function of many factors, such as observer, background noise, animal density or animal orientation. Alldredge *et al.* (in press) present an assessment of the influence of such factors for aural detections in a controlled experiment.

Fortunately, recent years brought the development of new technological solutions, which have been used to assist in the evaluation of distances in the field. These include for example laser range finders (e.g. Diefenbach *et al.*, 2003; Hounsome *et al.*, 2005), precise *GPS* portable units (e.g. Chen, 1998; Marques *et al.*, 2006) and radar/sonar technology (e.g. Harmata *et al.*, 1999), as well as new measurement and recording methods (e.g. Gordon, 2001; Southwell *et al.*, 2002; Diefenbach *et al.*, 2005). It is therefore plausible to think that the failure of this assumption will become both less frequent and less influential in the future. Nonetheless, there are still many situations in which measurement error is bound to occur, as for example for cetacean surveys, where most accurate methods fail, or bird surveys where most animals are only heard and not seen, and hence some subjective way to determine those distances is involved. Therefore it is important to understand the impacts of errors in the detected distances and how they influence final density and abundance estimates, as well as to find methods that account for their effect.

In the next section I present the different types of measurement error relevant in distance sampling. This is followed by an approach to correct the bias induced by measurement error, based on a multiplicative error model, in section 3.3. Most of the material covered in this section comes from Marques (2004), but also from Burnham *et al.* (2004), and for consistency with this reference I refer to the methods in this section as the PDF approach. This is in contrast to the methods of section 3.4, based on Burnham *et al.* (2004) and Borchers *et al.* (in prep a) and referred to as the likelihood approach, in which a likelihood incorporating an error model along with the usual detectability model is proposed. Methods in both sections 3.3 and 3.4 are illustrated by simulation and a real life application example. After a brief section about the estimation of measurement error model parameters from data (section 3.5), section 3.6 is devoted to 2 special cases of measurement error: (1) heaping, a special case of measurement error characteristic of distance sampling and (2) errors exclusively at large distances. This is followed by a final discussion section, where the methods presented are compared and some general conclusions drawn.

At first sight, the structure of this chapter might look cumbersome. The reason for this is that I decided to keep the two different approaches to deal with measurement error separate, to reflect the historical order in which these were developed. First the PDF approach and only afterwards the likelihood approach. Looking back on it, it seems unlikely that the PDF approach would have been developed had the likelihood approach arisen first. Nonetheless, there was a considerable part of this thesis time put into the development of the PDF approach.

In the following it is assumed that, except for measurement error, the usual distance sampling assumptions hold. For readability, detectability is considered to be a function of distance alone, but the methods might be extended to multiple covariate distance sampling (see Borchers *et al.*, in prep a, which extend the relevant formulation to include covariates).

3.2 Types of distance sampling measurement error

A large body of literature has been devoted to the study of measurement error in general regression models. Hence, we can place the measurement errors found in distance sampling in the context of other broader classes of measurement error. Using the definition of Carroll *et al.* (1995, p. 16), errors in distance sampling are nondifferential, in the sense that given the true distance, there is no information about the detection process in the error distances. On the other hand, we usually are interested in structural models, in the sense that parametric assumptions are made about the true distances distribution (Carroll *et al.*, 1995, p. 6).

Given a set of true detected distances (Fig. 3.1a), a number of situations can occur. Usually some measurement error is bound to be present, hence the no measurement error case (Fig. 3.1b) is unlikely. We can conceptually think of errors has having a random and a systematic component. In the case of unbiased errors (Fig. 3.1c and 3.1d), only the random component is present, i.e. the mean value of the distance measured (with error) is the true distance. The observed errors in real data sets might also result from a combination of random and systematic components (Fig. 3.1e). While under some conditions we might expect an error without systematic component (Fig. 3.1c and 3.1d), there is no reason to expect a purely systematic error (Fig. 3.1f), and hence I do not address such scenario in this work; however, if that was the case, a simple calibration experiment would solve the problem (e.g. Alpizar-Jara, 1997).

Although conceivably all models are wrong, it is useful to think of errors as being the result of either an additive or multiplicative model. An additive model for measurement error can be represented as

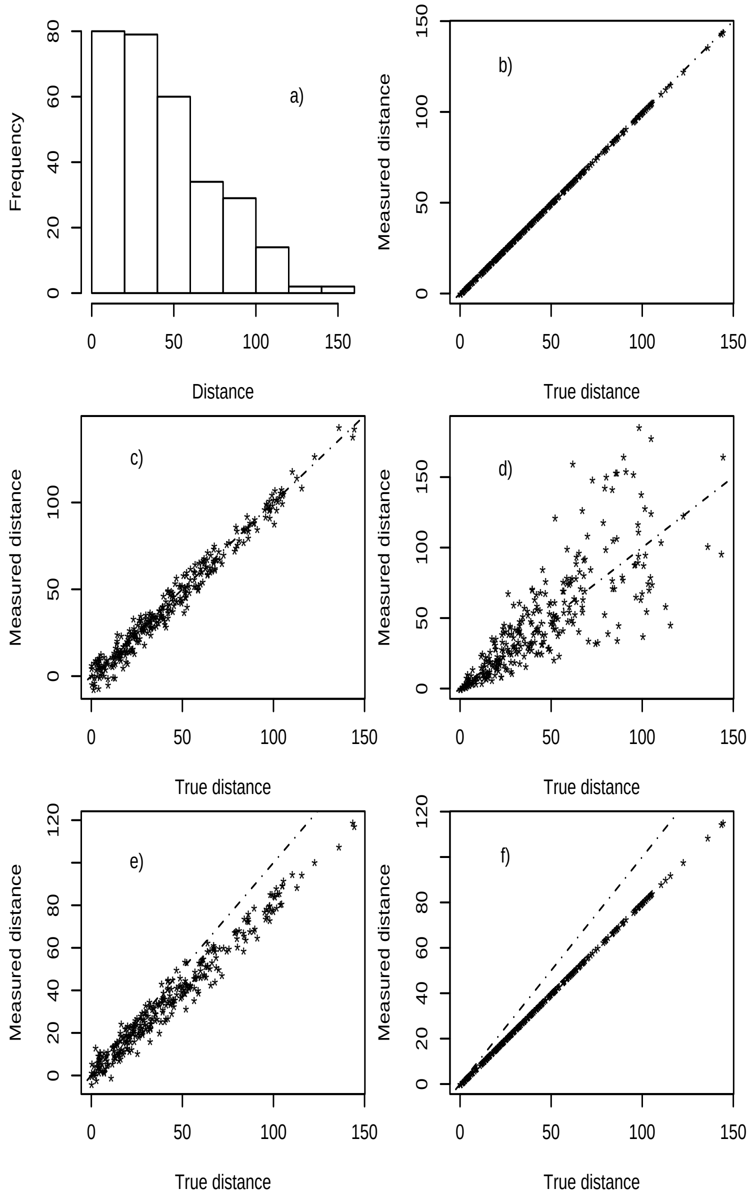

Figure 3.1: Different types of measurement error in distance sampling. a) A set of detected distances without error. b) to f) True versus measured distances for different measurement errors: b) No measurement error; c) Random only (additive) measurement error; d) Random only (multiplicative) measurement error; e) Random (additive) and systematic measurement error; f) Systematic only measurement error.

$$Y = X + R, \tag{3.1}$$

where X represents the true distance, Y the observed (contaminated) distance and R is the error, here assumed independent of X. The measured distances in figure 3.1c were generated using such a model.

Another plausible representation is a multiplicative model,

$$Y = XR \tag{3.2}$$

with X and R also independent. The measured distances in figure 3.1d were generated using such model. For simplicity I assume $X, Y \geq 0$, leading also to $R \geq 0$. This is sensible as negative distances are not possible for point transects, and for line transects one usually folds the process over the line, leading to positive distances only.

We usually think of *error* as the difference between a true and contaminated value. In fact, any error model can also be seen as an additive model. In particular, for the multiplicative error model above, expression 3.2 can be rewritten as

$$Y = X + R' \tag{3.3}$$

where $R' = X(R - 1)$, where the error is no longer independent of X. Hence, for the multiplicative error models used, the term *error* could also be used for representing R'.

The method used to measure the distances, and the corresponding associated errors, will usually be better represented by one of these models. If the measurement is dependent on the precision inherent to the measuring device, then the resulting error is likely to be additive in nature. An example of this is described by Chen

(1998), where a GPS unit is used, and the error is related to the GPS precision and should therefore be independent of X. If we consider a procedure where the error is likely to increase at larger distances, as with eyeball estimates, a multiplicative model might be more adequate. In this thesis the focus is on multiplicative error models, and I will be referring to these unless stated otherwise. I believe that under most distance sampling scenarios one would expect that the standard deviation in measurements increases as the true distance increases.

Throughout this chapter the terms *true distance*, *contaminated distance* and *error* refer respectively to X, Y and R. For consistency with the notation in Marques (2004), X is used for distances of both point and line transects. Note that this is in contrast with the notation used in the other chapters of this thesis.

3.3 Correcting the effect of measurement error - the PDF approach

In the following I derive a corrected estimator in the presence of measurement error, by assuming a multiplicative error model as in section 3.2.

The corrected estimator is obtained by multiplying a biased density estimator and an appropriate correction factor. While the biased density estimator is obtained as usual, based on the contaminated distances, the correction factor is derived as a function of the error distribution characteristics.

The methods are presented for the line transect case first, based on Marques (2004), and then extended to the point transect case, as in Burnham *et al.* (2004).

3.3.1 Line transects

3.3.1.1 Deriving corrected estimators

As described in section 2.2, in the case of line transects, we are interested in obtaining the following density estimator

$$\hat{D} = \frac{n\hat{f}_X(0)}{2L}. \tag{3.4}$$

If measurement errors are present, yet ignored, the conventional estimator is a biased estimator of density $(\hat{D}_y)$. In practice, we do not observe the X's, but the Y's, hence we effectively estimate density as

$$\hat{D}_y = \frac{n\ \hat{f}_Y(0)}{2L}. \tag{3.5}$$

The advantage of using a multiplicative error model is that, as I show below, the values of Y and X *pdf*'s, evaluated at distance 0, are proportional, and the proportionality constant is a function of the error structure. Hence, given knowledge about the error structure, we can correct an estimate based on the contaminated distances.

The distribution of the contaminated distances can be expressed by integrating, with respect to the error, the joint distribution of the contaminated distances and the errors, as

$$f_Y(Y) = \int_0^\infty f_{R,Y}(r, y)dr. \tag{3.6}$$

This expression can be rearranged, and based on standard random variable transformation, coupled with the independence assumption between X and R, the following expression follows

$$f_Y(Y) = \int_0^{\infty} f_X(\frac{y}{r}) f_R(r) \frac{1}{r} dr. \tag{3.7}$$

One is interested in the value of the *pdf* of detected distances evaluated at 0. For $Y = 0$, this expression simplifies to

$$\begin{aligned} f_Y(0) &= \int_0^{\infty} f_X(0) \frac{1}{r} f_R(r) dr & (3.8) \\ &= f_X(0) \mathbb{E}(\frac{1}{R}) & (3.9) \\ &= f_X(0) K_l. & (3.10) \end{aligned}$$

Therefore, using equation 3.10, together with equations 3.4 and 3.5, a corrected estimator for density is

$$\hat{D}_c = \frac{n \hat{f}_X(0)}{2L} = \frac{n \; \hat{f}_Y(0)}{2 \; L \; K_l} = \frac{\hat{D}_y}{K_l}. \tag{3.11}$$

This requires that $\mathbb{E}(R^{-1})$ exists, reflecting restrictions to possible models for R and values that K_l might take for each assumed distribution of R. Given the distribution of R is known, K_l can be evaluated.

In practice, the distribution of R is not known and hence needs to be estimated. Two approaches are possible to estimate K_l: (1) use an assumed distribution to estimate parameters via maximum likelihood, and then calculate the mean value involved or (2) consider a nonparametric estimator.

A suitable nonparametric estimator for K_l^{-1} is the harmonic mean of a sample of R's ($M_H(r_s)$),

$$\hat{K}_l^{-1} = M_H(r_s) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{r_s} \tag{3.12}$$

where S is the number of observations for which both true and error distances are available. The final corrected density estimator can be expressed as

$$\hat{D}_c = \hat{D}_y M_H(r_s). \tag{3.13}$$

3.3.1.2 Analysis guidelines and variance estimation

The approach presented has the advantage that standard software available for analyzing distance sampling data can still be used. The usual distances, here assumed contaminated, are used to calculate an estimate, $\hat{D}_y$, which given the measurement error, is likely to be biased. Then, provided we have some observations for which we have both true and contaminated distances, we can estimate K_l, and include it in the analysis as a multiplier, with the corresponding measure of precision.

If all observations could be collected using a method that is precise enough so that we can consider them the true distances, i.e., error free, the methods presented here would not be needed. More frequently, the methods used during the larger part of the survey need to be fast and cheap, hence sometimes the error associated with them is considerable. If at least for a representative subset of the data we can use a more precise, likely more expensive or time consuming method, we can use the information contained on these pairs of contaminated and true distances to estimate K_l.

Note that these pairs of contaminated/true distances can be a subset of observations from the survey or observations collected on a separate experiment, although in the latter case care needs to be taken to ensure that the experiment conditions are similar to those on the survey.

If a subset of the survey is used, D_y and K_l are not necessarily independent, and hence that dependence should be incorporated in the analysis. The simplest way to do this is to use a nonparametric bootstrap where the observations are resampled both for the estimation of density and the error correction. Note that in such a setting it might be easier if the field protocol is such that one collects say the first f distances of each line both with the "error-free" method and the usual method. Since usually the bootstrap procedure for density considers the transects rather than observations as sampling units, this protocol ensures that resampling yields a representative sample to estimate both quantities we are interested, accounting for possible non-independence of D_y and K_l.

The analysis is simpler in the case of a separate experiment, because then D_y and K_l are independent; under this setting, the nonparametric bootstrap is still an option, but we can also estimate D_y and K_l variances independently and then combine them using the delta method as

$$var(\hat{D}_c) \simeq \hat{D}_c^2[cv^2(D_y) + cv^2(\hat{K}_l)]. \tag{3.14}$$

3.3.1.3 Using models for the error

Given that the estimator of K_l based on the harmonic mean is used, there is no need to assume a specific distribution for R. However, it is useful to do so to gain insight on the effect of different error structures on the resulting bias.

In Marques (2004) I considered models with beta and gamma distribution as plausible candidates for describing measurement error, implementing the use of beta related variables for R. Assume the following models for the distribution of R

$$\text{model I: } R = 0.5{+}U$$

$$\text{model II: } R = 2\ U$$

where $U \sim \text{beta}(\theta_1, \theta_2)$ and $\theta_1, \theta_2 \in (0, +\infty)$. Models with either $\theta_1 < 1$ or $\theta_2 < 1$ are not useful, since situations in which these error distributions would arise are unlikely.

Hence I consider only the case where θ_1 and $\theta_2 \geq 1$. Under these models, a symmetric beta ($\theta_1 = \theta_2$) results in unbiased estimation of distances. The contaminated distances take values between 0.5 and 1.5 (model I) or 0 and 2 times (model II) the original value.

The resulting errors are therefore dependent on the values of the original observations, although X and R are independent. The beta family contains a wide choice of shapes, and depending on parameter values, $\mathbb{E}(Y)$ can be larger, equal or lower than $\mathbb{E}(X)$, therefore allowing simulation of a wide range of different situations.

As shown generally for multiplicative error models in the previous section, even when the observed distances are unbiased estimators of the true distances, the density estimator is still biased. This is due to the fact that observations are unbiased if $\theta_1 = \theta_2$, and the estimator of population density is asymptotically unbiased if K_l=1, but neither condition implies the other.

Expression 3.9 can be developed for both models. For model I

$$f_Y(0) = f_X(0) \int_0^{+\infty} \frac{f_R(r)}{|r|} \,\mathrm{d}r = f_X(0) \int_0^{+\infty} \frac{f_U(r-0.5)}{r} \,\mathrm{d}r \tag{3.15}$$

$$= f_X(0) \int_{0.5}^{1.5} \frac{1}{B(\theta_1, \theta_2)} \frac{(r-0.5)^{\theta_1 - 1}(1.5-r)^{\theta_2 - 1}}{r} \,\mathrm{d}r \tag{3.16}$$

and for model II

$$f_Y(0) = f_X(0) \int_0^{+\infty} \frac{f_R(r)}{|r|} \,\mathrm{d}r = f_X(0) \int_0^{+\infty} \frac{\frac{1}{2} f_U(r/2)}{r} \,\mathrm{d}r \tag{3.17}$$

$$= f_X(0) \int_0^{2} \frac{1}{B(\theta_1, \theta_2)} \frac{1}{2} \frac{(r/2)^{\theta_1 - 1}(1-r/2)^{\theta_2 - 1}}{r} \,\mathrm{d}r. \tag{3.18}$$

For model II, substituting $t = r/2$ and simplifying leads to

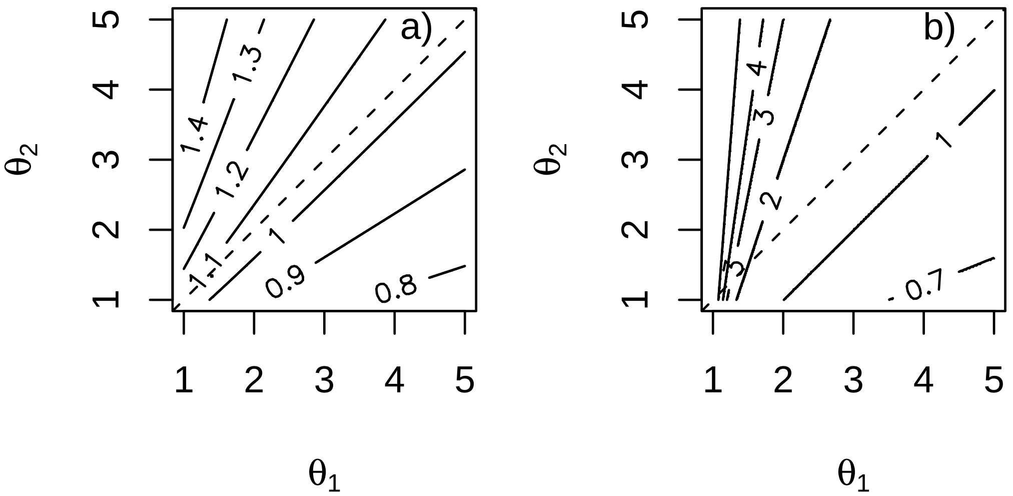

Figure 3.2: Values of the *PDF* correction K_l, as a function the parameters (θ_1, θ_2) of the error model: a) Under model I; b) Under model II. The dashed line indicates unbiased estimation of distances $(\theta_1 = \theta_2)$. Areas above and below the dashed line correspond respectively to underestimation and overestimation of distances. $K_l < 1$ corresponds to uncorrected density estimates being underestimated, and $K_l > 1$ corresponds to uncorrected density estimates being overestimated.

$$f_Y(0) = f_X(0)\frac{1}{2}\frac{(\theta_1 + \theta_2 - 1)}{(\theta_1 - 1)}. \tag{3.19}$$

The correction factor K_l can be calculated as a function of the error model parameters. The corresponding surfaces for models I and II are presented in figure 3.2. The effect of the error is more pronounced for model II for the same values of the parameters. In both cases, density estimates are positively biased if the error process is unbiased $(\theta_1 = \theta_2)$. In some cases, $\mathbb{E}(Y) > \mathbb{E}(X)$ results in positively biased density estimates. This reflects the impact of the specific form of error distribution on the estimation process. It is interesting to note that for some values of the parameters, $K_l = 1$, even though $\mathbb{E}(Y) > \mathbb{E}(X)$.

3.3.1.4 A simulation example

The performance of the proposed method was assessed by simulation. A known size simulated population was surveyed, the true distances to detected animals contaminated with errors of known structure, and the expected bias (and corresponding correction factor) were obtained. Then, using a subset of the data, the correction was estimated for each data set, and the corrected estimates compared to the ones ignoring the errors.

A population of 10000 animals was randomly (uniform coordinates in both dimensions) generated on a square with side 1000 meters ($D = 100$ animals/hectare). The study area was divided into 25 non-overlapping squares, and in each of these squares a transect of 200 meters was randomly selected. In each row of squares a transect was randomly generated for the first square and in the subsequent squares it was systematically placed with respect to the first one (see Figure 3.3). At the analysis stage, a truncation distance of 10 meters was used. To avoid edge effects, no transects were placed at less than 10 meters from the edge of the square. For every animal, a rejection method was used to decide if it was considered detected or not, based on a half-normal detection function ($\sigma = 5$). This process was repeated 100 times, resulting in 100 independent data sets. The average number of animals detected in each realization of the process was 593, standard deviation 20.1. The data generated were considered to be error free.

I then generated errors with the following distributions: beta(1,1); beta(3,2) and beta(5,5). The choice for these particular models was arbitrary, but had a rationale. The beta(1,1) was used as an extreme case, beta(5,5) as estimation of distances is unbiased, but density estimation is biased, and beta(3,2) as estimation of distances is biased, but nonetheless estimation of density should be unbiased. For each distance without error in these data sets, errors were independently generated, and introduced

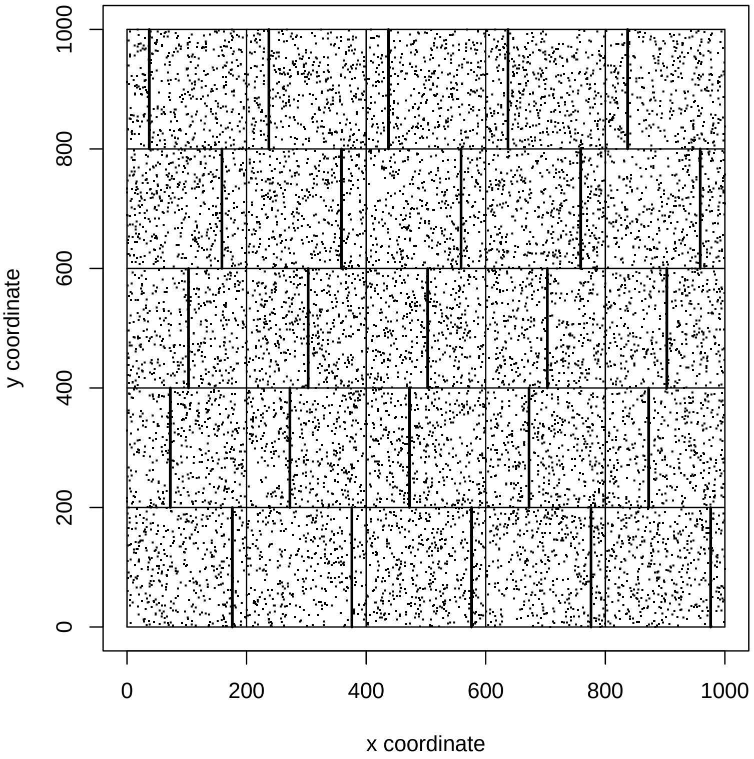

Figure 3.3: Schematics of the study area considered in the simulation study to access the performance of the PDF approach, with an example realization of the transects and animals. The survey design consisted of random placement of a transect in the first square of each row followed by systematic placement (with respect to the first square) of the remaining transects in each row.

as postulated above (for model I and II), leading to five contaminated sets. (The case of beta(1,1) for model II was not implemented, since K_l would be infinite.)

To preclude analyst influence, the data sets were analyzed using a conventional analysis in Distance 4 (Thomas *et al.*, 2002), as follows. The models considered for the detection function were half-normal+cosine ($HNcos$), uniform+cosine ($UNIcos$) and hazard rate+simple polynomial ($HRpol$), and the one with lowest AIC selected. The variance for encounter rate was calculated analytically based on replicate lines. In the analysis of contaminated data, the largest 5% of distances was truncated, as otherwise some models required several adjustment terms to provide an adequate fit of the data.

The analysis of the error free data led to an average estimated density of 98.6 animals per hectare, with a standard error of 0.65. The actual coverage for the 95% CI was 93%.

For the contaminated data sets, only 23 transects were used to estimate density, and the remaining 2 were used as a separate experiment, where true and contaminated distances were evaluated. This resulted on average in 516.2 (standard deviation 19.4) observations to estimate density and 49.6 (standard deviation 7.1) observations to estimate K_l. K_l was estimated using the harmonic mean estimator on the sample of R's resulting from the 2 transects, as well as based on the true beta model used to generate the errors, by maximizing the appropriate likelihood and then evaluating equations 3.16 or 3.19 by substitution of the true parameter values with their corresponding maximum likelihood estimates. The variance of K_l estimates, $var(K_l)$, was obtained by bootstrap (999 resamples). The variance for the corrected estimator, considering the nonparametric estimator for K_l, was then obtained by combining the variance of $\hat{D}_y$ and $\hat{K}_l$ using the delta method, using expression 3.14.

The true, mean estimated and mean observed K_l for each combination of model

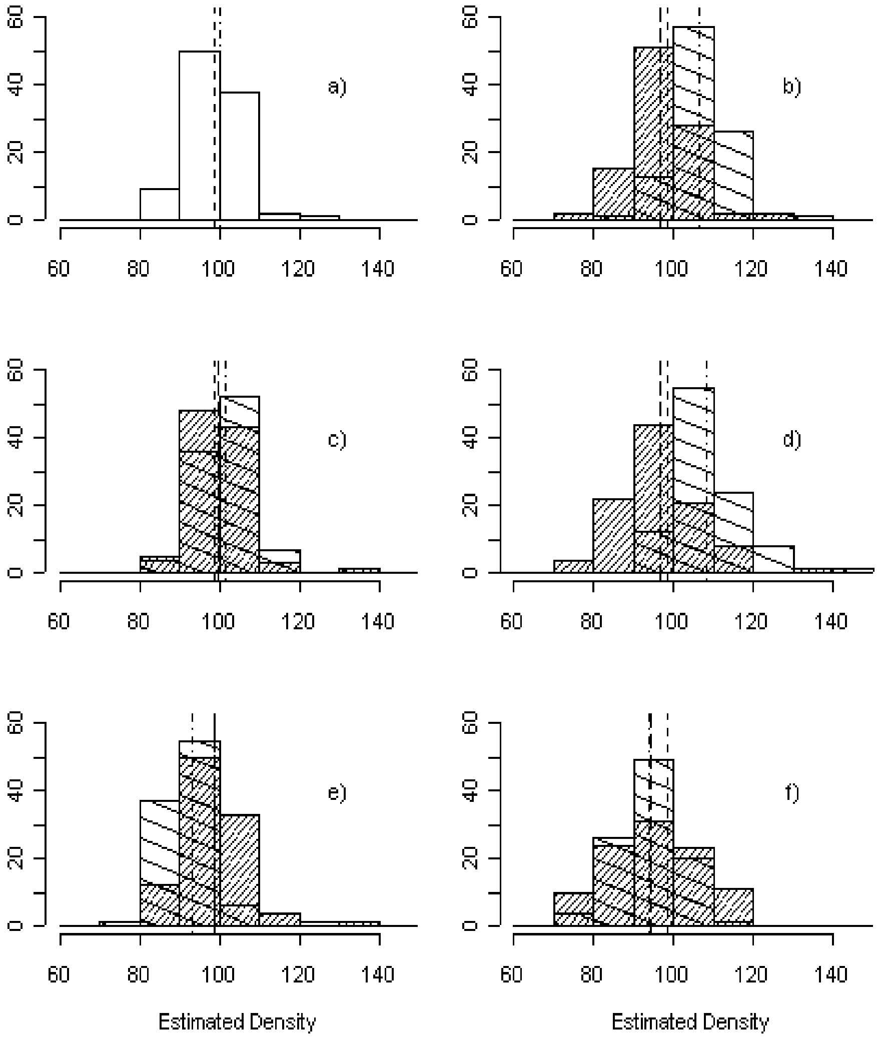

Figure 3.4: Results of the simulation exercise to evaluate the PDF approach to deal with measurement error. Error-based density estimates (lighter histograms) and the corrected estimates (darker histograms) using the harmonic mean estimator. a) True distances; b) Error beta(1,1), model I; c) Error beta(5,5), model I; d) Error beta(5,5), model II; e) Error beta(3,2), model I; f) Error beta(3,2), model II. Dashed line - mean value of estimates based on true distances. Dashed-doted line - mean value of estimates based on error distances. Long dashed line - mean value of estimates based on error distances, corrected with the PDF approach. True $D = 100$ animals/ha.

Table 3.1: Results of the simulation exercise to evaluate the PDF approach to deal with measurement error. True K_l (TK), mean estimated K_l using the harmonic mean (EHK) or the true beta model (EBK), and mean observed K_l (OK), under the combinations of errors and models (I - model I, II - model II) considered. Mean estimated density (D), density coefficient of variation (DCV) and coverage of 95% CI for density ($95\%CIC$), respectively for the corrected and uncorrected estimator. Mean estimated density based on true distances is 98.6 animals/ha. True density is 100 animals/ha.

	TK	EHK	EBK	OK	D^1	DCV^1	$95\%CIC^1$
beta(1,1), I	1.099	1.105	1.104	1.078	96.5/106.4	7.46/5.83	89/81
beta(5,5), I	1.024	1.021	1.021	1.030	99.4/101.5	6.76/6.76	95/94
beta(5,5), II	1.125	1.126	1.125	1.096	96.4/108.0	8.69/6.17	89/78
beta(3,2), I	0.944	0.943	0.943	0.941	98.5/92.8	6.83/6.13	94/75
beta(3,2), II	1.000	1.008	1.007	0.952	94.2/93.9	11.04/6.68	90/75

[1] Corrected analysis/uncorrected analysis

and error, along with the corresponding corrected and uncorrected mean density estimates, are presented in table 3.1. Also shown is the coverage of the 95% confidence interval, based both on corrected and uncorrected analysis.

The nonparametric estimator for K_l and the parametric beta-based estimator showed no differences, justifying the nonparametric estimator when the true model is unknown. There was an increased coverage with the use of the proposed correction in all cases. Figure 3.4 shows the uncorrected (i.e. error based density estimates) and the corrected density estimates using the harmonic mean estimator, showing that the correction reduced the bias in most cases. It can be seen that the results were very close to the expected ones, validating the predictions of the effects of errors and the proposed correction. However, in some cases, true K_l and observed K_l were slightly different. Especially in the case of beta(3,2) for model II, a K_l of 1 was expected but an average K_l of 0.952 was obtained. These unexpected results will be considered in the discussion. Note that even in this case coverage was increased, due to an increased variance related to the estimation of K_l.

3.3.1.5 A real life application

The *PDF* approach was applied to a survey of golf tees. Although clearly not a natural population, the setting used would not be any different than the one used for a real survey of, say, a plant population, with the advantage that truth was known[2].

Golf tee groups ($N = 250$) were randomly distributed in a study area of 1677.12 m^2, resulting in a density of 0.149 tees/m^2. Two strata, with different abundances (N_1=130 and N_2=120, in areas of respectively 1057.12 and 620 m^2), were surveyed for golf tees by 8 independent observers, which were considered as a single pooled observer, resulting in 125 sightings. The original tee data set includes group size, color and visibility, but these were ignored for the purpose of this study. The results here refer to tee group density, rather than individual tee density, but for simplicity this exposition refers only to tees and tee density. There were respectively 6 and 5 transects in each stratum, and the width of the transects was 4 m. Further details on this data set can be found in Borchers *et al.* (2002). Initial analysis of the data revealed a serious $g(0)$ problem. As $g(0)$ problems are a side issue for our purposes, I simply estimated it (and its variance) from the data and used it as a multiplier (Buckland *et al.*, 2001, p. 57) in Distance 4 (Thomas *et al.*, 2002). To mimic a real life application in which a separate experiment was conducted for measurement error assessment, I assumed that one transect in each stratum (with a total of 22 detections) was part of such an experiment, to estimate K_l. Thus, estimated distances (mean for those observers that saw each tee) and real distances were available for 22 observations. The remaining 9 transects (with 103 detections) were used in the usual way to derive a density estimate. Using Distance 4, estimates of density were obtained, both considering true distances and distances with errors (Table 3.2). The errors led to an overestimation of density of 16.8%. Using the harmonic mean of R,

[2] See Acknowledgements section for proper credit on the use of this data set.

Table 3.2: Analysis of the golf tees data set comparing the performance of the uncorrected estimator and the proposed PDF estimator. Results are shown considering the true distances, distances contaminated with errors (uncorrected estimator) and distances contaminated with errors with the proposed correction. True density is 0.149 tees/m^2.

Analysis	$\hat{D}$	DCV	D 95 % CI	$RMSE$
True distances	0.149	0.195	(0.100,0.222)	0.029
Error distances - D_y	0.174	0.195	(0.117,0.260)	0.042
Error distances - D_c	0.155	0.217	(0.100,0.238)	0.039

K_l was estimated as 1.123, and therefore a corrected point estimate of density was, using expression 3.13, 0.155. A bootstrap variance for K_l (999 bootstrap resamples) was calculated, and assuming K_l and D_y independent, I used $\hat{K_l}$ as a multiplier in the analysis. As can be seen from table 3.2, the corrected results now lie closer to the true values; 72% of the error bias in density was removed by using the correction, and the impact on precision was negligible.

Although the harmonic mean was used to estimate K_l, assuming either a beta or a gamma model leads to very similar estimates for K_l, respectively 1.115 and 1.117.

Although the CV of the corrected estimator is the larger, having to account for estimating an extra parameter, in terms of $RMSE$ we are better off using the correction than ignoring the effect of the errors.

3.3.2 Point transects

Although appropriate here for completeness, this subsection was developed by David Borchers, who generalized the line transect PDF approach by Marques (2004) to point transects. This section borrows on material in Burnham *et al.* (2004).

In the case of point transects the CDS estimator is

$$\hat{D} = \frac{n\ \hat{h}_X(0)}{2\pi k}. \tag{3.20}$$

If measurement errors are present, yet ignored, the conventional estimator is a biased estimator of density ($\hat{D}_y$). In practice, one does not observe the X's, but the Y's, hence density is effectively estimated as

$$\hat{D}_y = \frac{n\ \hat{h}_Y(0)}{2\pi k}. \tag{3.21}$$

Given equations 3.20 and 3.21, one would like to be able to express $h_Y(0)$ as a function of $h_X(0)$. Differentiating equation 3.7 leads to

$$\begin{aligned}
h_Y(y) &= \frac{df_Y(Y)}{dy} && (3.22)\\
&= \frac{d}{dy}(\int_0^\infty f_X(\frac{y}{r})\frac{1}{r}f_R(r)dr) && (3.23)\\
&= \int_0^\infty \frac{df_X(\frac{y}{r})}{dy}\frac{1}{r}f_R(r)dr) && (3.24)\\
&= \int_0^\infty \frac{df_X(\frac{y}{r})}{d\frac{y}{r}}\frac{1}{r^2}f_R(r)dr). && (3.25)
\end{aligned}$$

Therefore, we can express $h_Y(y)$, evaluated at 0, as

$$h_Y(0) = h_X(0)\int_0^\infty \frac{1}{r^2}f_R(r)dr = h_X(0)\mathbb{E}(\frac{1}{R^2}) = h_X(0)K_p. \tag{3.26}$$

Analogous to the case of lines, this expression requires that $\mathbb{E}(R^{-2})$ exists. Given the distribution of R, K_p can be evaluated. Therefore, a corrected estimator for density is

$$\hat{D}_c = \frac{n\ \hat{f}_Y(0)}{2\ L\ K_p} = \frac{\hat{D}_y}{K_p}. \tag{3.27}$$

As before, most commonly the distribution of R is not known and hence needs to be estimated. The same two approaches are possible to estimate K_p: (1) use the assumed distribution to estimate parameters via maximum likelihood, and then calculate the mean value involved or (2) to consider a nonparametric estimator.

A suitable nonparametric estimator for K_p^{-1} is the harmonic mean of a sample of R^2's ($M_H(r_s^2)$),

$$\hat{K_p}^{-1} = M_H(r_s^2) = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{r_s^2} \tag{3.28}$$

where as before S is the number of observations for which both true and error distances are available. The final corrected density estimator can be expressed as

$$\hat{D}_c = \hat{D}_y M_H(r_s^2). \tag{3.29}$$

3.3.3 Consequences of multiplicative errors on bias

By assuming a multiplicative error model we can derive some interesting conclusions about the effects of the corresponding error measurements.

The first is related to the influence of the error structure on the bias it promotes. According to Chen (1998) an unbiased error model leads to line transect underestimation of density, a fact then cited by Buckland *et al.* (2001, p. 264). The results in this section show that if a multiplicative model is assumed, unlike in the additive case described by Chen, the effect of unbiased errors, i.e. $\mathbb{E}(R) = 1$, is the opposite. Note that considering Jensen's inequality,

$$\mathbb{E}(\frac{1}{R}) \geq \frac{1}{\mathbb{E}(R)}, \tag{3.30}$$

where the equality is only achieved if R is a degenerate distribution with a point mass on 1, which means by definition no measurement error. Therefore, equation 3.9

shows that, considering a multiplicative model, unbiased errors lead to overestimation of density, which means that the effect of measurement error is model dependent.

A second conclusion is that, conditional on an unbiased multiplicative measurement error, the same error model will cause greater bias for point transect data than for line transect data. This is because under such conditions it can be shown that $\mathbb{E}(\frac{1}{R^2}) > \mathbb{E}(\frac{1}{R})$. In fact, given two random variables W and Z, the Cauchy-Schwartz inequality states that

$$[\mathbb{E}(WZ)]^2 \leq \mathbb{E}(W^2)\mathbb{E}(Z^2). \tag{3.31}$$

Hence, considering $W = 1/R$ and $Z = 1$, leads to

$$\left[\mathbb{E}\left(\frac{1}{R} \times 1\right)\right]^2 \leq \mathbb{E}(\frac{1}{R^2}) \times \mathbb{E}(1). \tag{3.32}$$

On the other hand, as $\mathbb{E}(\frac{1}{R}) \geq 1$ (Jensen's inequality with $\mathbb{E}(R) = 1$), then

$$\left[\mathbb{E}\left(\frac{1}{R}\right)\right]^2 > \mathbb{E}(\frac{1}{R}) \tag{3.33}$$

and finally, combining the expressions above leads to

$$\mathbb{E}(\frac{1}{R^2}) \geq \left[\mathbb{E}(\frac{1}{R})\right]^2 \geq \mathbb{E}(\frac{1}{R}) \tag{3.34}$$

proving the statement above.

Strictly speaking, the idea that the same amount of error is believed to lead to larger bias for point transects than for line transects (e.g. Buckland *et al.*, 2001, p. 264-265) is not valid for all error models. Whilst this might be true for most cases, one can easily come up with counter examples: consider the bias originated by a gamma model, with parameters such that the mean value of R is between 0.9 and 1.2 (hence errors go from underestimation to overestimation of distances, with unbiased errors

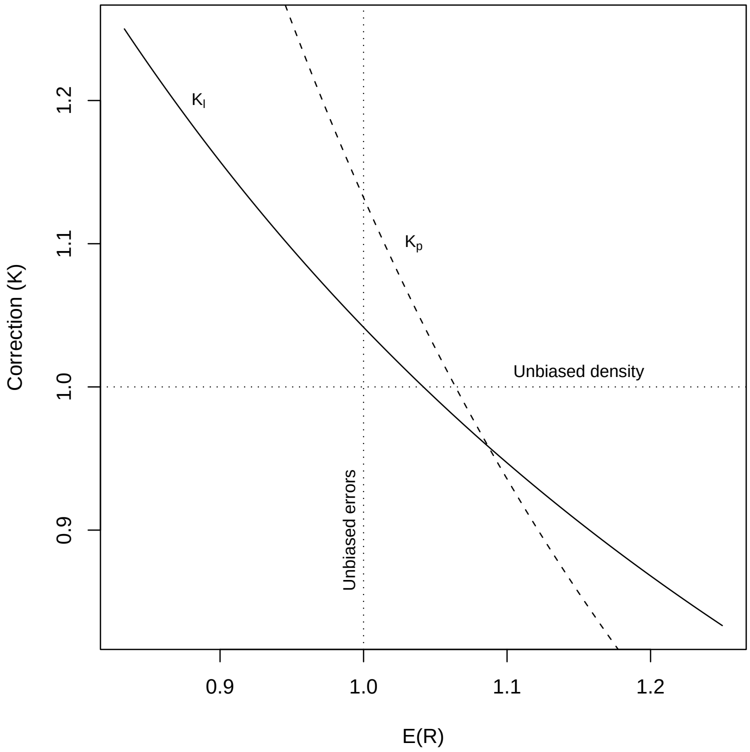

Figure 3.5: Illustrative example showing that the correction needed for lines might be more severe than the correction needed for points. Expected correction needed, for lines (solid line) and points (dashed line), as a function of the mean value of a gamma model for the error. Gamma scale parameter varying from $\frac{1}{20}$ to $\frac{1}{30}$, and shape parameter $= 25$. Horizontal dotted line represents unbiased density and vertical dotted line represents unbiased errors.

for $\mathbb{E}(R) = 1$). Whilst for most models considered, the correction is more extreme for points, for some values of the mean value of R above 1, the correction is further away from 1 for lines than for points, which means the bias is larger for lines (Figure 3.5). However, note that for such cases the correction is so close to 1 that for practical purposes it is like if no correction was needed, and hence this is an academic remark with likely no impact on real life studies, where measurement error is indeed likely to be more influential for points rather than lines.

It seems intuitive that provided the amplitude of the errors is small, their effect will be negligible. An important question to be solved is whether, in a given setting, we should worry about the effect of measurement errors or if, regarding the application at hand, the accuracy and precision are good enough to preclude significant bias.

Assume that the error is unbiased, meaning that although the error process introduces an extra variance component in the estimated distances, there is no consistent over- or underestimation of distances, which assuming a multiplicative error, implies that $\mathbb{E}(R)$=1. Considering a specific model for the error, we can make some comments on the expected bias as a function of the errors magnitude, namely as a function of its coefficient of variation (CV_R). I shall assume that $R \sim \text{gamma}(\sigma, b)$, with the following parametrization

$$f_R(r) = \frac{r^{b-1}e^{-\frac{r}{\sigma}}}{\sigma^b\Gamma(b)}, \quad \sigma, b > 0, \quad r > 0. \tag{3.35}$$

Simple algebra leads to the expression for the appropriate corrections for lines as

$$\mathbb{E}(R^{-1}) = \frac{1}{\sigma(b-1)} \tag{3.36}$$

and for points as

$$\mathbb{E}(R^{-2}) = \frac{1}{\sigma^2(b-1)(b-2)}. \tag{3.37}$$

Note that for the above gamma model parametrization, $\mathbb{E}(R) = b\sigma$, and $var(R) = b\sigma^2$, hence $CV_R = \frac{1}{\sqrt{b}}$, and as $\mathbb{E}(R) = 1$, $\sigma = \frac{1}{b}$. Using these, we can show that the correction is a function of the CV_R, as in the case of line transects

$$K_l = \frac{1}{1 - CV_R^2} \tag{3.38}$$

while for point transects

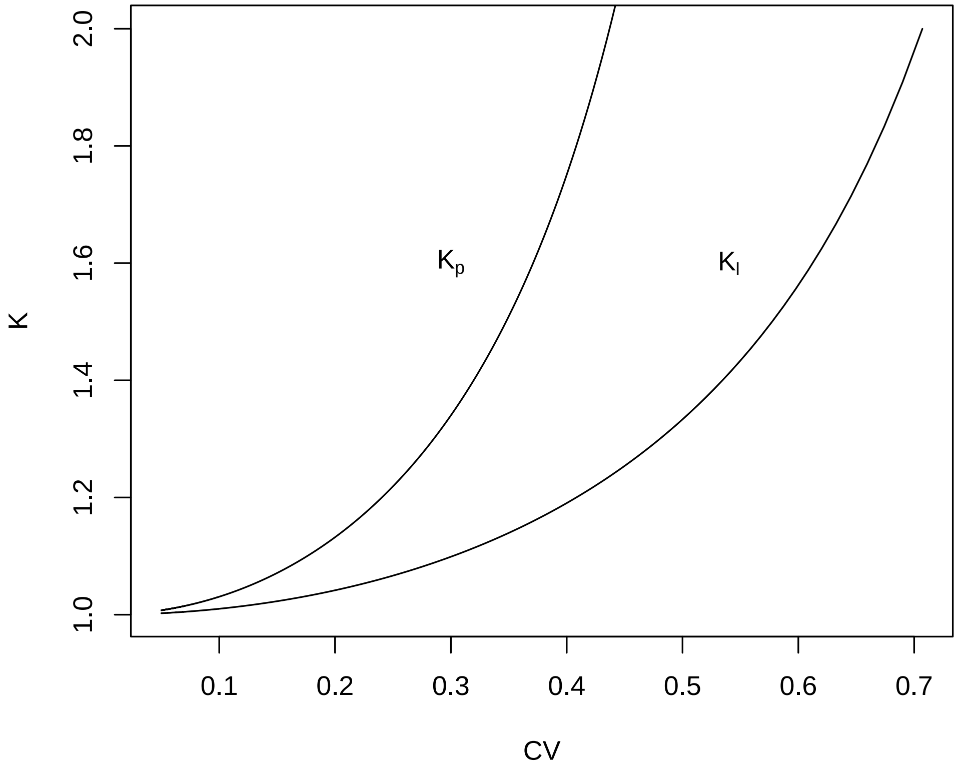

Figure 3.6: Illustration of the effect of the precision in distance measurement on the expected bias in density estimates: expected values of K, for lines (K_l) and points (K_p) as a function of the CV of the error, assuming an unbiased gamma model for R.

$$K_p = \frac{1}{(1 - CV_R^2)(1 - 2CV_R^2)}. \tag{3.39}$$

Hence we can plot the correction needed for points and lines as a function of the CV of a gamma model. As the CV of the errors increases, the overestimation of density increases exponentially (Figure 3.6). Despite this increase, in absolute terms, for a gamma model with CV of 20% the expected bias is around 4.2% for lines, which is reassuringly low bias for a fair amount of error, but 13.2% for points, which is already non-negligible.

3.4 Incorporating measurement error through a likelihood approach

The PDF approach to deal with errors is simple to use, and intuitive in the sense that a biased estimator is obtained and then a separate correction factor implemented, which allows standard distance sampling software Distance to be used. However, it is not flexible and seems to have problems under some settings (see section 3.3.1.4).

A more elegant and easily generalizable approach follows: a likelihood is derived that allows the use of data with measurement error to estimate the parameters of the detection function without bias. As before, a model for the error is needed. Although I have been a co-author on both Burnham *et al.* (2004) and Borchers *et al.* (in prep a), from which I borrow some material, the original idea of formalizing and developing these methods is due to David Borchers. A similar approach was likely used by Hiby *et al.* (1989), although neither the actual methods or the way to implement them were described in that paper.

3.4.1 A likelihood approach to accommodate measurement error

As was seen before, if the observed data are subject to measurement error but the error is ignored, the parameter (vector) of the detection function, $\underline{\phi_1}$, will be estimated with bias, and hence abundance estimation will be biased.

In this section, I derive a likelihood function that integrates a measurement error model, hence allowing the estimation of $\underline{\phi_1}$ without bias, based on the data with measurement error.

3.4.1.1 The proposed likelihood

As in the case of the PDF approach, we will need some information on the measurement error. I assume that a sample of n contaminated distances, $\underline{y} = (y_1, y_2, ..., y_n)$ is available, to estimate abundance. Additionally there is also available a sample of pairs of contaminated and true distances, numbered for convenience from $n+1$ to $n+J$, $(\underline{x}^+, \underline{y}^+) = \{(x, y)_{n+1}, (x, y)_{n+2}, ..., (x, y)_{n+J}\}$. These could be obtained in a separate experiment.

Assuming a parametric form for $Y|X$, indexed by a parameter vector $\underline{\gamma}$, we can use the J pairs of distances to maximize the following likelihood with respect to $\underline{\gamma}$

$$\mathcal{L}(\underline{\gamma}|\underline{y}^+, \underline{x}^+) = \prod_{i=1}^{J} \pi(y_{n+i}|x_{n+i}) \tag{3.40}$$

to obtain an estimate for the error model parameter. Given that, it is possible to use the n contaminated distances to estimate the detection function parameter vector $\underline{\phi_1}$ accounting for measurement error. Building on the model for $Y|X$, we can obtain the joint distribution of (Y, X), as we assume as usual a model for the distribution of X and $f(x, y) = f(x)f(y|x)$. Therefore we can also get the distribution of Y as

$$f_Y(y) = \int f(x, y)dx = \int f(x)f(y|x)dx. \tag{3.41}$$

Using equation 2.34, this expression can be further developed, leading to

$$f_Y(y) = \int \frac{g(x)\pi(x)}{\int_0^w g(x)\pi(x)\ dx} f(y|x)dx \tag{3.42}$$

which provides the basis for a likelihood from which estimation of $\underline{\phi_1}$ in the presence of measurement error is possible, conditional on the previously estimated measurement error parameter $\underline{\gamma}$, as

$$\mathcal{L}(\underline{\phi_1}|\underline{y},\underline{\gamma}) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \int \frac{g(x)\pi(x)}{\int_0^w g(x)\pi(x)\ dx} f(y_i|x)dx. \tag{3.43}$$

The implementation of this likelihood is done trough numeric methods. Care must be taken in the choice of a distance over which the above integral is evaluated. In practice, a distance is chosen such that it would be unlikely that any true distance would be larger than said distance.

The estimation could be made simultaneously for $(\underline{\phi_1},\underline{\gamma})$ by maximizing the joint likelihood

$$\mathcal{L}(\underline{\phi_1},\underline{\gamma}|\underline{y},\underline{y}^+,\underline{x}^+) = \mathcal{L}(\underline{\phi_1}|\underline{y},\underline{\gamma})\mathcal{L}(\underline{\gamma}|\underline{y}^+,\underline{x}^+). \tag{3.44}$$

It can be shown that the conditional estimators are asymptotically normal, but the maximum likelihood estimator from the joint likelihood is more efficient than that from the conditional likelihood (P. Jupp, unpublished material). However, in practice, there seems to be small efficiency gain by using the the joint likelihood since there is only information about the measurement error model in the pairs of distances (see also section 3.5.4 for preliminary results on this subject).

Once the estimates for $\underline{\phi_1}$ are obtained the estimation proceeds in the same way as for conventional methods. The probability of detecting an animal is estimated and an HTL estimator of density/abundance follows.

3.4.1.2 Variance estimation

I propose that resampling methods are used to obtain variance estimates and confidence intervals for the estimators involved. The simplest approach is to use a non-parametric bootstrap, as described for the PDF approach.

This resampling strategy means that the variance estimates are not conditional on the true distances used in the experiment, which could lead to a slight overestimation of variance.

As an alternative, a parametric bootstrap strategy could be implemented. For error models in the exponential family, not only point estimates $\hat{\underline{\gamma}}$ but also a corresponding variance-covariance matrix $\hat{\Sigma}_{\underline{\gamma}}$ can be readily obtained from standard GLM software. Hence, given that maximum likelihood estimators are asymptotically normally distributed, for each resample, the likelihood equation 3.43 can be maximized conditional on a random deviate from a normal distribution with mean $\hat{\underline{\gamma}}$ and variance-covariance $\hat{\Sigma}_{\underline{\gamma}}$. Variances and confidence intervals are then obtained in the same way as for the nonparametric bootstrap.

3.4.2 Assessing the methods by simulation

To assess the performance of the proposed methods a simulation exercise was conducted. Both line and point transects were considered, and estimates derived by the PDF approach were also included to compare the two methods.

3.4.2.1 Simulation settings

The animals were distributed randomly in a square with side long enough so that the probability of detection at the edge of the square was low ($< 10^{-4}$). This procedure was intended to avoid edge effects, since as stated before these are not a relevant issue in the context of this thesis.

I considered a half-normal detection function, both with a narrow ($\sigma = 0.125$) and a wider shoulder ($\sigma = 0.25$), as well as a hazard-rate (with $\sigma = 0.25$, $b = 4.3$), with a wider shoulder followed by a steeper slope (Figure 3.7).

Given these detection functions, total population sizes were set to obtain sample sizes (i.e. mean number of detected animals) that would be considered both “small” and “large”. Following the general guidelines in Buckland *et al.* (2001), small sample

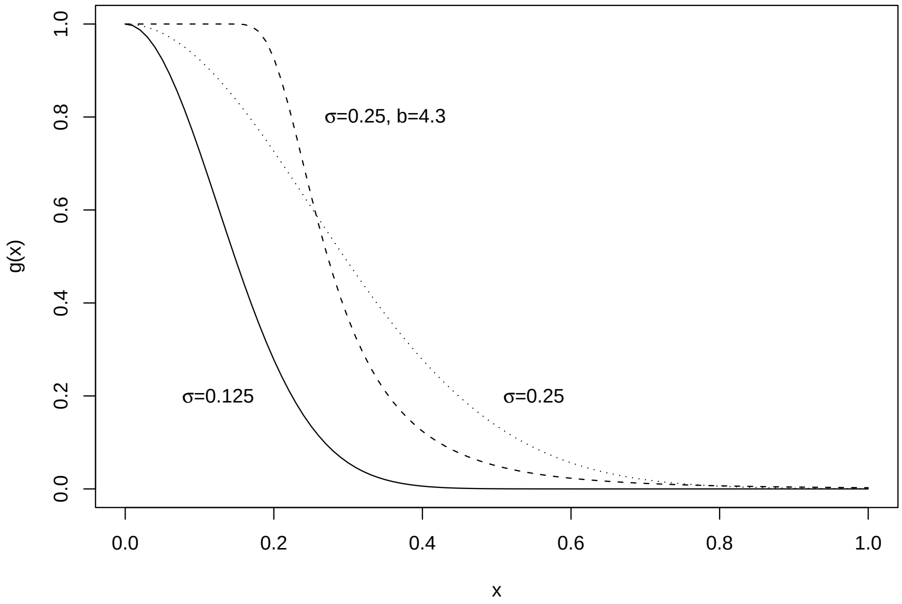

Figure 3.7: Detection functions used in the simulation exercise to evaluate the performance of the likelihood approach to deal with measurement error. The solid and dotted lines are half-normals with different scale parameters ($\sigma = 0.125$ and 0.25); the dashed line is hazard-rate ($b = 4.3$, $\sigma = 0.25$)

Table 3.3: Error models used in the simulation exercise to assess methods to deal with measurement error. $\mathbb{E}[Y|X]$ and $CV[Y|X]$ represent respectively the mean value and the coefficient of variation of $Y|X$. K_l is the true correction factor value for line transects; K_p is the true correction factor value for point transects. The values of $K_l - 1$ and $K_p - 1$ are the expected percentage biases for line and point transect estimators which ignore the measurement error.

Model	$\mathbb{E}[Y\|X]$	$CV[Y\|X]$	$K_l - 1$	$K_p - 1$
CV_{10}	X	10%	1%	3%
CV_{30}	X	30%	10%	34%
CV_{50}	X	50%	33%	167%
$0.8CV_{30}$	$0.8X$	30%	37%	109%
$1.2CV_{30}$	$1.2X$	30%	-8%	-7%

size was considered to be $n = 60$ and $n = 80$ respectively for line and point transects, and large sample size $n = 300$ for both lines and points.

The distribution of $Y|X$ was assumed to be gamma, and 5 different measurement error scenarios considered, as shown in table 3.3: in addition to 3 levels of unbiased errors (low, medium and high CV, respectively with CV= 10, 30 and 50%), I considered a situation with underestimation ($\mathbb{E}(Y|X) = 0.8X$) and with overestimation ($\mathbb{E}(Y|X) = 1.2X$) of true distances (both of these with CV's of 30%). Note that this last measurement error model, according to the bias predicted by the PDF approach, should lead to worst bias for lines than for points.

For each combination of sampler, detection function, sample size and measurement error model, 500 simulated data sets were generated, each consisting of the true (X) and contaminated (Y) detected distances. Each data set was analyzed considering the following approaches:

1. original error free data analyzed with conventional distance sampling, considering either the true detection function ($X_{TM}CDS$) or the best model chosen by minimum AIC ($X_{BM}CDS$), from a set of plausible models, namely and as for

the previous PDF approach simulation exercise, $HNcos$, $UNcos$ and $HRpol$. These were intended to act as controls. The true control corresponds to using the true model, but given in practice the true model is not known, I also used the best AIC model approach;

2. the same analysis as before but considering the contaminated data, respectively $Y_{TM}CDS$ and $Y_{BM}CDS$. These correspond to the situation where conventional methods are used ignoring the effects of measurement error;

3. contaminated data analyzed with the proposed likelihood methods, considering the detection function and error model as known ($Y_{TM}LIK$), using equation 3.43.

Based on the uncorrected estimates for the second approach, the corresponding results corrected using the PDF method were also calculated (referred respectively to as $Y_{TM}PDF$ and $Y_{BM}PDF$).

As a way to assess the performance of these different approaches the mean estimated % bias is presented,

$$100\frac{\sum_{k=1}^{500}\frac{\hat{N}_k-N}{N}}{500} \tag{3.45}$$

where $\hat{N}_k$ represents the estimated population size in the k^{th} simulation and N the true population size.

3.4.2.2 Simulation results

The mean observed bias for the scenarios considered in the simulation exercise are shown in tables 3.4-3.7. The analysis of these tables allows a number of different considerations:

- For unbiased measurement error with small $CV(10\%)$, the effect of ignoring measurement error seems negligible, and trying to correct for it brings no additional gain;

- As expected, for large unbiased measurement error or biased errors the effect of ignoring measurement error might be severe, with large bias shown for most scenarios tested;

- When analyzing data with measurement error, there is a noticeable difference in considering the true model or the best model when the true detection function is HN. In fact, although I predict that unbiased multiplicative errors should lead to overestimation of abundance, the opposite is observed when the true model is used, in contrast with the expected bias direction observed when the best model is used. This effect is much less noticeable for the HR. This is the result of the HN lack of flexibility, less evident in the more flexible 2 parameter HR, and reflects a shortcoming of the PDF approach further explored later: the method relies on the models used for the detection function being able to reflect the expected bias;

- While for large sample size scenarios the control bias is overall low, for small sample sizes there is evidence for some bias, especially in the case of the HR model. This is likely related to the choice of truncation distance. For CDS the usual strategy of no truncation effectively means that the largest distance is used for truncation, as we need a finite value to implement the numeric maximization of the likelihood. This procedure can be responsible for a small bias, as it has as consequence that a data point is necessarily present in the tail of the distribution (T. Marques, unpublished data);

- The bias for the PDF approach is usually lower when the best model strategy is used, rather than the true model. Nonetheless and overall the use of the PDF approach presents an improvement over simply ignoring measurement error for line transects, but the remaining bias shown makes it unlikely to be an adequate approach for point transects;

- There seems to be some consistency on the bias across different sample sizes, both for line and point transects;
- Considering the PDF approach, we would expect the same bias irrespectively of the true detection function. Results are consistent across the two different parameter values for the HN, but nonetheless there are considerable differences when the HN and HR results are compared;
- The likelihood approach seems to be unbiased for almost all scenarios shown, hence justifying this approach to integrate measurement error in distance sampling analysis.

Overall, the observed bias seems to be a function of the true detection function, sample size and measurement error used, coupled with the flexibility of the model used for the detection function.

3.4.3 A real life application

To illustrate the methods I considered the data from an aerial minke whale survey (referred to as NASS87), in which cue counting methods were used[3]. Although for estimating abundance cue counting methods require additional data, namely cue production rate, in terms of detection function/detection probability estimation, there is no difference between cue counting or point transects, and hence (all other things being equal) similar bias should be expected due to measurement error in both cases (see e.g. Buckland *et al.* 2001, p. 191-197, for further details about cue counting).

[3] See Acknowledgements section for proper credit on the use of this data set.

Table 3.4: Results of the simulation exercise to assess methods to deal with measurement error. Mean observed % bias from 500 simulations, for line transects, considering small sample size ($n = 60$), for the different analysis strategies used. Y stands for an analysis based on contaminated distances, and X for an analysis based on true distances. CDS stands for conventional methods, LIK for the likelihood approach, and PDF for conventional CDS estimates corrected using the PDF approach. TM indicates that the true detection function model was used, and BM that the best model chosen by minimum AIC was selected. See text for further details.

Detection	Error				$\mathbb{E}(n)$=60			
Function	Model	$X_{TM}CDS$	$X_{BM}CDS$	$Y_{TM}CDS$	$Y_{BM}CDS$	$Y_{TM}LIK$	$Y_{TM}PDF$	$Y_{BM}PDF$
	CV_{10}	-3.24	0.94	-3.32	1.64	1.63	-4.27	0.64
HN	CV_{30}	-4.80	0.52	-6.53	8.16	-0.17	-14.94	-1.58
$\sigma = 0.125$	CV_{50}	-3.44	0.3	-9.68	23.09	1.44	-32.26	- 7.68
	$0.8CV_{30}$	-4.56	0.71	17.67	35.77	0.28	-14.34	-1.16
	$1.2CV_{30}$	-4.32	1.82	-20.76	-9.46	1.26	-13.46	-1.12
	CV_{10}	-4.27	1.66	-4.23	1.17	0.72	-5.18	0.17
HN	CV_{30}	-3.53	0.4	-4.81	7.24	1.42	-13.38	-2.41
$\sigma = 0.25$	CV_{50}	-2.98	0.66	-7.94	24.38	2.74	-30.95	-6.71
	$0.8CV_{30}$	-3.56	-0.1	19.22	33.58	1.58	-13.21	-2.75
	$1.2CV_{30}$	-3.24	1.12	-20.61	-9.69	1.47	-13.30	-1.38
	CV_{10}	4.46	7.42	5.27	7.99	1.41	4.23	6.92
HR	CV_{30}	3.05	7.36	10.76	14.31	-0.16	0.79	4.03
$\sigma = 0.25$	CV_{50}	4.59	8.94	31.05	28.99	1.29	-1.71	-3.26
$b = 4.3$	$0.8CV_{30}$	4.50	8.04	42.28	43.57	2.27	3.58	4.52
	$1.2CV_{30}$	6.31	8.02	-4.59	-4.58	2.61	4.19	4.2

Table 3.5: Results of the simulation exercise to assess methods to deal with measurement error. Mean observed % bias from 500 simulations, for point transects, considering small sample size ($n = 80$), for the different analysis strategies used. Y stands for an analysis based on contaminated distances, and X for an analysis based on true distances. CDS stands for conventional methods, LIK for the likelihood approach, and PDF for conventional CDS estimates corrected using the PDF approach. TM indicates that the true detection function model was used, and BM that the best model chosen by minimum AIC was selected. See text for further details.

Detection	Error				$\mathbb{E}(n)$=80			
Function	Model	$X_{TM}CDS$	$X_{BM}CDS$	$Y_{TM}CDS$	$Y_{BM}CDS$	$Y_{TM}LIK$	$Y_{TM}PDF$	$Y_{BM}PDF$
	CV_{10}	-1.72	-0.5	-2.01	0.09	3.16	-4.86	-2.83
HN	CV_{30}	-3.10	-2.1	-7.40	19.46	2.12	-30.90	-10.85
$\sigma = 0.125$	CV_{50}	-3.55	-3.24	-17.89	72.11	0.75	-69.13	-35.46
	$0.8CV_{30}$	-2.25	-2.37	45.93	84.65	2.42	-30.18	-2.13
	$1.2CV_{30}$	-2.51	-1	-35.98	-16.97	1.72	-31.16	-10.78
	CV_{10}	-3.36	-1.68	-3.68	0.24	1.59	-6.48	-2.67
HN	CV_{30}	-4.06	-2.38	-8.36	12.85	1.08	-31.61	-15.79
$\sigma = 0.25$	CV_{50}	-3.43	-1.06	-19.07	73.22	0.48	-69.58	-35.04
	$0.8CV_{30}$	-2.47	-2.29	44.55	78.18	2.08	-30.84	-5.57
	$1.2CV_{30}$	-3.33	-1.27	-36.26	-17.55	1.33	-31.46	-11.40
	CV_{10}	3.19	5.58	5.25	7.59	0.27	2.19	4.46
HR	CV_{30}	2.86	6.7	23.07	26.06	-0.92	-8.15	-5.93
$\sigma = 0.25$	CV_{50}	4.13	5.78	81.26	86.38	0.13	-31.86	-30.11
$b = 4.3$	$0.8CV_{30}$	5.20	6.03	100.44	98.8	0.92	-4.09	5.37
	$1.2CV_{30}$	4.71	6.85	-12.67	-10.87	1.17	-6.10	-4.22

Table 3.6: Results of the simulation exercise to assess methods to deal with measurement error. Mean observed % bias from 500 simulations, for line transects, considering large sample size ($n = 300$), for the different analysis strategies used. Y stands for an analysis based on contaminated distances, and X for an analysis based on true distances. CDS stands for conventional methods, LIK for the likelihood approach, and PDF for conventional CDS estimates corrected using the PDF approach. TM indicates that the true detection function model was used, and BM that the best model chosen by minimum AIC was selected. See text for further details.

Detection	Error	$\mathbb{E}(n)$=300						
Function	Model	$X_{TM}CDS$	$X_{BM}CDS$	$Y_{TM}CDS$	$Y_{BM}CDS$	$Y_{TM}LIK$	$Y_{TM}PDF$	$Y_{BM}PDF$
	CV_{10}	-1.22	-1.96	-1.52	-1.24	0.03	-2.50	-2.22
HN	CV_{30}	-0.6	-1.19	-3.95	4.85	0.56	-12.60	-4.58
$\sigma = 0.125$	CV_{50}	-1.61	-1.68	-10.33	16.28	-0.35	-32.74	-12.79
	$0.8CV_{30}$	-0.89	-0.79	19.93	31.85	0.36	-12.69	-4.01
	$1.2CV_{30}$	-0.62	-1.19	-19.81	-12.49	0.66	-12.43	-4.43
	CV_{10}	-0.78	-1.84	-1.06	-1.03	0.48	-2.04	-2.01
HN	CV_{30}	-1.15	-0.86	-4.35	5.71	0.07	-12.95	-3.8
$\sigma = 0.25$	CV_{50}	-1.04	-0.46	-9.9	17.98	0.11	-32.42	-11.52
	$0.8CV_{30}$	-1.18	-0.82	19.7	31.95	0.19	-12.86	-3.94
	$1.2CV_{30}$	-0.99	-1.41	-20.12	-12.68	0.2	-12.77	-4.64
	CV_{10}	0.96	5.79	1.71	6.31	0.43	0.71	5.26
HR	CV_{30}	0.22	2.03	6.03	9.76	-0.39	-3.51	-0.12
$\sigma = 0.25$	CV_{50}	0.37	2.04	18.71	21.74	-0.68	-10.97	-8.7
$b = 4.3$	$0.8CV_{30}$	0.30	2.02	32.30	37.45	-0.59	-3.68	0.07
	$1.2CV_{30}$	0.68	2.2	-11.37	-8.5	-0.14	-3.22	-0.08

Table 3.7: Results of the simulation exercise to assess methods to deal with measurement error. Mean observed % bias from 500 simulations, for point transects, considering large sample size ($n = 300$), for the different analysis strategies used. Y stands for an analysis based on contaminated distances, and X for an analysis based on true distances. CDS stands for conventional methods, LIK for the likelihood approach, and PDF for conventional CDS estimates corrected using the PDF approach. TM indicates that the true detection function model was used, and BM that the best model chosen by minimum AIC was selected. See text for further details.

Detection	Error			$\mathbb{E}(n)$=300				
Function	Model	$X_{TM}CDS$	$X_{BM}CDS$	$Y_{TM}CDS$	$Y_{BM}CDS$	$Y_{TM}LIK$	$Y_{TM}PDF$	$Y_{BM}PDF$
	CV_{10}	-0.8	-1.83	-1.32	-0.34	0.83	-4.19	-3.25
HN	CV_{30}	-0.64	-3.58	-7.23	16.66	1.09	-30.77	-12.95
$\sigma = 0.125$	CV_{50}	-1.61	-2.2	-19.45	72.87	0.09	-69.72	-35.18
	$0.8CV_{30}$	-1.3	-2.8	44.01	81.62	0.57	-31.10	-3.74
	$1.2CV_{30}$	-1.49	-2.63	-36.29	-18.97	0.09	-31.50	-12.93
	CV_{10}	-1.04	-1.8	-1.62	-0.13	0.52	-4.49	-3.04
HN	CV_{30}	-1.1	-1.68	-7.48	19.54	0.95	-30.95	-10.80
$\sigma = 0.25$	CV_{50}	-1.72	-1.54	-18.92	75.53	0.47	-69.52	-34.18
	$0.8CV_{30}$	-1.95	-2.34	42.63	84.28	-0.32	-31.76	-2.33
	$1.2CV_{30}$	-0.53	-2.11	-35.58	-18.93	0.95	-30.73	-12.88
	CV_{10}	0.39	1.75	2.44	3.2	-0.96	-0.55	0.20
HR	CV_{30}	1.46	0.77	19.43	18.24	-0.41	-10.88	-11.77
$\sigma = 0.25$	CV_{50}	0.49	0.95	63.47	68.2	-1.57	-38.54	-36.93
$b = 4.3$	$0.8CV_{30}$	0.64	0.84	85.05	86.07	-1.13	-11.46	-1.39
	$1.2CV_{30}$	1.17	1.67	-17.24	-16.5	-0.28	-11.01	-10.28

The number of detected cues during NASS87 was n=148. Borchers *et al.* (in review) present a complete analysis of this data set, and the reader is referred to them for further details on the survey. I focus here on the estimation of the probability of detection.

This data set was also analyzed by Hiby *et al.* (1989), who estimated a measurement error CV of 35%. The following analysis, based on the proposed methods, is conditional on this measurement error CV, which is assumed unbiased and having a gamma distribution.

The first approach was to analyze the data ignoring measurement error, using CDS. I used Distance 5 (Thomas *et al.*, 2005), selecting as candidate models for the detection function the half-normal and hazard-rate key functions, combined respectively with cosine and simple polynomial adjustment terms ($Y_{BM}CDS$). The results are presented in table 3.8. For the HR model, adjustment terms led to considerably higher AIC, hence the results are not shown.

The same detection function models were used, now considering the proposed methods with the measurement error model being gamma and assumed unbiased, i.e. $Y|X \sim Ga(0.35^2, 0.35^{-2})$. The corresponding results ($Y_{BM}LIK$) are also shown in table 3.8.

Resampling procedures were used for variance estimation: a nonparametric bootstrap routine was implemented in R (R Development Core Team, 2006), with the transects as resampling units (999 resamples), and 95% confidence intervals obtained by the percentile method (see section 2.3.2).

It is important to note that the AIC can not be used to compare across the different approaches, because they are conditional on different error models, one corresponding to no error, and the other to an error with 35% CV.

Ignoring measurement error, the half-normal with 1 cosine adjustment seems the

Table 3.8: Estimated detection probabilities ($\hat{P}$) from the NASS87 cue counting survey data, and corresponding coefficient of variation ($\hat{CV}_{\hat{P}}$) and 95% confidence intervals (95% CI), ignoring measurement error ($Y_{BM}CDS$) and accounting for it using the likelihood approach ($Y_{BM}LIK$). ΔAIC values also shown, allowing comparison within each estimation method. Detection function models, $g(r)$, are: HN - half-normal key with no adjustments; $HNcos$ - half-normal key with one cosine adjustment term; HR - hazard rate key with no adjustments.

Method	$g(r)$	$\hat{P}$	$\hat{CV}_{\hat{P}}$	95% CI	ΔAIC
$Y_{BM}CDS$	HN	0.128	17.3	(0.101;0.196)	26.2
	$HNcos$	0.082	17.7	(0.07;0.133)	0
	HR	0.102	24.0	(0.054;0.166)	2.6
$Y_{BM}LIK$	HN	0.110	17.8	(0.088;0.171)	0
	$HNcos$	0.096	19.9	(0.077;0.158)	1.3
	HR	0.137	22.5	(0.105;0.244)	6.4

best model for the data. However, given that measurement error is accounted for, the half-normal with no adjustments becomes the best model. This result is interesting on its own, suggesting that an otherwise (assumed true) simpler detection function became more complex (more parameters needed to describe it) due to measurement error. Given a error model with 35% CV, we would expect that if ignored, abundance was overestimated, hence the probability of detection underestimated. Therefore, the fact that $\hat{P}$ decreases when the errors are accounted for, considering the HN alone, was surprising, although the much larger AIC value associated with the HN model while not accounting for the errors is a clear indication that such model was not a good fit to the data to begin with.

A standard kernel smooth of the 999 bootstrap P pseudo-estimates for the best model, for each of the 2 approaches, illustrates well the effect of ignoring measurement error (Figure 3.8).

Recall that abundance estimates are proportional to $1/\hat{P}$. Hence, an abundance

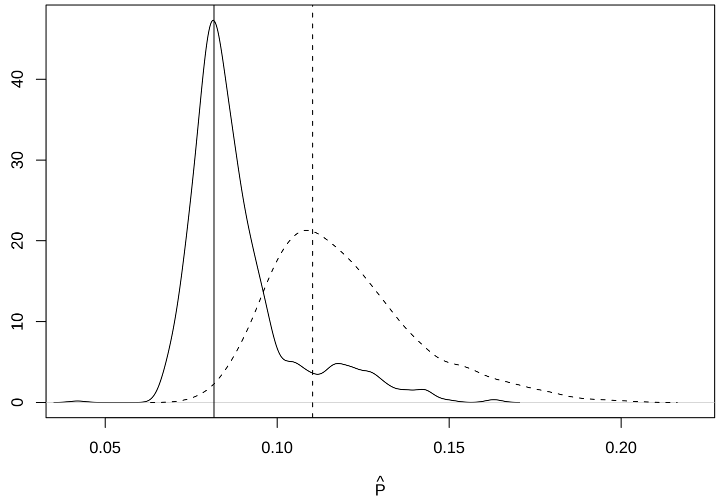

Figure 3.8: Kernel density estimates for the bootstrap pseudo-estimates of P from the NASS87 cue counting survey data, considering the half-normal+cosine detection function ignoring measurement error (solid line) and the half-normal detection function with the 35% CV gamma measurement error incorporated in the likelihood (dotted line). Vertical lines represent the corresponding $\hat{P}$ point estimates.

estimate for NASS87b ignoring measurement error would be around 34% higher (100[1/0.082-1/0.11]/[1/0.11]) than the best estimate accounting for measurement error.

This clearly reinforces the idea that, despite being frequently ignored, measurement error can have a considerable effect in distance sampling surveys, especially when point transect or cue counting methods are used.

3.5 Estimating the measurement error model

The most straightforward way to estimate the parameters of a measurement error model is to use a set of true and contaminated distances. Hopefully these will be collected during the actual survey, but if a separate experiment is set up with this objective in mind, the experiment conditions should mimic realistically the survey conditions. The danger of introducing bias in the main study due to inadequate error models estimated from independent data is discussed by Carroll *et al.* (1995, p. 11-12). However, one could also consider the estimation of the error measurement based on duplicate detections, given some additional assumptions on the error.

In this section several ideas relevant for the estimation of the measurement error from data are presented. After relating the error models involved in the PDF and likelihood approaches, I elaborate further on the estimation of the measurement error model using: pairs of true and contaminated distances considering an error model (1) separately or (2) in simultaneous with the detection function model, and (3) based on duplicate distances. While (1) and (3) are relevant for the PDF and likelihood approaches, (2) is only relevant for the latter. As before, the focus is on multiplicative error models.

3.5.1 Relation between R and $Y|X$

The PDF approach relies on specification of a model for the error R (and assuming that R and X are independent), while the likelihood approach relies on specification of an error model through the conditional distribution of Y given X. However, given a multiplicative error model (and R and X independent), there is no real difference in the underlying process, as assuming either one of these implicitly defines the other. By assuming a model for R, we constrain $Y|X$ to be in the same parametric family. One can obtain the conditional distribution of Y given X as a function of R distribution as

$$F_{Y|X}(y|X=x) = F_{XR|X}(y|X=x) = P[XR < y|X=x] = P[R < \frac{y}{x}] = F_R(\frac{y}{x}) \tag{3.46}$$

which leads to the corresponding *pdf* being

$$f_{Y|X}(y) = \frac{d\ F_{Y|X}(y)}{dy} = \frac{d\ F_R(y/x)}{dy} = \frac{1}{x} f_R(\frac{y}{x}). \tag{3.47}$$

Two special cases are presented. If the errors have Gaussian distribution, $R \sim N(\mu, \sigma)$, then equation 3.47 leads to

$$f_{Y|X}(y) = \frac{1}{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}[\frac{\mu - y/x}{\sigma}]^2} = \frac{1}{\sqrt{2\pi(\sigma x)^2}} e^{-\frac{1}{2}[\frac{\mu x - y}{\sigma x}]^2} \tag{3.48}$$

which means $Y|X = x \sim N(\mu x, \sigma x)$. It is worth noticing that care must be taken when using Gaussian errors in this context, as negative distances are likely unless models are constrained to have small variance. The only reason I presented it here as a special case is because others have used Gaussian errors in this context (e.g. Chen, 1998).

On the other hand, if the errors are gamma, $R \sim Ga(\sigma, b)$, then equation 3.47 leads to

$$f_{Y|X}(y) = \frac{1}{x} \frac{(\frac{y}{x})^{b-1} e^{-\frac{y/x}{\sigma}}}{\Gamma(b)\sigma^b} = \frac{y^{b-1} e^{-\frac{y}{\sigma x}}}{\Gamma(b)(\sigma x)^b} \tag{3.49}$$

which shows $Y|X = x \sim Ga(x\sigma, b)$.

3.5.2 Error model estimation using pairs of true and contaminated distances

Given a sample of (X, Y) data, we can assume a model for the measurement error R, obtain a sample of R's $(\frac{Y}{X})$ and then proceed to maximize a likelihood based on the assumed model, from which model parameters are estimated. Predefined functions like *fitdistr* in software R (MASS library) could be used. For some models, like the Gaussian, closed form estimators for the parameters of interest are readily available.

On the other hand, and especially if more complex scenarios are involved, for example if errors are also a function of other covariates, one could estimate the error model parameters in a regression context, using standard available software.

If Gaussian or gamma models are used it is also possible to fit them using standard general linear models, e.g. using functions *lm* or *glm* in software R, to estimate the parameters of the error model, by regressing Y on X with the appropriate link function and family model, and a zero intercept model. Note that if a more complicated measurement error structure is assumed, namely with a constant bias ξ independent of the true distance, say $Y = \xi + XR$, this framework allows to estimate it by using a non zero intercept model, but I have not considered that case here.

For the Gaussian error model, the standard *lm* function with weights= X^{-2} (given that the variance is proportional to the square of the true observations) can be used. If the regression coefficient is significantly different from 1, there is evidence for biased errors. The residual standard error of the regression is an estimate of the Gaussian error standard deviation.

For the gamma, the *glm* function will provide parameter estimates, by using the gamma family with the identity link. The regression coefficient will be the product of the shape and scale parameters (recall that in the gamma, the mean value is the product of the shape and scale parameters). The shape parameter can be estimated from the dispersion parameter (scale=1/dispersion) in the *glm* output. Note, however, that although the mean value is adequately estimated, the *glm* function, with family specified as gamma, uses an approximation which only provides a crude estimate of the shape parameter, and the *gamma.shape* (or *gamma.dispersion*) function in the *MASS* library should be used instead for better quality estimates. If the regression coefficient is not significantly different from 1 the null hypothesis that the gamma error is unbiased is not rejected.

3.5.3 Error model estimation using duplicate data

For the case of $MRDS$ (i.e. multiple observers) methods, we usually have a sample of contaminated distances Y. But for those animals which were duplicates, i.e. detected from two platforms, we have a sample of duplicate detections, $Y_j = (Y_{j1}, Y_{j2}),\ j = 1, 2, ..., J$. Intuitively, there seems to be information in these pairs of distances about the measurement error, which could be used to estimate the measurement error model. See Burnham *et al.* (2004, p. 375-376) for details on incorporating a measurement error model directly in a $MRDS$ analysis.

To make use of the duplicates information (without the true distances), we need to consider some further simplifying assumptions (note that some of these can be tested by comparing e.g. number of times distances are larger for one or for the other platform), namely that:

- the measurement error is common to both platforms;
- the two platforms are unbiased (i.e. $\mathbb{E}(R) = 1$), or at least the same common bias can be assumed (but then, if that bias is unknown, it is not enough to know the variance of the measurement error to correct for its effect);
- errors are independent across platforms;
- measurement error is not dependent on a detection being a duplicate or not;
- the animal is detected by both observers at the same true distance.

Given that we assume the errors to be unbiased, we just need to estimate a variance for R. Two different estimation procedures are possible, as described below:

- Given X, the $var(Y)$ is $X^2 var(R)$. Hence, if we calculate the variance within each pair Y_j, those should be an estimate of $X_j^2 var(R)$. Therefore, the slope of the regression line of X^2 on $var(Y_j)$ is an estimate of the variance of R. Note that because the true distance X is not known, we need to estimate it by the mean of the duplicates.

- On the other hand, defining the scaled difference of the j^{th} duplicate distances to be

$$Y_j^{ds} = \frac{Y_{j1} - Y_{j2}}{\frac{Y_{j1}+Y_{j2}}{2}} \tag{3.50}$$

we have that $\mathbb{E}[Y_j^{ds}]$=0 and $var[Y_j^{ds}] = 2var(R)$, hence half the estimated variance of Y_j^{ds} is an estimate of the variance of R.

A simulation example of the use of this second method to estimate the variance of Gaussian and gamma multiplicative errors is shown in figure 3.9. This plot is merely for illustration purposes: I considered to have 10000 pairs of duplicate distances for each CV shown, a clearly artificial example, but which removed most of the Monte Carlo variation for visual display.

However, as it can be seen in figure 3.9, there is some bias for larger CV's, present for both Gaussian and gamma errors. Nonetheless the bias tends to be small even for moderate CV's. The bias pattern is very similar for the first method (results not shown).

3.5.4 Using a joint likelihood for detection function and measurement error model estimation

In the simulation and example sections of the likelihood approach, the proposed methods were used conditional on a given error model, assumed known. In the first case because the objective was to access the methods performance under a favorable scenario, and in the second case because there was no experimental data to estimate the

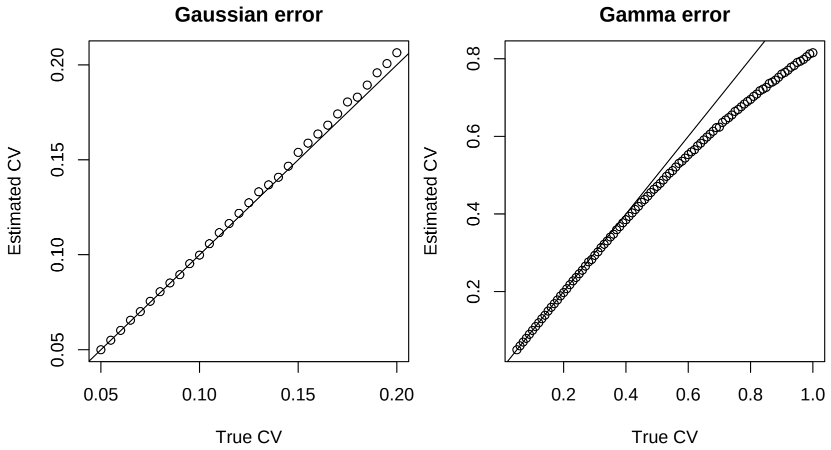

Figure 3.9: Using duplicate distances to estimate the error model: estimated coefficient of variation (CV) as a function of true CV of gaussian (left plot) and gamma (right plot) error models. The estimated CV was obtained using half the variance of the standardized duplicate differences (see text for further details). The $y = x$ line, which would indicate no bias in the estimation of the CV, is shown for reference.

error model parameters. Hence, the results were obtained through the maximization of the likelihood equation 3.43, where $\underline{\gamma}$ is assumed known.

Nonetheless, because a known measurement error model is in practice an artificial scenario, a more interesting case will be the one where the measurement error model parameters are also estimated from the data.

To assess the performance of the methods under such conditions, some extra simulations were carried out, considering for illustration one of the survey scenarios previously used in section 3.4.2: point transects, with $\mathbb{E}(n)$=300 and HN detection (σ=0.25), unbiased gamma errors with 30% CV. In addition to the (on average) 300 detected distances, a sample of (on average) 32 pairs of true and contaminated distances, with the same measurement error as in the survey, was used for the estimation of the measurement error model. As before, 500 data sets were simulated, and the

Table 3.9: Results of simulation exercise to compare the different approaches to estimate density without assuming the measurement error model parameters as known. Shown are the mean % bias and corresponding CV of estimates for different estimation strategies: $Y_{TM}CDS$ - ignoring the measurement error; $Y_{TM}LIK$ - true error model parameter used; $Y_{TM}LIK_{3.40}$ - error model parameter estimated using equation 3.40; $Y_{TM}LIK_{3.44}$ - joint likelihood equation 3.44 used for simultaneous estimation of error model and detection function parameters.

	$Y_{TM}CDS$	$Y_{TM}LIK$	$Y_{TM}LIK_{3.40}$	$Y_{TM}LIK_{3.44}$
% bias	-8.20	0.099	0.451	0.633
CV	0.094	0.091	0.144	0.144

data analyzed considering the error distances using: (1) true error model; (2) error model parameter estimated by a conditional likelihood (equation 3.40), then using equation 3.43, conditional on the estimates obtained for the error model parameter, to get the detection function parameter and (3) error model parameters estimated by the joint likelihood (equation 3.44). The observed bias and CV's of the 500 estimates are shown in table 3.9; also shown for comparison is the bias corresponding to the $Y_{TM}CDS$ estimates.

This initial set of simulations suggests that by estimating the model parameters the bias remains low. Nonetheless, and as one would expect, the variance in density estimates increases when the estimation of measurement error model is necessary. Being considerably faster to run, and easier to implement, the conditional likelihood seems to produce similar results to the joint likelihood. This is likely because the information from the error model comes from the experiment data (pairs of true and observed distances), while most of the information about the detection function comes from the survey data. Hence treating the two sets of information separately seems to have little impact in the final estimates.

3.6 Special measurement error cases

Due to their characteristics, two special cases of measurement error deserve a closer look: (1) heaping, which corresponds to rounding distances to favored values, and (2) cases in which errors occur mainly at large distances, but smaller distances can be measured accurately.

3.6.1 Heaping: rounding to favored distances

A chapter addressing measurement error in distance sampling would not be complete without a reference to heaping, an often reported effect of rounding to somewhat preferred distances (e.g. Anderson and Pospahala, 1970; Rosenstock *et al.*, 2002). If the method for obtaining distances does not produce a precise measurement and there is any subjectivity involved in the process (e.g. visual estimation of distances), it is inevitable that the histogram of detected distances will show some distances that occur much more frequently than what would be expected, while some others are rarely recorded, if at all. An example of data with substantial heaping, collected as part of a hare point transect survey[4], is shown in figure 3.10.

Favored distances are usually multiples of 1, 5, 10 or 100, depending on the scale of measurements being made. Humans have a preference for round numbers, and in the absence of any better alternative, those are used. Although very frequent in real data, heaping has to be strong to have a clear effect on density estimates. Even the analysis of the raw data, if heaping is not severe, should not be problematic. If it is thought heaping might have a considerable influence in the results, we can use judicious grouping of distances so that heaped values are approximately at the center of the distance intervals (Buckland *et al.*, 2001, p. 109-110).

For illustration of the consequences of heaping, a small simulation example follows.

[4] See Acknowledgements section for proper credit on the use of this data set.

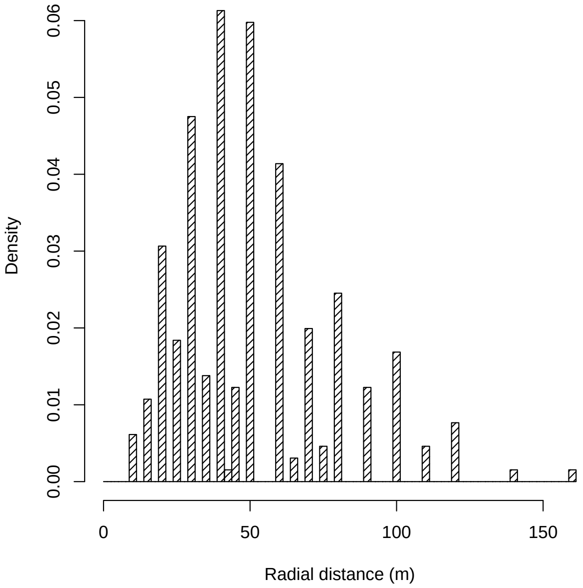

Figure 3.10: A real life point transect data example (hares in Northern Ireland), showing strong heaping at multiples of 5 and 10 meters.

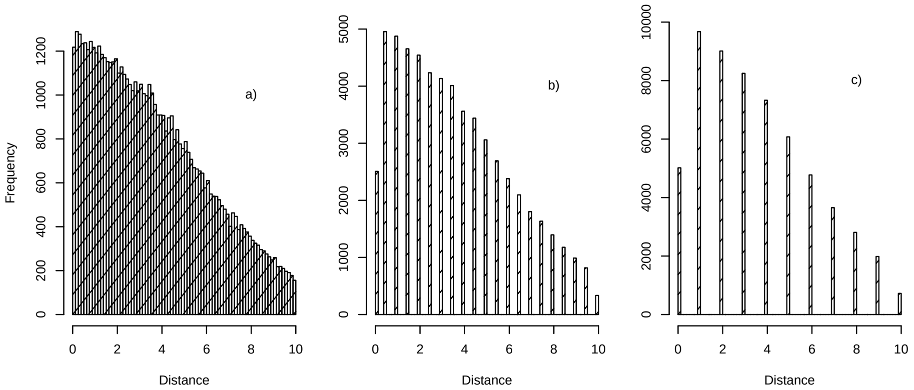

Figure 3.11: Histograms of simulated data for testing the effect of analyzing distance data ignoring heaping. a) Original data; b) Distances recorded as the closest multiple of 0.5 with respect to original distance; c) Distances recorded as the closest multiple of 1 with respect to original distance. Note that in b) and c) the first histogram bar is (roughly) half the height of the second because it corresponds to values heaped at 0, i.e. $X < 0.25$ for b) and $X < 0.5$ for c), while all the other bars span values over 0.5 or 1 meter intervals, respectively for b) and c).

Using the same simulations as in section 3.3.1.4 (100 line transect data sets), I have introduced two levels of heaping in the original data (Figure 3.11). All distances were recorded as the closest multiple of 0.5 (strong heaping) and closest multiple of 1 (very strong heaping), and analyzed the data using the automated procedure described before. A note about figure 3.11b,c: because the first histogram bar corresponds to true distances < 0.5 m or < 0.25 m, while all the others represent true distances in the vicinity on either side of the corresponding number, the first bar is only accounting for half the width of the others, and hence it looks approximately half the height of the second bar.

The consequences of heaping for either density estimates or associated variance were negligible (Table 3.10).

Even if heaping does not lead to large changes in density estimates, it can be

Table 3.10: Results of simulation exercise to assess the effect of heaping in the estimation of density. Mean estimated density $\hat{D}$ and respective CV for the analysis performed in Distance 4 considering the true distances and distances with strong and very strong heaping. True density is 100 animals/ha.

Analysis	$\hat{D}$	CV
True distances	99.02	0.0554
Strong heaping	98.92	0.0563
Very strong heaping	99.30	0.0562

responsible for some lack of independence between detections, and so special care must be given to the interpretation of goodness of fit measures. An example is presented in Marques *et al.* (in press).

However, heaping can be considered more like a hint for other problems, rather than a problem on its own. The presence of heaping usually means that the method used to measure distances was poor, and in such cases we can only hope that other errors, both random and systematic, were avoided.

The worst problem occurs when considerable heaping at 0 is present in the data, leading to a spiked detection function. This could be expected under several scenarios, like: (1) marine mammal surveys, where animals are detected at large distances from the observation platform, and rounding angles to 0 leads to heaping of perpendicular distances at 0; (2) Surveys on paths along which visibility is very good, coupled with animal movement, like flying birds on road surveys or small mammals along transects cut in dense forest; (3) incorrect definition of the measurement to be made, say a cluster recorded at 0 distance if any of the cluster members are on the line, when the true distance recorded should be the distance to the center of the cluster and (4) an imprecise definition of the transect. This might preclude appropriate estimation, usually leading to overestimation of abundance (provided the spike is really an artifact). The best way to deal with this problem is to avoid it, since at (and close to) zero, distances should not be difficult to measure. It is important to plot your data at an early stage of data collection, to identify potential problems; it is much simpler to identify and remove heaping at the data collection stage than to salvage the data analysis.

Smearing was introduced by Butterworth (1982) as an attempt to deal with heaping (not only in sighting distances but also in the recording of angles). The idea is to replace the preferred distances by plausible distances, derived by sampling from the vicinity of the original distance in a sensible way. Several options have been proposed (e.g. Hammond, 1984; Buckland and Anganuzzi, 1988). Ideally, one would want to sample proportionally to the true detection function, in the vicinity from which the heaped data was actually coming, leading to the new values being more often larger than the original heaped values; however, the rounding error tends to increase with distance, which means that more often a larger value is rounded down than a smaller value rounded up to a preferred value, and the two tend to cancel out. This led to Buckland and Anganuzzi (1988) recommendation that a uniform smearing be used, as it is much simpler and not necessarily worse. The choice of smearing parameters has been, so far, based on *ad hoc* methods.

It seems difficult to come up with simple models that represent heaping in an adequate way. Nonetheless, provided an experiment was designed to assess it, it seems to be possible to estimate an error model for heaping, namely using intervals rather than exact distances and estimating the probabilities of a distance being placed in given intervals, given the original interval the distance was in. Because it is unlikely for the presence of heaping without more general measurement error issues, it does not seem satisfactory to pursue methods which deal solely with this issue, rather than methods which deal with general measurement error issues and to some extent also accommodate heaping.

3.6.2 Errors exclusively at large distances

In many practical situations, the accurate measurement of larger distances might be harder than the measurement of small ones. This occurs for a number of reasons, including because it is harder and more time consuming to go to the actual detection location, or by having vegetation or other obstacles between the sighting and the line or point, or because one is not exactly sure of where the animal is/was when seen, animals at larger distances are not detected visually, etc.

It seems therefore useful to understand what would happen to distance sampling estimates if the measurement error was such that it was small/negligible for distances smaller than x^*, while for distances larger than x^* it followed some measurement error model as the ones presented in previous sections.

Hence, under this conceptual setting, the distances with measurement error could be modelled as

$$Y = \begin{cases} X & 0 \leq x \leq x^* \\ (X - x^*)R + x^* & x^* < x \leq \infty. \end{cases}$$

Note this model means that for line transects $f_X(0) = f_Y(0)$ and for point transects $h_X(0) = h_Y(0)$, hence no bias should arise, provided that the model used for the detection function is flexible enough to model the composite detection function (resulting from the true detection function and the measurement error). As x^* tends to 0 this model tends to the standard $Y = XR$ formulation. The larger x^*, the smaller the influence of the errors should be on the estimated density.

However, the situation might not be as clear cut in practice. A simulation exercise follows, using the same simulation settings as in section 3.4.2, for HN with $\sigma = 0.125$

and $E(n) = 300$, with unbiased gamma errors generated in the same way as before (30% CV case), but now using the error model described above. An example of this measurement error, considering $x^* = 0.05$, is shown in figure 3.12a. To assess the influence of different x^*, I increased x^* from 0 to around the maximum observed distance, hence going from the previous scenario of measurement error for all distances to no measurement error. The best detection function model was chosen accordingly to minimum AIC, and for each x^* I performed 1000 simulations. The observed % bias as a function of x^* is shown both for point (Figure 3.12c) and line transects (Figure 3.12d), with the corresponding bias considering the true distances used as a control. Note the bias in the control, caused by truncation issues (T. Marques, unpublished data).

Note that the values of percentage bias observed here for $x^* = 0$ correspond to those in Tables 3.6 and 3.7 for $Y_{BM}CDS$ (respectively 4.85 and 16.66; they are not exactly the same due to Monte Carlo variation).

For both line and point transects, just as was seen before for the case when all distances are contaminated, for small x^*, the effect of the random error leads to abundance overestimation, but there is a point at which the effect changes sign and we observe a small underestimation of abundance.

This example represents only an illustration, and should not be taken as a general result. Considering the example on its own, it might seem that we would be better off if we were able to record all distances with no error up to say 0.06 (bias ≈ 0) rather than to say 0.12. But this apparent counterintuitive conclusion stems from the fact that we are looking at one particular combination of simulation settings. Over all (infinite) possible situations, it seems likely that the larger x^*, the smaller the effect of the measurement error in distances larger than x^*.

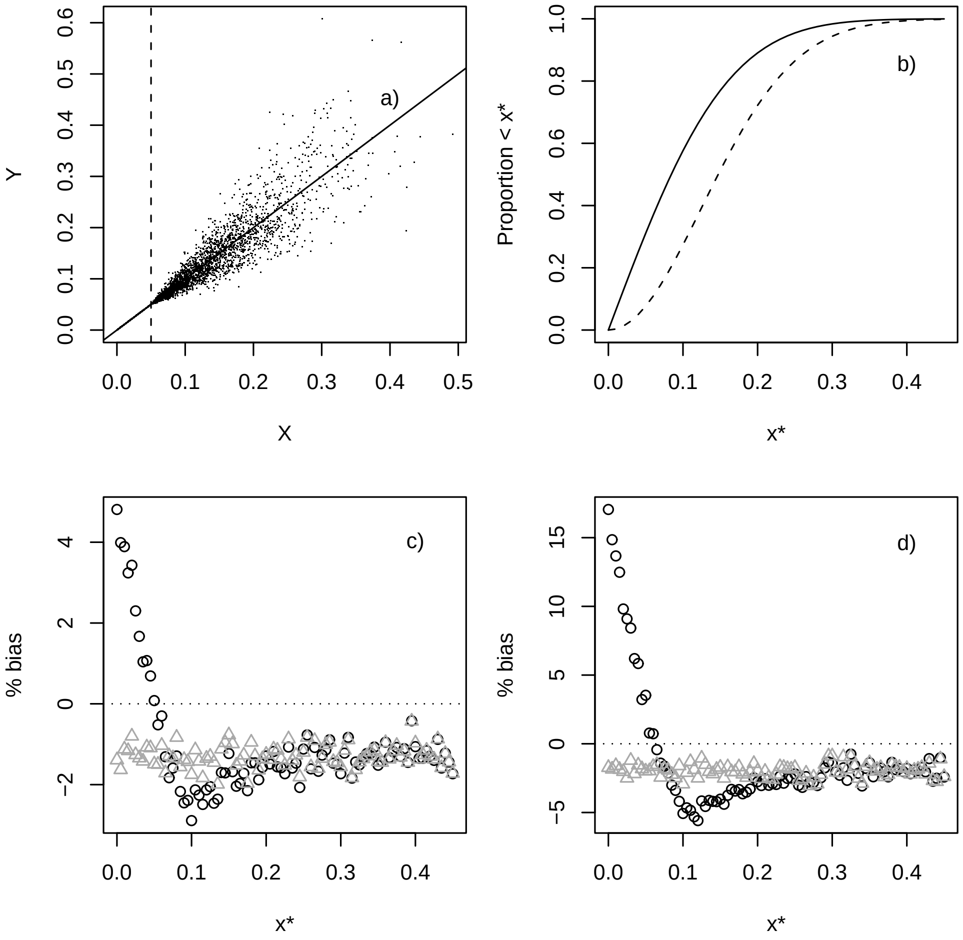

Figure 3.12: Errors exclusively at large distances. a) An example of errors exclusively at large distances, considering the minimum true distance at which errors might occur $x^* = 0.05$ and a gamma unbiased model with 30% CV. b) Proportion of detected distances smaller than x^*, for point (dashed line) and line transects (solid line), as a function of x^*. Results of the simulation exercise to evaluate the effect of errors at large distances, considering the same setting as in a). c) Line transects; d) Point transects. Percentage bias as a function of the value of x^*, with x^* varying over a range of values that corresponds to measurement error for all distances to virtually no measurement error. Black circles represent the mean % bias of the analysis based on error distances, grey triangles the mean % bias of a control analysis of the data without error.

Point transects seem to require a larger x^* to measurement error having a negligible impact. Everything else being the same, a given x^* corresponds to a smaller proportion of the data being error free for point transects compared with line transects, due to underlying geometry of points versus lines (cf. figure 3.12b). From the simulation presented here, it appears that over 75% of the distances need to be error free so that measurement error is negligible, but nonetheless, and not surprisingly, in general an increase in x^* leads to bias reduction.

This *all or none* measurement error example is clearly an oversimplification of real life situations, but it helps understanding the consequences of the measurement error, and how reducing the measurement error, even if only at smaller distances, is beneficial. Practitioners usually like to adopt standard protocols for data collection, with the same procedure always used for a given task. Nonetheless, if all distances can not be measured accurately, but the smaller ones might be, given some (hopefully not too large) extra amount of work, then this is a better option than to measure them all imprecisely just to be consistent in the measurement method.

3.7 Discussion

Despite being a potential severe source of bias, both for line transects but especially for point transect and cue counting methods, measurement error is still treated as a lesser problem compared to the other key assumptions. Although it stands to reason that for most survey settings $g(0) < 1$ or movement might be responsible for the larger proportion of the potential bias, the results in this chapter clearly demonstrate that the problem should not be ignored.

The proposed methods to deal with measurement error, the PDF and the likelihood approach, represent alternatives to the suggestions of Chen (1998) and Chen and Cowling (2001) as possible ways of tackling the problem. The main advantage of the PDF approach is that it can be implemented in the commonly used software Distance, using multipliers. However, the poor performance of this method, especially for point transects, compared to the likelihood approach, does not permit its recommendation for general use.

The disappointing results for the PDF approach arise due to the combination of two factors: model mispecification, and models used for the detection function that are not flexible enough. The simulation exercises consider a true detection function for data generation; then either that detection function or one chosen by AIC from a set of functions is used to model the data contaminated with the measurement error. However one of the consequences of measurement error is that the resulting distribution of detected distances no longer shares the distribution of the original detection function, as might be seen from equation 3.7. Using this same equation one can actually look at the shape of the resulting detection function and *pdf* in the presence of measurement error. As an example, I consider the contamination of a half-normal detection function with a gamma unbiased error with 50% CV (Figure 3.13). It is clear that the resulting detection functions, given the contaminated errors, are no longer half-normal, with a shape more closely resembling an exponential distribution. This in turn means that the models for the detection function can not fit the resulting shape appropriately, hence the bias, predicted by the PDF approach, does not occur in practice. Hence, when the correction is applied we tend to over correct and we get a bias going in the opposite direction (in this case, underestimation due to the correction versus the expected overestimation ignoring the correction, cf. Tables 3.4-3.7). If one were to use as candidate models for the contaminated distances something like the negative exponential, or even a (very flexible) kernel density estimator (as used by e.g. Mack and Quang, 1998), the PDF approach might be less biased. Note that this means also that the PFD approach is more sample size dependent, because the more data you have available the more likely you are to select a more flexible model that will predict the expected bias. This will be the case when using AIC as a criterion to choose amongst models, as with less data simpler models will be chosen even if the overall fit might not be great.

The results from the simulation exercise were encouraging with respect to the likelihood approach. A key advantage of it relative to the PDF approach is that provided we can define a $Y|X$ model, we are not restricted to a specific error structure (like the multiplicative one), and hence this is a much more flexible framework, that can be integrated with more complex scenarios, like the $MRDS$ methods (see Burnham *et al.*, 2004, p. 375-376 for details). While far superior, the likelihood approach might be difficult to implement. The likelihoods involved are not straightforward, very slow to run compared with conventional methods, and currently available software does not provide the user with the possibility of including an error model. Multiple covariate likelihoods were not implemented, and although these do not present any major additional analytic issues, even more numerical problems are likely to occur, as tends to happen when $MCDS$ methods are used, compared to their CDS counterparts.

A possible much simpler, and preferably exploratory alternative, given that some information on the error structure can be obtained, is as follows: (1) Use the n contaminated error distances to estimate a detection function; then repeat many times: (2) Simulate n distances from such detection function, and contaminate them with errors with the assumed error structure; (3) Analyze the true and contaminated distances, and calculate the difference between the estimated densities. This will allow a first feeling for the influence of measurement error in the situation at hand. From these differences, one could calculate a mean correction value and its variance, and use it as a multiplier on the original analysis. Note that with this procedure, we implicitly assume the effect of the errors on the estimated detection function based

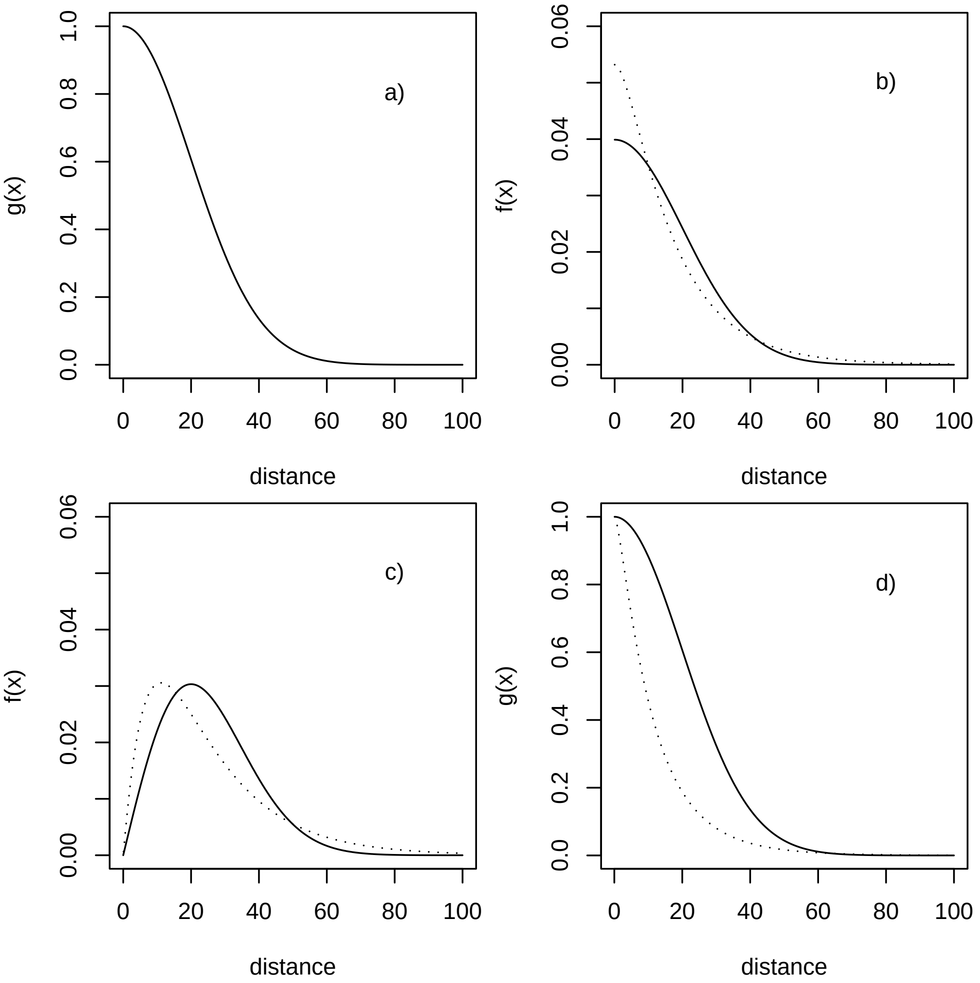

Figure 3.13: The effect of a measurement error example (unbiased gamma with 50% CV) in the detected distances probability density function (*pdf*). a) True detection function; b) Corresponding *pdf* (solid line) and contaminated distances *pdf* (dotted line) for line transects; c) Corresponding *pdf* (solid line) and contaminated distances *pdf* (dotted line) for point transects; d) True (solid line) and contaminated (dotted line) apparent detection function for point transects.

on the error distances is the same as the effect of the errors on the true detection function, which might not be true. A sensitivity analysis, using slightly different detection functions from the one estimated based on the contaminated distances, will enable one to determine if that is a plausible assumption. A similar procedure is described in Marques *et al.* (2006), to assess the potential bias of using an imprecise distance measurement method (combining inclinometer and altimeter recordings to get distance measurements) compared to a more precise one (GPS device with distance measurements obtained in a geographic information system).

When collecting information about measurement error using an experiment rather than doing it during the actual survey, extreme care must be taken to ensure that the error process during the survey can be estimated. As an example, Williams *et al.* (2007) report that "...Preliminary evidence suggested that an observer differed in the ability to judge distance to fixed, continuously-visible cues and ephemeral, cetacean cues, which calls into the question the common practice of using marker buoys as cetacean proxies in distance-estimation experiments...", which clearly shows that this might be harder to do than one may anticipate.

I hope that this work has helped to raise awareness of the potential impact of measurement error in the detected distances on distance sampling abundance estimates. Investigators conducting distance sampling studies should always consider the training of observers at measuring distances (as in e.g. Baldi *et al.*, 2001; Tobias and Seddon, 2002), which has been shown to reduce error and bias in distance estimates (Alldredge *et al.*, in press). Additionally, I recommend the use of best possible technology to aid in the process of distance measurement.

Some measurement error is likely to be present even if the best technology at hand is used. However, under many circumstances, it should be possible to reduce the error magnitude to such levels that have negligible consequences on the estimates.

The results presented suggest that unbiased measurement error with CV's of around 10% or less should have negligible effect on estimates (certainly so when one considers all the other possible sources of bias in a real life distance sampling survey). If one can achieve such a standard then that is undoubtedly the recommended approach. The take home message is that it is better to avoid errors at the data collection stage than to rely on analysis methods to cope with measurement errors. The everyday improvement in technology leads us to believe that the failure of this assumption might become less frequent in the future.

The option to correct for bias due to measurement error needs to be seen in the light of the corresponding bias versus variance trade off (e.g. Carroll *et al.*, 1995). Because we need extra parameters to model measurement error, an increase in the variance of the estimates when accounting for measurement error is likely. Nonetheless, provided additional data is available to estimate the error model, one might actually obtain better precision once the error model is included in the analysis. Considering that obtaining the smallest variance possible for density/abundance estimates is a primary objective of any survey, the option to avoid additional analysis methods if modifications to the field procedures are likely to avoid these problems is further justified.

Measurement error and $g(0) < 1$ share the fact that people tend to be overly confident on their capabilities, because there is no obvious way to get feedback on one's performance. In face of the results in this chapter, it seems reasonable for any large survey that uses distance sampling to attempt at least some assessment of the measurement error involved. Using Mark Twain's words, “It ain't what you don't know that gets you into trouble. It's what you know for sure that just ain't so”.

Chapter 4

Non random allocation of samplers and density gradients

4.1 Introduction

A key assumption of CDS methods is that the distribution of animals in the covered area, with respect to distance from the samplers, is known by design. This means that, once the geometry of the samplers is accounted for, one can interpret the numbers of detected animals at different distances from the transect as a result of the detection function alone, hence allowing the estimation of detection probabilities and the corresponding distance sampling abundance estimators.

The justification for this known distribution of animals by design rests on the fact that *a sufficient number of samplers are placed over the study area independently of animal distribution, following some previously specified random design.* Hence we can anticipate problems if we consider either a small number of transects or if transects are placed in such a way that their locations are not independent of the animals' locations.

Under given logistical constraints, transect allocation might be far from random. The use of platforms of opportunity, like fishing or tourist boats which scientists might use for conducting cetacean line transects (e.g. Marques, 2001), is an example where that might happen. These vessels travel along given routes because there are more fish, or more whales, or it is the shortest path between two ports, etc. This fact alone suggests that inferences from such data, with respect to a wider region, must be made with extreme caution.

The use of spatial models and multiple covariate distance sampling has been put forward as a way to circumvent some of the shortcomings of this sort of data (e.g. Hedley, 2000; Marques, 2001). With a spatial model, we predict abundance surfaces as a function of spatially explicit covariates, so provided we have good coverage we might still derive accurate estimates by integrating under the estimated density surface, even if transect placement was non random. Note however that if spatial coverage is bad, estimation to a wider region becomes an extrapolation exercise, and again care should be taken to decide to what extent it is valid.

The methods developed in subsequent chapters are relevant in situations in which transects are laid along a linear feature, with respect to which animals might present a *density gradient*. For convenience I stress the distinction between an absolute density gradient, $D(x)$, which represents density at a given distance from the linear feature, and a relative density gradient, which is only proportional to it, $d(x)$. Most commonly *density gradient* is used to describe the latter, the shape of the distribution of animals with respect to the linear feature involved, rather than absolute density, but when it is not clear from the context which is being referred to, the distinction will be made explicitly.

Although laying transects along linear features that might have an impact on the distribution of the population of interest is discouraged, it is nonetheless a situation that occurs much more frequently than one would hope, and as the subsequent material will show, it can have a strong impact in the quality of the estimates. Nonetheless, many published studies either ignore the potential problem (e.g. Simons *et al.*, 2006), briefly mention it (e.g. Ogutu *et al.*, 2006), or acknowledge it properly but do not present solutions to solve it, considering the advantages to outweigh the disadvantages (e.g. Heydon *et al.*, 2000). Presumably, the problem is even more pervasive in the grey literature, which reinforces the need to study the problem.

This kind of sampling, referred to as "convenience sampling" by Anderson (2001), is usually criticized, but under some circumstances might be the only approach possible to sample a population. Examples include point transects along a road, sampling along a river, cetacean surveys from shore, bird radar data or krill acoustic surveys. In each of these, it stands to reason that the underlying distribution of animals with respect to samplers is not known *a priori*, because it depends on how the animals respond respectively to the road, the river, the shore line, their altitude distribution and depth distribution. Therefore, strictly speaking, CDS methods should not be used in these cases.

When we rely on sampling to estimate abundance, we usually can think of a two stage process. In the first stage we estimate abundance in the covered area(s), and in the second stage we scale that estimate up to the wider survey region. The use of non-random samplers has a clear impact on the second stage, because if the sampling was not random there is no justification for scaling up on the basis of design properties: there is no valid argument to justify that the covered areas are representative of the wider survey region we are usually interested in making inferences about (e.g. Hiby and Krishna, 2001). This will be so, irrespective of the method used to sample the covered area, be it distance sampling, mark recapture, plot sampling, or any other alternative. As described above for spatial methods, one alternative might be to choose a model based approach to perform the inferences over the wider region. The bias related to this second stage is not the key focus of this work, because this problem is common to all sampling methods, and this thesis is focussed on distance sampling.

Nonetheless, the use of a spatial model might help in obtaining better estimates for uncovered areas (see section 4.4).

On the other hand, and more relevant for the methods in this thesis, if we consider distance sampling, the density gradient with respect to the transects might also be responsible for bias in the first stage, estimation of abundance in the covered areas. Note this is in contrast with a situation in which an approach is used that does not rely on a known distribution of animals with respect to the samplers, like plot sampling or mark recapture.

Hence, in the case of distance sampling, it is likely that the use of these linear features for transect placement might promote a density gradient in the available distances for detection with respect to samplers, which precludes the estimation even for the covered areas, because the $APTA$ condition does not hold. This is described in detail in the next section.

4.2 Bias arising due to density gradients

Consider that line transects are placed along a road (i.e., the road is the transect), and interest lies in 3 different species: the first ignores the road (SpIg), the second avoids the road (SpAv), and the third is strongly attracted by the road (SpAt). The use of words like avoidance and attraction solely describe the distribution of animals with respect to the road, and do not assume any particular animal behaviour towards the road.

Hypothetical density gradients $d(x)$ for each of the species involved are shown in figure 4.1 (left column). For simplicity, the detection function is the same for all species (Figure 4.1, center column). Assume that these density gradients are ignored, the survey is conducted, and the observed distances are recorded as usual (Figure 4.1, right column). SpIg should be estimated without bias (at least for the vicinity of the road), since there is no effect at all from the road in the process, the usual uniform distribution holds and CDS is adequate. However, given that the detected distances are interpreted as a function of detectability alone - the distances available for detection are assumed uniform - SpAt and SpAv estimates will be biased. The probability of detection of SpAt will be underestimated, as it seems that many animals away from the line were missed, when in fact they were not there to detect to begin with. On the other hand, the probability of detection of SpAv might be overestimated, as it seems that a larger proportion of animals was seen at large distances, when in fact they were more animals there to begin with. Hence, for the covered area, SpAv abundance is underestimated and SpAt abundance is overestimated.

It is therefore clear that ignoring the effect of availability of distances could lead to potential bias, if the $APTA$ condition does not hold. In such cases, alternative methods that either incorporate the availability process or do not depend on it should be considered.

Note that, even for truly random transects with respect to the animal distribution, if a small number of samplers is used, the $APTA$ condition should hold on average (over many realizations of the design), but a single realization of the survey process may result in an actual (unobserved) availability process which is far from optimal. This is important, because for any given survey only a single realization of the design is usually available, hence one should aim to have a large number of lines.

More formally, what is at stake is that when considering CDS methods, detection probability is assumed to be given by

$$P = \int_0^w g(v)\pi(v)dv = \mathbb{E}[g(v)] \tag{4.1}$$

where the detection function parameters are estimated from the conventional likelihood,

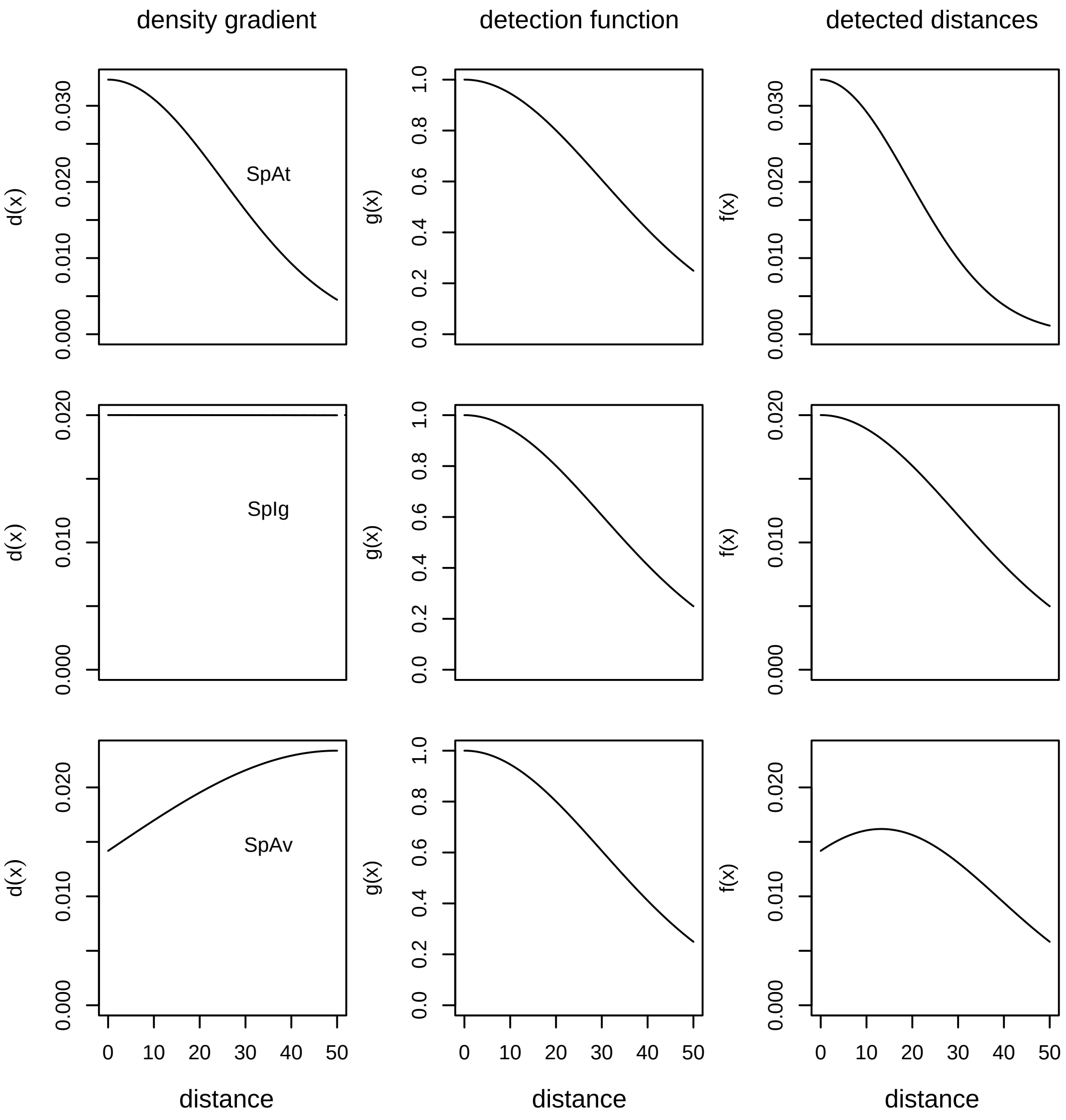

Figure 4.1: Examples of possible density gradients (left column) obtained by placing transects along a road, considering a common detection function (center column), and corresponding perceived detection function (right column) by assuming a uniform distribution of distances available for detection. The first row represents a species attracted to a road, SpAt, the second a species which ignores the road, SpIg, and the third a species which avoids the road, SpAv.

$$\mathcal{L}(\underline{\phi_1}|\underline{v}) = \prod_{i=1}^{n} f(v_i) = \prod_{i=1}^{n} \frac{g(v_i)\pi(v_i)}{\int_0^w g(v)\pi(v)\ dv} \tag{4.2}$$

but further it is assumed that $\pi(v)$ is known, and equal to $\frac{1}{w}$ or $\frac{2r}{w^2}$, depending respectively on considering line or point transects. Given that it is not the case and a density gradient exists, ignoring it and naively using these expressions leads to biased estimators. Ways to deal with this problem are the focus of the remainder of this thesis.

In the presence of a density gradient, two strategies exist to avoid the corresponding bias: (1) consider a method that does not depend on the density gradient or (2) use a method that, instead of assuming it known, models the density gradient and incorporates it in the estimation procedure. Of course, to implement these, one needs some extra data, because the traditional distances alone represent a total confounding of the availability for detection and detection process.

In the next section, previous work done by others related to the first strategy is covered. This is followed by a section presenting some ideas useful for the second strategy, that might be used in conjunction with methods described in subsequent chapters. In those chapters implementations of methods considering specific survey settings are presented.

4.3 Using multiple observers to avoid assumptions on the availability of distances

The first strategy to deal with the presence of a density gradient with respect to the transects is to use a method that does not depend on it.

As mentioned before, approaches based on combining mark-recapture and distance sampling data have been put forward to relax the $g(0) = 1$ assumption, by allowing

$g(0)$ estimation. Some of the methods used in that context are also relevant here, as described briefly below.

I borrow heavily here from the material and notation of Laake and Borchers (2004). I no longer assume detection certain on the line or point, and to distinguish from the previous context the detection function is defined as $p(v) = p(0)g(v)$, with $p(0) \leq 0$, but with $g(0) = 1$. Consider Laake and Borchers (2004) equation 6.8 (which assumes no covariates other than distance, but see their equation 6.16 for those),

$$\mathcal{L}_w = \prod_{i=1}^{n} \frac{Pr[\underline{w}_i|v_i]}{p.(v_i)} \tag{4.3}$$

where $Pr[\underline{w}_i|v_i]$ is the probability of an individual capture history and $p.(v_i) = p_1(v_i) + p_2(v_i) - p_1(v_i)p_{1|2}(v_i)$ is the probability that at least one observer detects animal i, where $p_j(v_i)$ is the probability that observer j $(j = 1, 2)$ detects animal i given that the animal is at distance v_i, and $p_{j|3-j}(v_i)$ is the conditional probability of observer j detecting an animal, given that observer $3 - j$ detected it; note that independence of the 2 observers detections is not assumed. The probability of individual capture histories $\underline{w}_i$ is given by

$Pr[\underline{w}_i = (1, 0)|v_i] = p_1(v_i)[1 - p_{2|1}(v_i)]$

$Pr[\underline{w}_i = (0, 1)|v_i] = [1 - p_{1|2}(v_i)]p_2(v_i)$

$Pr[\underline{w}_i = (1, 1)|v_i] = p_1(v_i)p_2(v_i)$.

This likelihood allows the estimation of the detection function without having to specify any $\pi(v)$. The maximization of this likelihood can be done either through direct numerical maximization or using a standard logistic regression, with an iterative procedure to account for the animals missed by both observers, as proposed by Buckland *et al.* (1993b) and described in detail in Burnham *et al.* (2004, p. 361-363).

Note that, without allowing the direct estimation of $\pi(v)$, given the estimation of $p.(v)$ and the observed distances, one can obtain $\pi(v)$ non-parametrically (see Laake and Borchers, 2004, p. 145, for an example). In the following this was not implemented or investigated further, because the focus was intended to be on distance sampling approaches rather than mark-recapture approaches (with distance as covariate). See Fletcher and Hutto (2006) for an example with double counting of birds from a river by using a canoe as the survey platform. Note that in this example the inference is restricted to birds along the river (the covered area) and no extrapolation to wider areas is attempted.

4.4 Incorporating an availability model

4.4.1 Building on *MRDS* approaches

It is important to make a distinction between capture-recapture methods with distance as a covariate, and a true *MRDS* approach. The method described in the previous section, corresponding to the former approach, uses a mark-recapture likelihood ($\mathcal{L}_w$) but not the distance sampling likelihood[1] ($\mathcal{L}_v$), not capitalizing on the knowledge of $\pi(v)$ (Laake and Borchers, 2004, p. 147).

An appealing idea is to combine the information contained in the mark-recapture data with the information combined in the distance sampling data. While there might be disadvantages from using only the $\mathcal{L}_w$ likelihood component, given that the information contained in the $\mathcal{L}_v$ component is not used to provide information about the detection function shape (and hence not being in the realm of what Laake and Borchers (2004) refer to as true *MRDS*), combining $\mathcal{L}_w$ and $\mathcal{L}_v$ comes at a cost. We need to either (1) assume $\pi(v)$ as known, and then we can consider point independence, or (2) we need to assume full independence and a model for $\pi(v)$ (see Laake and Borchers, 2004, p. 117-120, for details about full/point independence).

The first case is preferred if the main issue is that animals on the transect are

[1] which corresponds to some variant of equation 4.2.

missed, but the survey design is such that $\pi(v)$ can be safely assumed as known. Then one can proceed as usual in a point independence analysis, maximizing separately $\mathcal{L}_w$ and $\mathcal{L}_v$ (e.g. Borchers *et al.*, 2006). If one assumes point independence, which is recommended from the $MRDS$ perspective alone, we can no longer estimate $\pi(v)$, without additional info on $\pi(v)$, because once again we rely on the information about the shape of the detection function being in $\mathcal{L}_v$ only, and hence confounded with $\pi(v)$.

The second case might be preferred if the main problem is the unknown $\pi(v)$ due to design issues. One can assume a model for $\pi(v)$ and estimate the corresponding parameters by numerical maximization of $\mathcal{L}_w\mathcal{L}_v$, but the price to pay is that point independence must be abandoned. To the best of my knowledge this has not been implemented in practice, likely because full independence is not realistic under most scenarios and, until now, when $MRDS$ methods are used the main concern is undetected animals on the line rather than unknown density gradients.

Hence in the presence of a density gradient, using only $\mathcal{L}_w$, or $\mathcal{L}_w\mathcal{L}_v$ with full independence, will allow inferences in the covered area without assuming known $\pi(v)$, even if it this approach opens the door to the usual problems associated with conventional mark-recapture methods (such as unmodelled heterogeneity). It is nonetheless worth noting that by combining some of the ideas presented in the remaining chapters with $MRDS$, one can potentially estimate $\pi(v)$ with the point independence assumption, provided some additional information on $\pi(v)$ is available (see Buckland *et al.* (in press) and also the general discussion chapter for possible extensions to these methods, incorporating estimation of $\pi(v)$ along with $\mathcal{L}_w\mathcal{L}_v$).

4.4.2 Considering an explicit density gradient with respect to the samplers

As described in the previous section, considering $\mathcal{L}_w\mathcal{L}_v$ with full independence, is conceptually different from using $\mathcal{L}_w$ alone. Rather than considering methods that do not assume known $\pi(v)$, an alternative approach to deal with the presence of a density gradient is to use a method that explicitly accounts for it. This corresponds to assuming a model for the density gradient and being able to collect appropriate data such that the model parameters can be estimated. This will be the focus of chapters 5, 6 and 7 which build on existing methods to estimate abundance in the covered areas.

Unlike for conventional methods, it is useful to describe the spatial setting under which the methods are derived. I assume that transects (either lines or points) are placed along a linear structure with respect to which the animals present a density gradient. Hence, after transect placement, one can describe the animal distribution as a function of the distance to the transects.

Consider a Cartesian coordinate system, where $D(x, y)$ represents density at location (x, y). Note that in the following I assume no measurement errors, and in contrast with the previous chapter y is no longer used to represent an error distance. I assume without loss of generality that the linear feature runs along the y dimension, at $x = 0$. The focus is in the vicinity of the linear feature, defined as a strip of width $2w$ centered on it, with mean density D. Further, I conceptually fold the process along the linear feature. Consider $D(x)$ to be the absolute density gradient, describing density at a distance x from the linear feature. Note that $D(x) \propto d(x)$, where $d(x)$ is the relative density gradient that represents availability for detection as a function of distance from the line. For identifiability I define $d(x)$ to be a *pdf*, i.e.

$$d(x) = \frac{D(x)}{\alpha} = \frac{D(x)}{\int\limits_0^w D(x)dx} \tag{4.4}$$

which means that in the case of line transects $d(x) = \pi(x)$, but this is not true for point transects due to the point geometry (see section 6.2.2, equation 6.4, for further details).

Assume for a moment that we were able to obtain an estimate for the mean density in the vicinity of the linear feature (i.e., the area within w of the line), $\hat{D}$, and that as part of the estimation process we were also able to estimate $d(x)$. Chapters 5 and 6 provide examples for estimation in the covered areas along linear features. Provided a random selection of units along the linear feature is used, estimates are valid for the vicinity of the linear feature on the basis of the design.

Considering $D(x)$ to be the rate of a point process on $(0, w)$, it follows that

$$\mathbb{E}(N) = \int\limits_0^w D(x)dx \tag{4.5}$$

and hence

$$\mathbb{E}(D) = \frac{\int\limits_0^w D(x)dx}{w} \tag{4.6}$$

which, considering equation 4.4, leads to the following estimator for $D(x)$

$$\hat{D}(x) = \hat{D}w\hat{d}(x). \tag{4.7}$$

The intuitive idea underlying the algebra is that the area under the function $D(x)$, the animal density as a function of distance x from the transects, is the same as the area under the constant function of height D, the average density in the vicinity of the road (see Figure 4.2), and hence is estimable given $\hat{D}$ and $\hat{d}(x)$.

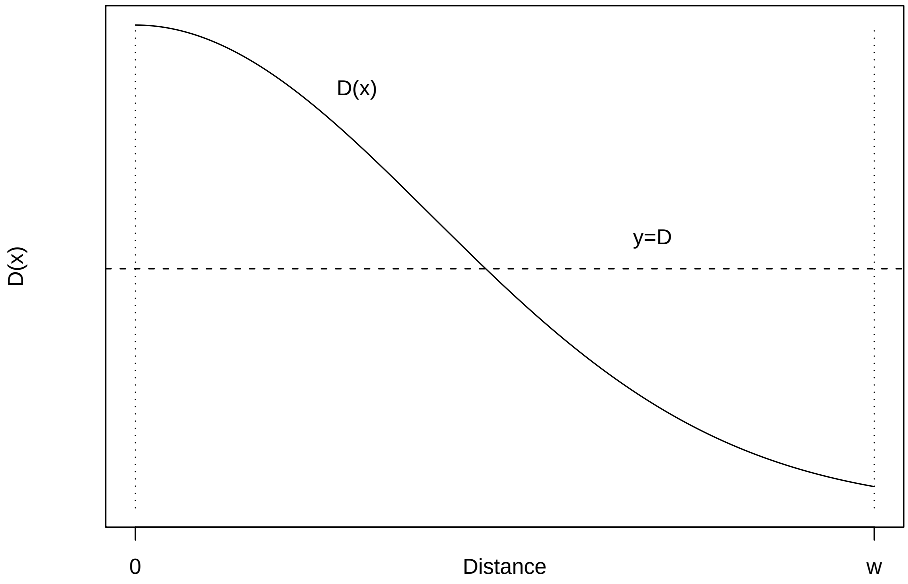

Figure 4.2: Schematic representation of the relation between the average density in the vicinity of the linear feature, D, and the function that describes density at a distance x from the linear feature, $D(x)$. Note the area under $D(x)$ and under the line $y = D$ must be equal, allowing the estimation of $D(x)$ from $\hat{d}(x)$ and $\hat{D}$.

In contrast with the design-based approach used in conventional methods, this leads to a model-based scaling up of the density estimate for the wider survey region. Despite being based on extrapolation of the estimated density gradient beyond the range of distances in the data, if one can assume that the influence of the linear feature has disappeared at w, then $\hat{D}(w)$ is a better density estimate, for the wider region not covered, than $\hat{D}$. The inspection of the estimated $d(x)$ will allow assessment of whether or not that is a reasonable assumption.

An estimate for the abundance in the survey region is the sum of the abundances in the area in the vicinity of the linear feature (N_c, obtained with methods like those described in chapters 5 and 6) and in the area away from the linear feature (N_c^-, which if the influence of the linear feature has disappeared at w, might be estimated as described in the previous paragraphs), i.e.

$$\hat{N} = \hat{N}_c + \hat{N}_c^- = A_c\hat{D} + A_c^-\hat{D}(w) \tag{4.8}$$

where A_c represents the covered area and A_c^- represents the area not covered.

Variance estimates could be obtained using the nonparametric bootstrap. The sampling units should be lines (e.g. independent stretches of road), even if the samplers are points along lines, as points along a given line will typically be non independent.

In the next two chapters, I present two distinct ways to estimate abundance in the vicinity of linear features. Despite being generally discouraged, this is a relatively common situation in published distance sampling applications, mostly for logistical reasons. Numerous examples of such a setting, with transects placed along existing roads or trails, can be found in the literature (e.g. Borralho *et al.*, 1996; Baldi *et al.*, 2001; Hiby and Krishna, 2001; Boano and Toffoli, 2002; Ruette *et al.*, 2003; Ward *et al.*, 2004; Dörgeloh, 2005; Bårdsen and Fox, 2006). For line transects, placed along a linear feature to which the animals respond to, we are dependent on at least some secondary transects to provide information on the density gradient (Chapter 5). By contrast, for point transects along such a linear feature, we can use information in the sighting angles with respect to the linear feature to provide information about the density gradient (Chapter 6). It is then natural to combine these two approaches, by having survey points along a linear structure and also some secondary transects perpendicular to the structure to provide additional information about the density gradient. This, as well as some other possible extensions, are presented in Chapter 7.

Chapter 5

Line transects with density gradients

5.1 Introduction

In this chapter I consider the case where a distance sampling line transect survey is performed along a linear structure with respect to which the animals might present a density gradient.

Because the information on availability for detection (the density gradient) and detectability is completely confounded in the conventional perpendicular distances, and hence the usual uniformity assumed by design is not necessarily reasonable, bias will arise in the presence of a density gradient. Therefore one must collect additional independent information about one of these two processes if they are to be disentangled.

The rationale behind the methods described in this chapter is to have at least some secondary transects perpendicular to the linear structure, from which one can directly estimate the density gradient. Usually it will be clear from the context, but if confusion is possible the transects along the linear feature are referred as the primary transects.

The layout of this chapter reflects the learning process that I went through during the PhD, in fact through my research in general. I start by presenting in section 5.2 an initial approach to the problem, dealing with the problem in a more intuitive way, using material from Marques and Buckland (2005). Then, section 5.3 places the problem into a wider general framework, in which parametric models are assumed for the processes involved (detection and availability) and maximum likelihood is used to estimate model parameters and draw inferences. In section 5.4 a brief discussion about these methods is presented, opening the door to the subsequent chapter, where the focus is on a similar survey setting but this time with point transects as samplers.

5.2 A naïve approach to the problem

In this section, nonparametric kernel estimators are used to model the density gradient and the product of the density gradient and detection function. Intuitively, by dividing these resulting models one should obtain an estimate of the detection function. Rather than a general approach, this constituted a first attempt to deal with the problem.

5.2.1 Proposed methods

In the conventional setting, the detected distances are the result of two functions, reflecting two independent processes: the detection process and the availability process. The two processes are separated by assuming uniformity in the availability process. This implies that the shape of the probability density function *pdf* of detected distances is solely a function of the detection process, allowing their separation.

The methods proposed here require the collection of independent data about the availability process. This information can be obtained by placing some secondary transects perpendicular to the primary ones. These transects should be at least as long as the largest distances that are expected to be observed in the primary transects (or, if inferences for areas away from the linear feature are important, transects should extend to the region where the effect of the linear feature has disappeared). Modelling the distances along the secondary transects allows us to separate the information due to detection and due to availability on the primary transects, leading to corrected estimators of abundance.

Recall that, as described in section 4.4, modelling the density gradient not only allows estimation in the covered area (of size $2wL$), but it might also help to obtain estimates in a wider survey region that we usually are interested in.

5.2.1.1 Estimating the number of animals in the covered area

The distances detected on the primary transects are the combination of the density gradient, $d(x)$, and the detection function, $g(x)$. Recall that for line transects $d(x) = \pi(x)$, and hence in this chapter only $\pi(x)$ is used. Defining $f(x)$ as the *pdf* of the detected distances, it follows that

$$f(x) = \frac{1}{\mathbb{E}[g(x)]}\pi(x)g(x) \tag{5.1}$$

where $\frac{1}{\mathbb{E}[g(x)]}$ represents a normalizing constant ensuring that $f(x)$ is a *pdf*. Rearranging the previous expression leads to

$$\mathbb{E}[g(x)] = \frac{\pi(x)g(x)}{f(x)}, \quad \forall\ x \in (0, w) \tag{5.2}$$

and assuming $g(0) = 1$ (as in conventional methods), it follows that

$$\mathbb{E}[g(x)] = \frac{\pi(0)}{f(0)}. \tag{5.3}$$

Once we obtain estimates of $\pi(x)$ and $f(x)$, we can estimate $\mathbb{E}[g(x)]$, and hence derive an estimator for $g(x)$, as

$$\hat{g}(x) = \hat{\mathbb{E}}[g(x)]\frac{\hat{f}(x)}{\hat{\pi}(x)}. \tag{5.4}$$

On the other hand, the usual estimator for the number of animals in the covered area is given by

$$\hat{N_c} = \frac{n}{\hat{P}} = \frac{n}{\hat{\mathbb{E}}[g(x)]} = n\frac{\hat{f}(0)}{\hat{\pi}(0)}. \tag{5.5}$$

The data collected on the primary transects can be used to estimate $f(x)$, while $\pi(x)$ can be estimated using the distances, from the linear feature, to animals seen in the secondary transects. If the influence of the linear feature on density at distance w from the linear feature has disappeared, estimation for the wider survey region might be obtained by using equation 4.8.

5.2.1.2 Modelling the *pdf*'s

The procedures described require the modelling of $f(x)$ and $\pi(x)$, both *pdf*'s. Resulting from the product of availability and detectability, the shape of $f(x)$ is unknown, and nonparametric approaches are a plausible alternative.

I considered here the use of a Gaussian kernel to model the *pdf*'s. Seber (1986) seems to have been the first to suggest kernels to model the *pdf* of detection distances, and they were subsequently used in that context for example by Buckland (1992) and Chen (1996). Mack and Quang (1998) present a comprehensive treatment of kernel methods for line and point transects. The key reference for kernels as a general method for density estimation is still Silverman (1986).

To avoid problems at the boundary $x = 0$, I used signed distances rather than positive distances (as in e.g. Mack and Quang, 1998).

5.2.2 Illustration of the methods by simulation

To evaluate the properties of the proposed methods a simple simulation exercise was conducted, in which both sampling and modeling were simulated to mimic a real life survey where transects were laid along a linear feature, and a set of secondary transects perpendicular to it was available. To illustrate the methods the density gradient was such that animals tended to be considerably more common near the linear feature than away from it.

5.2.2.1 Simulation settings

A rectangular survey region with 40 ha (1000 m by 400 m) was considered, longitudinally crossed at the center by a single primary transect (L=1000 m). The truncation distance w was assumed to be 100 m, i.e., the covered area is half of the total survey region. A few 100 m perpendicular secondary transects were randomly set, independently on both sides of the primary transect, from which a direct sample of $\pi(x)$ could be obtained. The total number of animals in the survey region was $N = 2000$, of which 1000 were distributed at random according to a constant background density, and 1000 were distributed randomly as a function of the distance from the transect, according to a Gaussian distribution, with $\sigma = 20$ (Figure 5.1). This ensures that at $w = 100$ the density gradient is constant, allowing inferences outside the covered area as described in the previous section. The detection function was assumed to be half-normal. I considered two different detection functions, $\sigma = 25$ and $\sigma = 45$ (Figure 5.1).

Using a rejection method, the sampling process was simulated for the primary transect, and the detected distances used to estimate $f(x)$. To allow the estimation of $\pi(x)$ 200 distances were collected on the secondary transects.

Both *pdf*'s were estimated using a Gaussian kernel. The choice of the appropriate

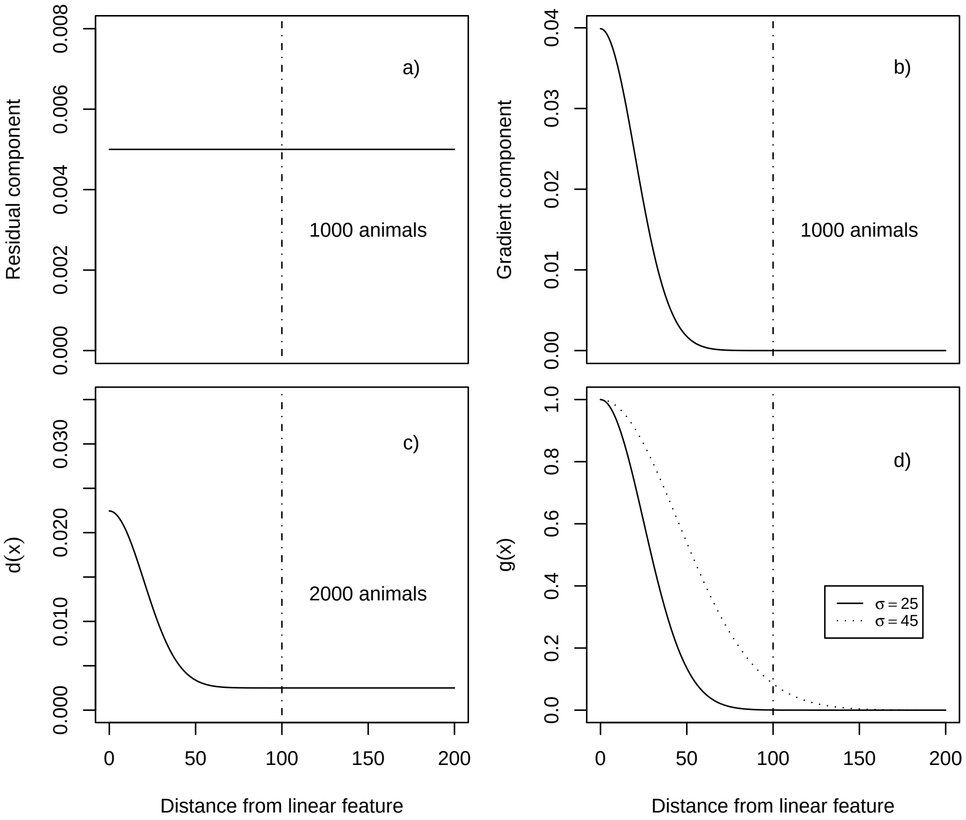

Figure 5.1: Graphical representation of the density gradient and detection function considered in simulations. a) Background density component (corresponds to density away from the linear feature along which the transect was placed); b) Gradient component; c) The density gradient resulting from combining a) and b); d) The detection functions used. The dot-dash line represents the boundary of the covered area.

bandwidth was outside the scope of this work, so I considered arbitrary values, chosen to be close to the default values of the *density* function in R (R Development Core Team, 2004), when analyzing some trial sets of simulated data.

Estimates of abundance for the covered area and survey region were obtained using expression 5.5 and 4.8, respectively. For direct comparison, the estimates obtained using the conventional methods ignoring the gradient, i.e. using directly expression 2.11 based solely on the data from the primary transects, were also calculated. Each situation was repeated 1000 times.

5.2.2.2 Simulation results

The abundance estimates resulting from the simulation exercise are shown in figure 5.2. When ignoring the gradient we observe an overestimation of density that can be over 100% of the true value. This overestimation is a function of the slope of the density gradient, and increasing the slope should lead to even worse results. We can see that incorporating the availability model in the analysis removes most of the bias present when it is ignored. The price to pay is a decrease in precision. The comparison of coefficients of variation (CV) for the corrected and uncorrected estimators shows that the bias decrease is accomplished at the expense of at least 2 times higher CV's, but considering $RMSE$ as an overall measure of quality, the proposed methods substantially outperform the conventional approach (Table 5.1).

5.3 A likelihood approach to the problem

The previous approach fails to capitalize directly on the fact that there is a process common to the data collected in the primary and secondary transects, namely the density gradient.

As an alternative obvious option (although after the fact, most good options tend

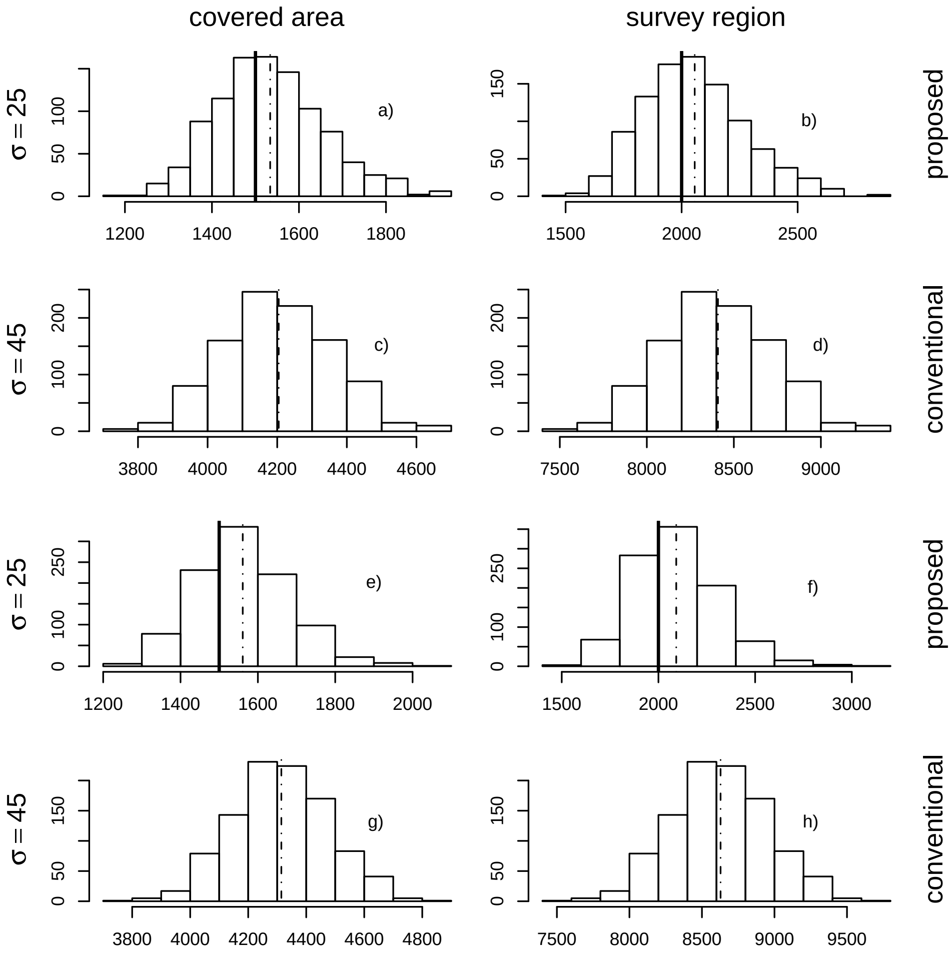

Figure 5.2: Results from the simulation exercise to evaluate the methods proposed to account for the density gradient using the information on secondary transects. a) to d) corresponds to the half-normal detection function with σ=25, while e) to h) to σ=45. a) and e) Estimated N in the covered area, using the kernel based methods; b) and f) Estimated N in the survey region, using the kernel based methods; c) and g) Estimated N in the covered area, using conventional methods; d) and h) Estimated N in the survey region, using conventional methods. Solid line: true N. Dot-dash line: mean estimated N.

Table 5.1: Coefficient of variation (CV) and root mean square error ($RMSE$) for the 1000 estimates obtained by the conventional and proposed methods to deal with density gradients based on kernels, for both detection functions considered in the simulation exercise.

		σ=25		σ=45	
	Method	Covered area	Survey region	Covered area	Survey region
CV	Conventional	0.0376	0.0376	0.0379	0.0379
	Proposed	0.0802	0.1059	0.0770	0.1040
$RMSE$	Conventional	2708.961	6416.485	2818.953	6636.477
	Proposed	127.5622	224.7865	134.6694	236.153

to seem obvious!), once this problem is cast into a general likelihood framework it is straightforward to estimate parameters for both models at once and proceed as usual.

A further advantage of such an approach is that it can be generalized to more complicated situations, capitalizing on the likelihood framework.

5.3.1 Proposed methods

Note that by specifying $\mathbb{E}[g(x)]$, equation 5.1 can be rewritten as

$$f(x) = \frac{g(x)\pi(x)}{\int_0^w g(x)\pi(x)dx}. \tag{5.6}$$

Consider that the detection function and density gradient are indexed respectively by the parameter vectors $\underline{\phi_1}$ and $\underline{\phi_2}$. Because the two functions only appear as a product, this *pdf* is not enough to derive a likelihood to estimate parameters from both the detection function and the density gradient, using the distances $\underline{x_p} = (x_{p,1}, x_{p,2}, ..., x_{p,n_p})$ collected along the primary transects. However, the distances $\underline{x_s} = (x_{s,1}, x_{s,2}, ..., x_{s,n_s})$, collected along the secondary transects, provide a sample from $\pi(x)$, which can be used to estimate the density gradient parameter using

$$\mathcal{L}(\underline{\phi_2}|\underline{x_s}) = \prod_{j=1}^{n_s} \pi(x_{s,j}). \tag{5.7}$$

Assuming that the data collected in the primary and secondary transects are independent (which must be enforced by proper design and survey methods), a likelihood combining the two data sources follows naturally, which will allow the joint estimation of the parameters of both processes, as

$$\mathcal{L}(\underline{\phi_1}, \underline{\phi_2}|\underline{x_p}, \underline{x_s}) = \mathcal{L}(\underline{\phi_1}|\underline{\phi_2}, \underline{x_p})\mathcal{L}(\underline{\phi_2}|\underline{x_s}) = \prod_{i=1}^{n_p} \frac{g(x_{p,i})\pi(x_{p,i})}{\int_0^w g(x)\pi(x)dx} \prod_{j=1}^{n_s} \pi(x_{s,j}). \tag{5.8}$$

Given the parameter estimates, the usual estimate for P follows as

$$\hat{P} = \int_0^w \hat{g}(x)\hat{\pi}(x)dx \tag{5.9}$$

and we can therefore obtain the corresponding estimator for density (see equation 1.2). As before, given that we estimate the density gradient, we can use it to estimate density at a given distance from the linear feature, which leads to a model based abundance estimation for the wider survey region (see section 4.4, equation 4.8).

5.3.2 Illustration of the methods by simulation

5.3.2.1 Simulation settings

To provide a direct comparison with the performance of the initial approach, the same simulation settings as used in section 5.2.2 were considered.

Note that in the previous simulation section, the way in which animals were distributed was rather *ad hoc*, with half the animals being allocated according to a Poisson process (the background constant density) and half the animals distributed according to a half-normal density gradient. Because now we need to implement a likelihood, it is useful to view the resulting relative density gradient in a parametric form

$$d(x|\underline{\phi_2}) = \frac{\tau + \exp(-\frac{x^2}{2\sigma^2})}{\int_0^w \tau + \exp(-\frac{x^2}{2\sigma^2})dx} = \frac{\tau + \exp(-\frac{x^2}{2\sigma^2})}{\nu} \qquad (5.10)$$

where $\underline{\phi_2} = (\tau, \sigma)$ and τ is proportional to the background density, and hence $\frac{\alpha\tau}{\nu}$ is the density at distances for which the linear structure no longer has an influence on density.

5.3.2.2 Simulation results

The coefficient of variation and root mean square error of the estimates, for the methods proposed in this section, are contrasted in table 5.2 with those obtained for conventional methods and those proposed in section 5.2 based on kernel density estimation. Note that differences in the first two lines of this table, with respect to table 5.1, are only due to Monte Carlo variation.

The proposed methods based on maximum likelihood are slightly less biased and considerably more precise than those derived by the initial intuitive approach to the problem.

5.4 Discussion

The methods presented in this chapter represent a first attempt to estimate animal abundance using line transects in situations where animals present a density gradient with respect to these. Our results show that modelling the availability to detection is a possible approach to remove the bias present when such availability is not uniform. That modeling exercise can be done provided that some data can be collected in secondary transects perpendicular to the main ones, independently of the sighting distances obtained in the primary transects. The comparison with conventional

Table 5.2: Coefficient of variation (CV) and root mean square error ($RMSE$) for the 1000 estimates obtained by the conventional and proposed methods to deal with density gradients, both based on kernels and on maximum likelihood, for both detection functions considered in the simulation exercise.

		σ=25		σ=45	
	Method	Covered area	Survey region	Covered area	Survey region
CV	Conventional	0.0422	0.0422	0.0404	0.0404
	Kernel	0.0788	0.1063	0.0816	0.1081
	Likelihood	0.0460	0.0808	0.0456	0.0819
$RMSE$	Conventional	2827.9	6654.0	2819.4	6637.2
	Kernel	137.8	243.1	140.0	241.3
	Likelihood	71.8	163.5	71.5	166.4

methods also shows that considerable bias can occur if an existing density gradient is ignored. This bias will be a function of the density gradient, and if this is unknown, even the direction of the bias might be unknown, rendering estimates not only useless but potentially dangerous if used to support management decisions.

Some questions remain that need to be resolved to implement the initial approach to the problem. The main question is related to the options taken to estimate the *pdf*'s involved. Silverman (1986) shows that bias in kernel estimators is proportional to bandwidth. In principle, bandwidth would decrease as sample size increases, and that is the justification that authors use to claim that estimators are asymptotically unbiased. As large sample sizes in distance sampling are not common, the issue might be relevant here. The choice of different bandwidths for estimation of $\pi(x)$ and $f(x)$ might be responsible for different bias in estimating $\pi(0)$ and $f(0)$, and consequently bias in the abundance estimates. Silverman (1986) also shows that the bias is proportional to the second derivative of the function being estimated, which adds extra difficulty in the case of the *pdf*'s considered here, given that we are estimating their values at 0 distance, likely a function maximum. (Note however that in situations where density increases with distance from the line the maximum may be away from zero.) As the second derivative is negative at a maximum, some underestimation is expected, a fact observed in other work using kernels in a distance sampling context (e.g. Chen, 1996; Mack and Quang, 1998), but not clearly attributed to this fact. Note that in this case, we want to estimate κ (see expression 5.5), a quotient of two estimated quantities, and so the final bias is the result of the relative bias of its components. While in simulations arbitrary values were used for bandwidth, producing sensible results, some bias was still present after the correction, and further investigation is needed to address this cause.

Another question not dealt with explicitly here is the variance of the naïve estimator, fundamental for the application of the methods. The simplest approach is to use resampling methods, like the nonparametric bootstrap as described in Borchers *et al.* (2002), but nonetheless it should be possible to find approximate analytical expressions for the variance by using standard kernel theory results and the delta method.

The estimation of the *pdf*'s involved might be done using other methods. A possible alternative is the traditional semi-parametric approach introduced by Buckland (1992), as implemented in standard distance sampling software Distance, but without the monotonicity constraints. Other options include the use of series expansions, log-splines or splines.

The main inconvenience of this intuitive approach stems from the fact that the modeling of $f(x)$ and $\pi(x)$ is done independently. Because of that, sample perturbations mean that the estimated detection function might not be a monotonically decreasing function of distance (as one might expect), and probability of detection, at some distances, can be estimated as being higher than 1, which corresponds to an inadmissible estimate (although this is unlikely to happen in practice with reasonable sample size). This suggests that there might be advantages to finding a more parsimonious way to deal with the issues involved, and that is also why some of the issues have remained unanswered. Although an interesting first approach to the problem, it seems *a priori* unlikely that these kernel based methods would perform better than those based on maximum likelihood.

The initial approach to the problem ignored the use of any type of parametric form either for the detection function or the density gradient. Because of that, while the initial results of section 5.2.2 were interesting and showed that the approach was worth pursuing, the smaller bias was obtained at the cost of a much larger variance, which is not ideal.

The idea of extending the methods to a likelihood framework followed naturally, and the results shown indicate that there might be substantial advantages in doing so. Of course this approach is more appropriate if the investigators have a good idea of the potential shape of the density gradient. Even if several models could be tested, based on some model selection criteria, it is a sounder approach to have *a priori* a reasonable expectation about the form of the density gradient. When that knowledge is not available a priori, a possible intermediate approach is to use the data collected in the secondary transects only, with some nonparametric density estimation technique, say kernel density estimation, to get a first feel for the shape of the density gradient, and then choose flexible parametric (or semi-parametric) models that span shapes similar to those obtained by the nonparametric method.

A further advantage of embedding the methods in a likelihood framework is the easy generalization of these to more complicated situations. A simple example is the situation in which the detection function is not constant in the secondary transects, but it is a function of the distance to the primary transects. This might be the case if habitat changes as a function of distance from the linear feature. For the initial approach one might be forced to use only distances in the shoulder of the detection function from the secondary units, hence guaranteing that the animals considered are a representative sample of the density gradient. Cast in a likelihood framework, one can extend the simple formulation presented here, to include a distance sampling likelihood (i.e. accounting for a detection function) for the data collected in the secondary transects. This flexibility and further examples will be illustrated later in the thesis (see Chapters 7 and 8).

The main disadvantage of the methods presented is that the additional information must be obtained from secondary transects, when in many cases the rationale for using a non-random design in the first place is that it is difficult or impossible to travel away from the linear feature. Therefore it would be a tremendous advantage if one was able to derive methods which would only rely on the data gathered from the primary transects, but still cope with density gradients with respect to it. Whilst that is not simple to do for lines[1], I show in the next chapter a method based on point transects that has in this characteristic its main appeal.

[1] It is worth noting that it should be possible to derive such methods for line transects, given distance from the line and angle with respect to the linear feature at first sighting, together with the additional assumption that the detection function is independent of angle. The extent to which this assumption might hold in practice in the case of line transects, due to non uniform search patterns, might however cast some doubts over such procedure. Nonetheless, as suggested by Hans Skaugh (pers. communication), appropriate field methods might help insuring such assumption.

Chapter 6

Point transects with density gradients

6.1 Introduction

The methods in this chapter are intended to be used in situations similar to those of the previous chapter, i.e. when transects are placed along a linear feature, with respect to which the animals might respond to. However, this time point rather than line transects are used as (primary) transects. The original motivation for this work came from point transects placed along riparian habitats, but the illustrative example used is from a survey of Northern Ireland hares (*Lepus timidus hibernicus*) performed along roads[1]. The ideas should be relevant under any context where a density gradient exists with respect to a feature along which point transects are placed.

The rationale behind the methods is to use the angle of sighting with respect to the linear feature in addition to the usual radial sighting distance, to estimate simultaneously the parameters of both the detection function and the density gradient. This allows for a non uniform distribution of the animals with respect to the linear feature, unlike conventional methods that (erroneously in this case) assume a uniform distribution. The methods assume that the detection function is independent of the

[1] See Acknowledgements section for proper credit on the use of this data set.

sighting angle with respect to the linear feature.

Hence, unlike the methods presented in the previous chapter, one can obtain the necessary information to implement these methods without having to spend extra resources with additional data collection, a considerable advantage if these resources are scarce.

6.2 Proposed methods

6.2.1 Preliminaries

Consider k point transects laid along the linear feature, with the usual distance sampling assumptions holding (see section 2.5). Assume the detection function is, as usual for point transects, a function of the radial distance alone, with associated vector of unknown parameters $\underline{\phi_1}$.

Consider a Cartesian coordinate system, where $D(x, y)$ represents density at location (x, y). A linear feature is present at $x = 0$, and the focus is on a strip of width $2w$ centered on the linear feature (i.e., truncation distance $= w$), with mean density D.

Assume, without loss of generality (for estimation of the parameters of interest, see below), that point centers are all located at (0,0). After point placement, we can describe the relative location of animals in space with respect to this point. Let $r = \sqrt{x^2 + y^2}$ represent the radial distance to the animals from an observer at a point center, and $\theta = \arcsin \frac{x}{r}$ the sighting angle with respect to the linear feature (see Figure 6.1a).

To avoid having to account for circle geometry, I start by considering the square of side $2w$ centered on the point. The process is conceptually folded into one quadrant with respect to both axes $x = 0$ and $y = 0$. As before, consider $D(x)$ to be the absolute density gradient, representing density at distance x from the linear feature,

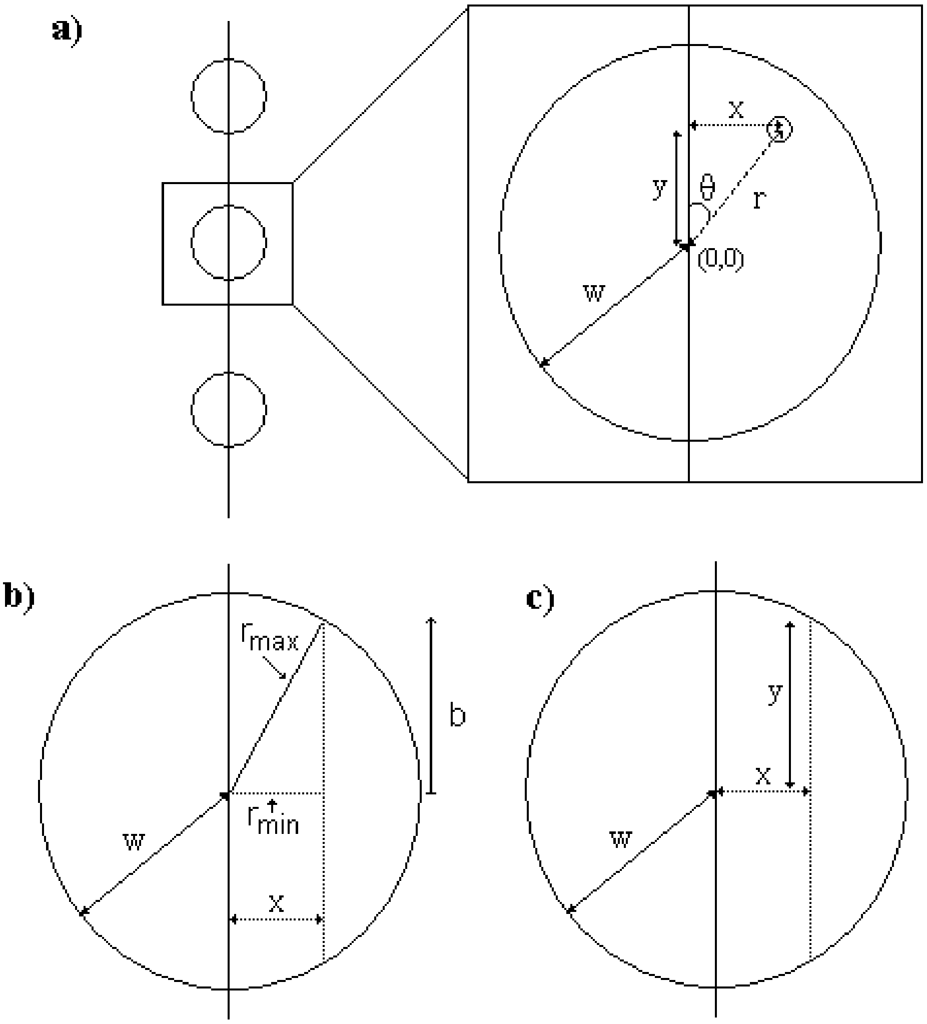

Figure 6.1: A linear feature along which point transects are placed, with the distances involved in the definition of the methods shown. a) Points along the linear feature. A point is blown up showing details of the distances to a given animal, represented as a circle (x, perpendicular distance with respect to the linear feature; y, distance along the linear feature; r, radial distance to the animal; θ, sighting angle with respect to the linear feature); b) Given $X = x$, r can take values between $r_{min} = x$ and $r_{max} = w$; c) Given $X = x$, y is assumed uniform in $(0, \sqrt{w^2 - x^2})$, for $0 \leq x \leq w$.

given by (once folded along the linear feature)

$$D(x) = \int_{-w}^{w} D(x,y)dy + \int_{-w}^{w} D(-x,y)dy, \quad 0 \leq x \leq w. \tag{6.1}$$

For simplicity, I work with the relative density gradient function $d(x) = D(x)/\alpha$, such that $d(x)$ is a *pdf* (this has no implication on the methods because, as it will become apparent later, there is no information in the likelihood used to allow the estimation of the proportionality constant). One will need to assume a model for the relative density gradient, with associated vector of unknown parameters $\underline{\phi_2}$. This parameter vector will be estimable from the data as described below. For readability, I often drop the density gradient and detection function parameters (respectively $\underline{\phi_2}$ and $\underline{\phi_1}$) in the notation.

No assumption is made about the animals' distribution along the linear feature, but I assume that, after point placement, $D(y)$ is a constant independent of y, i.e.

$$D(y) = \int_{-w}^{w} D(x,y)dx + \int_{-w}^{w} D(x,-y)dx = D, \quad 0 \leq y \leq w. \tag{6.2}$$

The right hand side of the equation holds if the y distribution on $(-w, w)$ is an odd function. For example, if trend in density in the y direction is assumed linear over the short distance $2w$, then this requirement is met. Random point placement along the line ensures, by design, that density within the covered squares of side $2w$ is representative of the $2w$ width strip centered on the linear feature.

Although usually the recorded data in the field would be radial distances (r) and sighting angles (θ), any two of (x, y, r, θ) are sufficient to implement the proposed methods.

6.2.2 Bivariate likelihood

As presented earlier, conventional point transect methods involve the maximization of a likelihood (with respect to $\underline{\phi_1}$ alone) based on the *pdf* of detected distances in the circle. A model is assumed for the detection function $g(r)$ and $\pi(r) = \frac{2r}{w^2}$ is assumed known by design. One can generalize this likelihood to a bivariate likelihood, which assumes that r and θ are independent and the detection does not depend on θ, leading to

$$\prod_{i=1}^{n} \pi(\theta_i) \frac{g(r_i)\pi(r_i)}{\int\limits_R g(r)\pi(r)dr} = \mathcal{L}_\theta \mathcal{L}_r. \tag{6.3}$$

Since $g(r)$ and $\pi(r)$ appear as a product, they can not separately be estimated from $\mathcal{L}_r$, the CDS likelihood. $\mathcal{L}_\theta$, not usually considered because there is no direct interest in $\pi(\theta)$, allows us to test whether $\pi(\theta)$ is uniform, which one would expect under CDS. Failure of such uniformity could indicate the need for the methods described in this chapter.

The basis of the methods proposed here is a bivariate *pdf* which uses the dependence between r and θ, a consequence of the non uniform density gradient, to allow the joint estimation of the parameters of $g(r)$ and $\pi(r)$.

Consider $\pi(x, y)$ to be the joint *pdf* of animal locations (x, y) (whether detected or not) in the quarter circle[2]. Given the definition of the relative density gradient, the *pdf* of the perpendicular distances (detected or not) in the quarter circle of radius w is

[2] Keep in mind that this process is considered in a quarter circle, rather than a full circle like for conventional point transects, due to the conceptual folding in the x and y dimensions. However, this has no practical impact on the estimation process.

$$\pi(x) = \frac{d(x)\varphi(x)}{\int\limits_0^w d(x)\varphi(x)dx}, \quad 0 \le x \le w \tag{6.4}$$

where $\varphi(x) = \frac{b}{w} = \frac{\sqrt{w^2-x^2}}{w}$, and b is the half-length of the chord parallel to the transect at perpendicular distance x from the transect (Figure 6.1b). Note that $\varphi(x)$ accounts for the circle geometry.

On the other hand, $\pi(y|x)$ is assumed uniform (by design, see equation 6.2), and hence

$$\pi(y|x) = \frac{1}{\sqrt{w^2 - x^2}}, \quad 0 \le y \le \sqrt{w^2 - x^2} \tag{6.5}$$

leading to the joint distribution

$$\pi(x, y) = \pi(y|x)\pi(x) = \frac{1}{\sqrt{w^2 - x^2}} \frac{d(x)\varphi(x)}{\int\limits_0^w d(x)\varphi(x)dx} = \frac{d(x)}{\int\limits_0^w d(x)\varphi(x)dx}. \tag{6.6}$$

The field data usually comprise radial distances and angles, hence it is convenient to represent equation 6.6 in terms of polar coordinates rather than Cartesian coordinates. Therefore, considering the random variable transformation $R = \sqrt{X^2 + Y^2}$ and $\theta = \arcsin \frac{X}{\sqrt{X^2+Y^2}}$, with inverse $X = R\sin\theta$ and $Y = R\cos\theta$, leads to

$$\pi(r, \theta) = \frac{rd(r\sin\theta)}{\int\limits_\theta d(r\sin\theta)\varphi(r\sin\theta)r\cos\theta d\theta}. \tag{6.7}$$

A change in the integration variable on the denominator $r\sin\theta = x$ (which just reverses the changes introduced by the above random variable transformation in the integral) leads to

$$\pi(r, \theta) = \frac{rd(r\sin\theta)}{\int\limits_0^w d(x)\varphi(x)dx}. \tag{6.8}$$

The advantage of this is obvious when we write down the joint *pdf* of radial distances and sighting angles (now for the detected animals only) as

$$f(r,\theta) = \frac{\pi(r,\theta)g(r)}{\int\limits_R \int\limits_\theta \pi(r,\theta)g(r)d\theta dr} \tag{6.9}$$

because it leads to a much simpler expression due to the integral in the denominator of expression 6.8 canceling out, leading to

$$f(r,\theta) = \frac{r\ d(x)g(r)}{\int\limits_0^w \int\limits_0^{\frac{\pi}{2}} r\ d(r\sin\theta)g(r)d\theta dr}. \tag{6.10}$$

Note that, apart from a constant that is not a function of the parameters, this turns out to be the same likelihood that one would end up with if starting from $\pi(x,r)$ rather than $\pi(r,\theta)$ (see appendix A for further details), hence leading naturally to the same parameter estimates.

6.2.3 Estimating density in the covered circles

Given the n observed (r,θ) pairs, assuming a parametric form for the density gradient and the detection function, we can use the joint distribution of θ and r to build a likelihood that can be maximized to estimate the unknown parameters ($\underline{\phi_1}$ and $\underline{\phi_2}$) as

$$\mathcal{L}(\underline{\phi_1},\underline{\phi_2}|\underline{r},\underline{\theta}) = \prod_{i=1}^{n} \frac{\pi(r_i,\theta_i)g(r_i)}{\int\limits_R \int\limits_\theta \pi(r,\theta)g(r)d\theta dr}. \tag{6.11}$$

As for the conventional case,

$$P = \int\limits_0^w g(r)\pi(r)dr \tag{6.12}$$

although now the distribution of r must be obtained through integration of the joint distribution of (r,θ), with respect to θ,

$$\begin{aligned}
\hat{P} &= \int\limits_0^w g(r)\pi(r)dr && (6.13)\\
&= \int\limits_0^w g(r)\int\limits_0^{\frac{\pi}{2}} \pi(r,\theta)d\theta dr && (6.14)\\
&= \int\limits_0^w g(r)\int\limits_0^{\frac{\pi}{2}} \frac{rd(r\sin\theta)}{\int\limits_0^w d(x)\varphi(x)dx}d\theta dr && (6.15)\\
&= \frac{1}{\int\limits_0^w d(x)\varphi(x)dx}\int\limits_0^w\int\limits_0^{\frac{\pi}{2}} r\ g(r)d(r\sin\theta)d\theta dr. && (6.16)
\end{aligned}$$

Given the MLE's for the parameters of interest, one can replace these in the previous expression, leading to the following estimator for P:

$$\hat{P} = \int\limits_0^w \hat{g}(r)\hat{\pi}(r)dr. \tag{6.17}$$

A density estimator is then obtained using standard HTL estimators as described in section 1.2.1 (see equation 1.2). Note, however, that the resulting estimate for D applies only to the covered circles and yet our interest is in density in the strip of width $2w$ centered on the linear feature. I address this issue in the next subsection.

6.2.4 Estimating density in the vicinity of the linear feature

To get a density estimate valid for the vicinity of the linear feature we need to estimate the probability of an animal being in the circle, given that it is in the square of side $2w$ that contains it. That probability is given by

$$P_{c|s} = \frac{\int\limits_0^w d(x)\sqrt{w^2 - x^2}dx}{\int\limits_0^w d(x)wdx}. \tag{6.18}$$

Note that for constant $d(x)$, as in CDS, $P_{c|s} = \pi/4$. We can estimate $P_{c|s}$ using the estimated density gradient. Hence the density estimate in the $2w$ width strip along the linear feature is

$$\hat{D} = \frac{n}{a\hat{P}\hat{P}_{c|s}} \tag{6.19}$$

where a is the area of the k squares containing the covered circles ($k4w^2$). This is a Horvitz-Thompson-like (HTL) estimator, *sensu* Borchers *et al.* (2002, p. 143-144): the probabilities involved are estimated instead of known by design as in conventional Horvitz-Thompson estimators.

This density estimator is valid only in the strip of width $2w$ centered on the linear feature, and in general it will not be representative of density in the wider survey region, because the linear feature vicinity is not a random sample representative of the wider region. If an estimate of abundance in areas away from the linear feature is required the approach described in section 4.4 might be adopted.

6.2.5 Variance estimation

A straightforward approach for obtaining variance estimates for the proposed estimators is a nonparametric bootstrap, resampling at the point level. Note that this assumes that the points are independent sampling units, placed randomly along the linear feature. Details of a general bootstrap procedure readily extendable to this case can be found in section 2.3.2 (see also Buckland *et al.*, 2001, p. 82-84).

6.3 A simulation experiment

In this section a simulation exercise to evaluate the performance of the proposed methods is described.

6.3.1 Simulation settings

Consider a square study area of side 200 m, the center of which is, without loss of generality, at (0,0). There is a linear feature along the y-axis (x=0). Animals have a density gradient in the x direction and are uniformly distributed in the y direction. To test the proposed methods, a design with one single point transect located at the center of the study area is considered. Regarding the detection and density gradient function estimation, this is equivalent to many points along a linear feature, which for analysis purposes are essentially stacked on top of each other (i.e., density at the single point is the sum of densities at all the points).

A constant population size of $N = 1000$ animals was simulated. The density gradient was assumed to be one of four types, determined by two factors (see Figure 6.2a-b): (1) animals either avoid or prefer the linear feature and (2) with either a hazard-rate (HR) or half-normal (HN) based density gradient. Therefore, for each animal, the x coordinate was generated from these distributions, while the y coordinate was generated from a uniform on (-100,100). The detection function was assumed half-normal (Figure 6.2c). Given the detected radial distances, r, a truncation distance of $w = min(100, max(r))$ was used.

The main simulation scenarios considered are shown in table 6.1, along with the true P and $P_{c|s}$ (respectively equations 6.17 and 6.18 evaluated with the true parameter values) associated with them. For each scenario, the following procedure was repeated 100 times: the animals' positions were generated according to the density

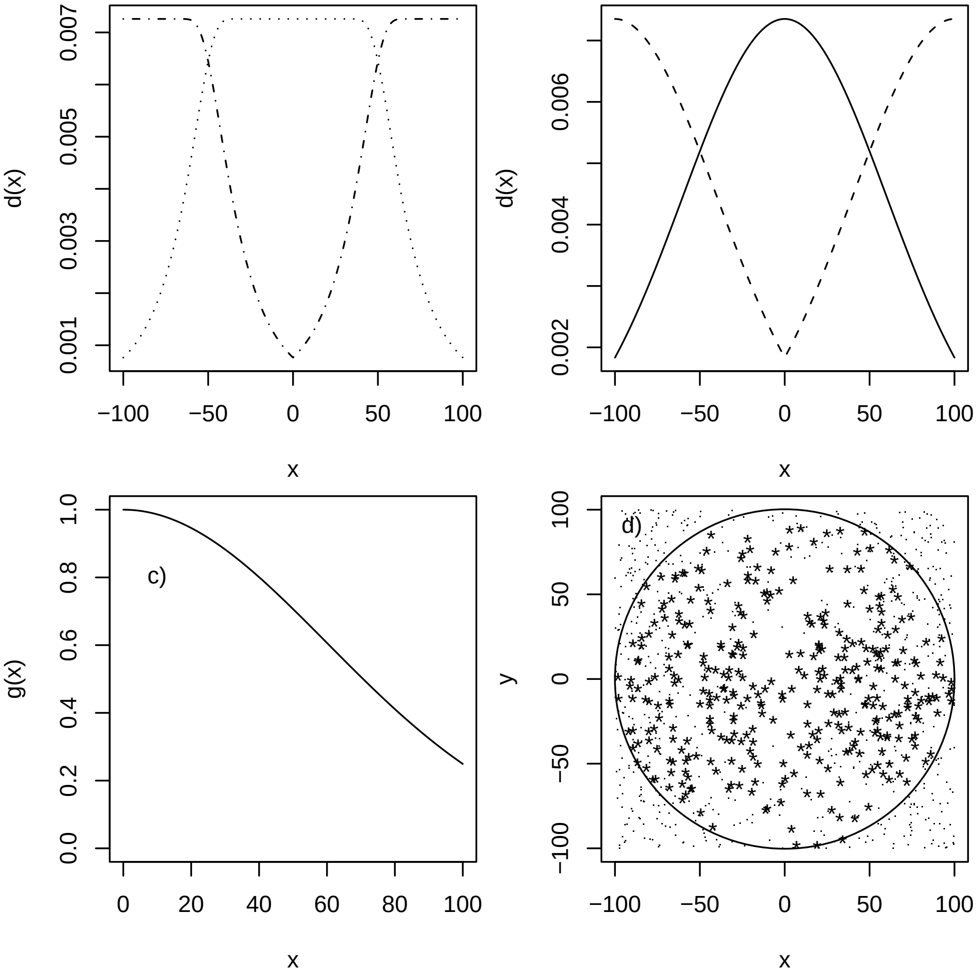

Figure 6.2: Simulation settings details. a) Example of hazard-rate density gradients, used in scenarios 3 and 4; b) Example of half-normal density gradients, used in scenarios 1-2 and 7-8; c) Example of half-normal detection function, used in scenarios 1-6 and 11-12; d) Realization of a population under scenario 2, with the detected animals shown as *. Linear feature is along $x = 0$.

Table 6.1: Simulation scenarios considered, as a function of density gradient $d(x)$ and detection function $g(x)$ parameters. HN stands for half-normal and HR for hazard rate, with the respective model parameters in brackets. $Atr = T$ for animals attracted to linear feature, and $Atr = F$ for animals avoiding it. P is the probability of detecting an animal, given that it is in the circle. $P_{c|s}$ is the probability of an animal being in the circle, given that it is in the square that contains it. $\mathbb{E}(n)$ is the average number of detected animals by simulation. True abundance is N=1000 for all scenarios.

Scenario	$d(x)$	Atr	$g(x)$	P	$P_{c\|s}$	$\mathbb{E}(n)$
1	HN(60)	T	HN(60)	0.576	0.861	496
2	HN(60)	F	HN(60)	0.496	0.716	355
3	HR(60,4.3)	T	HN(60)	0.586	0.891	522
4	HR(60,4.3)	F	HN(60)	0.477	0.702	335
5	HN(30)	T	HN(60)	0.629	0.952	599
6	HN(30)	F	HN(60)	0.409	0.583	238
7	HN(60)	T	HN(30)	0.215	0.861	185
8	HN(60)	F	HN(30)	0.131	0.716	94
9	HN(30)	T	HN(30)	0.279	0.952	266
10	HN(30)	F	HN(30)	0.055	0.583	32
11	HN(900)	T	HN(60)	0.541	0.786	425
12	HN(900)	F	HN(60)	0.540	0.785	424

gradient, the detection process simulated and the relevant distances to detected animals' used to estimate abundance (N), either ignoring the density gradient as in CDS or using the methods outlined above. For illustration purposes, a realization of a population under scenario 2 is shown (see Table 6.1 for the parameters associated with each scenario number), together with the point transect position and the location of detected animals, in figure 6.2d.

Table 6.2: Results of the simulation exercise to evaluate the proposed methods to deal with density gradients. Mean estimated abundance (N), associated standard deviations (s_N), intervals spanning 95% of the simulated values (95%ISV) and mean observed % bias (%B) for each of the main simulation scenarios, considering the proposed and conventional methods. True abundance is N=1000.

	Proposed Method				Conventional Method			
Scenario	N	s_N	95%ISV	%B	N	s_N	95%ISV	%B
1	999	89	981,1016	-0.1	1448	96	1429,1467	44.8
2	1010	141	982,1038	1.1	593	52	583,604	-40.7
3	998	84	981,1015	-0.2	1588	107	1566,1609	58.7
4	1018	161	986,1050	1.8	465	38	457,472	-53.5
5	998	63	986,1010	-0.2	2344	149	2314,2373	134.4
6	987	281	931,1042	-1.3	308	18	304,311	-69.2
7	1016	212	974,1059	1.7	1459	169	1426,1492	45.9
8	1075	457	984,1165	7.5	525	66	512,538	-47.5
9	979	114	956,1002	-2.1	2488	202	2448,2529	148.8
10	1381	1562	1071,1691	38.1	74	20	69,78	-92.7
11	952	96	933,971	-4.8	986	89	968,1003	-1.4
12	1026	86	1009,1043	2.6	995	72	980,1009	-0.5

6.3.2 Simulation results

The performance of conventional and proposed methods under different scenarios can be compared in table 6.2, which shows the estimated population size and corresponding mean observed % bias.

The improvement over conventional methods is substantial for all scenarios tested, showing that the proposed methods are a desirable approach. However, the estimator was still appreciably biased in scenario 10. This results from a combination of the nature of the estimator involved and small sample size. As pointed out by others (see e.g. Borchers, 1996; Marques and Buckland, 2003), HTL estimators are positively biased, and this bias can be considerable if the probabilities in the denominator are small, as in scenario 10 (cf. Table 6.1). Also, given that all the scenarios were identical with respect to true abundance, sample size was proportional to the true probabilities involved, and hence scenario 10 presented an expected sample size (32) that is smaller than what one would recommend for analysis of real data sets.

To evaluate this issue further I considered two additional simulation exercises: (1) 24 additional scenarios were run, in which P took values in between those of scenarios shown in table 6.1 and (2) 30 additional scenarios were run, with the same parameters as Scenario 10, but increasing true abundance from 1200 to 7000 (corresponding $\mathbb{E}(n)$ from 38 to 224). In figure 6.3a the bias in estimated abundance, considering the proposed methods, is shown as a function of true P, for all 36 scenarios considered. It can be seen that values of P of around 15-20% or greater are required for reliable estimates. Note that the bias of the proposed methods is nonetheless considerably smaller than that for conventional methods, for all scenarios run. The effect of sample size on estimation bias is clear in figure 6.3b, where for sample sizes larger than around 150 the bias becomes small.

The density gradient in scenarios 11 and 12 is such that, for all practical purposes, it could be taken to be uniform. It is reassuring to see that the proposed methods still work relatively well under the more conventional setting, even if there is not much information in the data to estimate the density gradient parameter, leading to large variance in estimates of ϕ_2. However, their use would not be recommended in such situations because an extra parameter (vector) needs to be estimated leading to worse precision than for the conventional methods (cf. Table 6.2). Note that the slight bias present in the simulation exercise for these scenarios is a consequence of using the true model for estimation. Depending on the density gradient assumed, when the parameters are estimated, you can only get a non-uniform gradient in one direction, hence you would underestimate density if you assumed an attraction gradient with respect to the linear feature, and vice versa. If the gradient direction

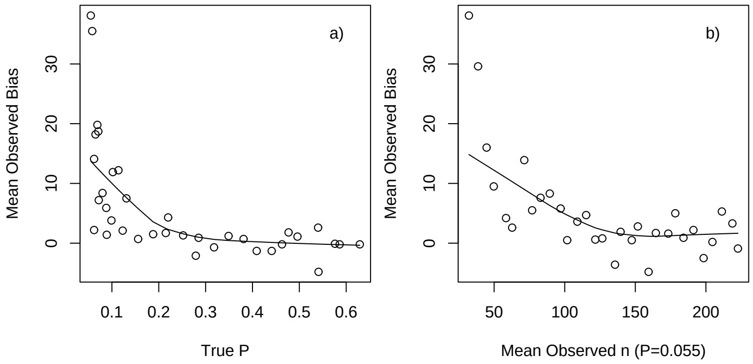

Figure 6.3: Mean observed % bias as a function of a) True P; b) Sample size. In a) data represents the 12 scenarios in table 6.1, plus 24 extra scenarios with values of P spanning values between those. In b) data corresponds to Scenario 10 with N spanning from 1000 to 7000. The line in both panels is a standard lowess smooth.

was also estimated this would not be the case.

6.4 Applying the methods to a hare survey data

In 2005, a survey of hares was carried out in Northern Ireland[3]. A total of 5421 point transects were surveyed, and due to logistic constraints these were placed along roads. A total of 210 hare clusters were detected, corresponding to 314 detected hares. For the sake of simplicity, I assume that the interest is in estimating cluster density.

These animals tend to avoid field boundaries, and this was easily seen from inspection of the distance data (Figure 6.4a), which shows hares clearly avoiding the vicinity of roads. The methods proposed in the previous sections were applied to estimate density within the covered region, and also (using equation 4.7) to estimate density at $w = 150$ m from the road. Estimates were compared with those from CDS.

I considered 4 alternative models: (M1) a conventional analysis, using a HN detection function (a cosine adjustment term was added for fit improvement, based

[3] See Acknowledgements section for proper credit on the use of this data set.

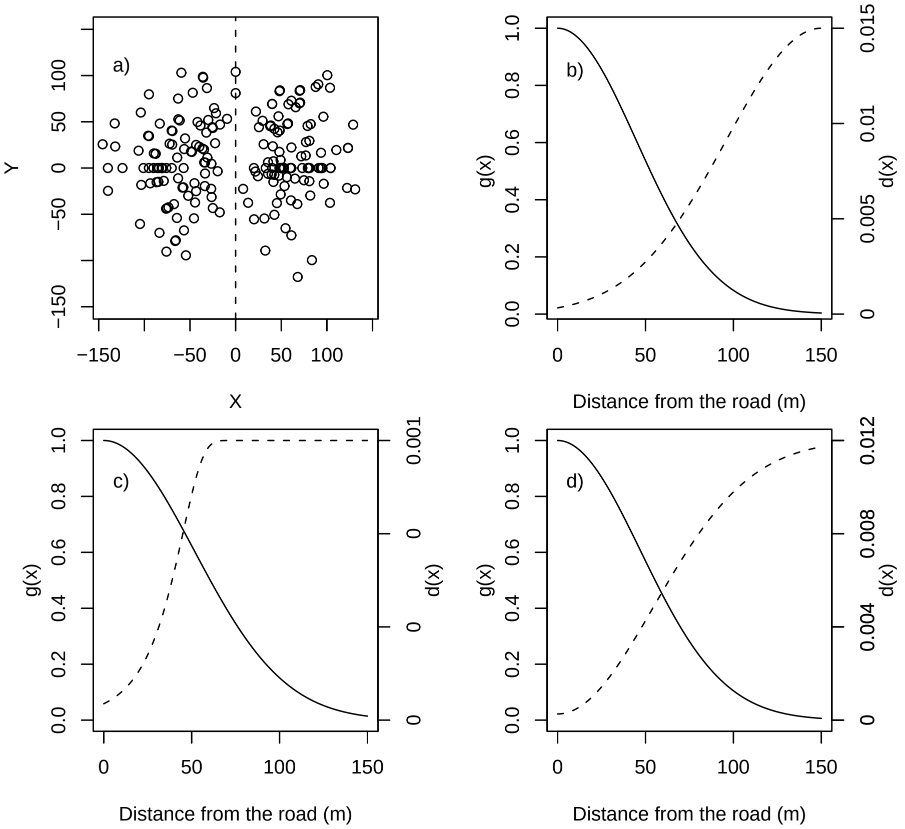

Figure 6.4: Data and estimated models in the hares example. a) The positions of the detected hares with respect to the point, considered to be centered at (0,0). The dashed line represents the road position with respect to the point. Data have been truncated so that $r < 150$ m; b) to d) The estimated relative density gradient (dashed line) and detection function (solid line), considering: b) a half-normal; c) an hazard-rate; and d) a normal based density gradient.

on minimum AIC), implemented in Distance 5 (Thomas *et al.*, 2005); for the proposed methods I used a HN detection function with a density gradient where animals avoid the road, with avoidance modelled using either (M2) a HN or (M3) a HR, as in the simulation section. I also considered a model for the density gradient (M4) as

$$d(x) = \frac{1 - \beta \times (f(x, \sigma))}{\int_0^w 1 - \beta \times (f(x, \sigma))dx} \tag{6.20}$$

where $f(x, \sigma)$ represents the normal density, with mean 0 and standard deviation σ. Hence model M4 is also based on the normal distribution as M2, but compared with the previous two, it allows easier parameter interpretation, with α representing density at distances where the linear feature influence has disappeared (recall $\alpha d(x) = D(x)$). This model also allows the relation of the animals towards the linear feature to be estimated: for attraction, $\beta < 0$, while for avoidance $\beta > 0$, while $\beta = 0$ simplifies to a uniform.

The results obtained for these models, after numerical maximization of the appropriate likelihoods, are shown in table 6.3. Variances were obtained using the standard empirical variance estimator for the conventional analysis and a bootstrap resampling procedure, considering points as the resampling units, for the proposed methods.

These results show that the conventional method underestimates abundance considerably, with any of the non uniform density gradient models leading to over 100% larger density estimates than the conventional methods, for the strip of width $2w = 300$ m centred on the road. Furthermore, any of these models shows that density estimates ignoring the density gradient will result in severe underestimation of density for the wider region. The considerable difference in point estimates obtained depending on the model used for the density gradient was somewhat disappointing. The fit of model M4 is preferred by AIC, and $\Delta AIC = 3.9$ for model M3, the second best model, which shows a strong preference for M4; for comparison, $\Delta AIC > 80$

Table 6.3: Results of the analysis of the hare data, considering 4 alternative models (M): (M1) CDS ignoring density gradient; proposed method with animals avoiding the road according to a (M2) HN gradient, (M3) HR gradient and (M4) a normal based gradient. AIC: Akaike Information Criterion, with corresponding AIC weights shown inside brackets; $d(x)$: estimated parameters for the density gradient; $g(x)$: estimated parameters for the detection function; $\hat{P}$: estimated probability of detecting a hare cluster, given it is in the circle; $\hat{P_{c|s}}$: estimated probability of a cluster being in the circle given it is on the square containing it; $\hat{D}$: estimated cluster density in the strip of width $2w = 300$ m centred on the road; $\hat{D}_{150}$: estimated cluster density at 150 m from the road. Corresponding 95% confidence intervals for density estimates are shown inside brackets.

M	ΔAIC	$d(x)$	$g(x)$	$\hat{P}$	$\hat{P_{c\|s}}$	$\hat{D}$	$\hat{D}_{150}$
M1	-	-	66.45,-0.11	0.42	-	1.24	1.24
						(0.71,2.15)	(0.71,2.15)
M2	3.83	54.2	44.8	0.075	0.626	8.673	19.249
	(0.126)					(4.99,14.41)	(10.01,35.02)
M3	7.87	106.3,8.1	51.4	0.166	0.723	3.382	4.437
	(0.017)					(2.19,7.46)	(2.68,12.60)
M4	0	134.2,54.7	47.1	0.106	0.671	5.727	10.074
	(0.857)					(2.95,13.54)	(3.98,40.38)

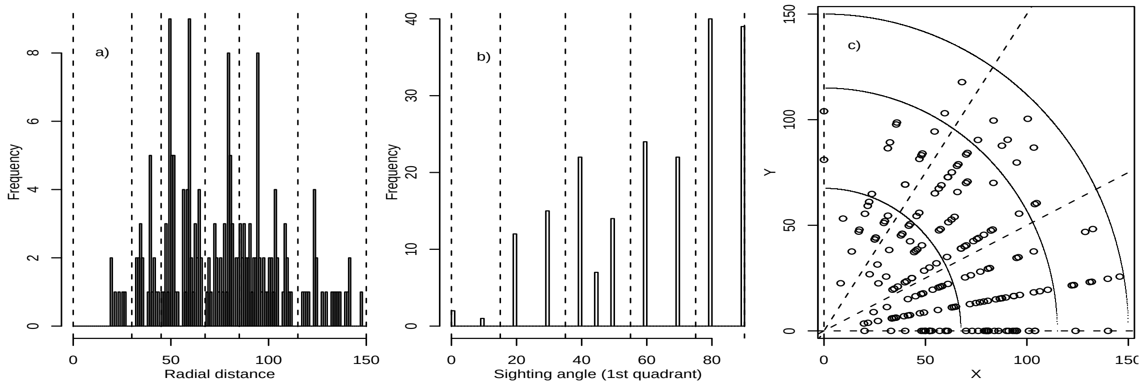

Figure 6.5: Observed data used for the chi-squared goodness-of-fit (GOF) tests for the hares data. a) Radial distances; b) Sighting angles; c) Two dimensional GOF, considering both radial distances and sighting angles. Dashed lines and solid quarter circles represent the cut points used in the GOF tests.

for models considering the conventional uniform density gradient. Model M3 is not supported by the data (cf. Table 6.3). M4 is arguably this is the most interesting of the 3 models considered, suggesting that it is likely that the effect of the road is almost absent for distances close to $w = 150$ m. Unfortunately, this model's 95% confidence limit is very large, reflecting poor precision.

To try to further discriminate between models, absolute goodness-of-fit (GOF) was assessed using a standard chi-square test. Due to considerable heaping of radial distances and especially of sighting angles, interval cut points were chosen by visual inspection of the data. The GOF tests were carried over the radial distances, the sighting angles and both dimensions. The data and the corresponding cut points used are shown in figure 6.5. The observed and expected counts for each model, as well as other relevant details for the chi-square tests, are shown in tables 6.4-6.6.

The tests based on radial distances or both radial distances and sighting angles suggested that the models considered were adequate. The poor fit indicated for tests on the sighting angles alone is most likely due to measurement error in angles. The severe heaping is evident in figure 6.5, with angles restricted to multiples of 10°

Table 6.4: Results of the Chi-square goodness-of-fit test with respect to the radial distances. Obs: Observed counts. Exp: Expected counts, according to respective model. Chi: Chi-square parcel corresponding to a given interval. Density gradients are half-normal (M2), hazard-rate (M3) and normal (M4) based. Summary values are the test statistic (χ^2), corresponding degrees of freedom (*df*) and associated P-value.

Distance		M2		M3		M4	
Cut points	Obs	Exp	Chi	Exp	Chi	Exp	Chi
(0,30)	6	11.21	2.42	8.01	0.51	8.04	0.52
(30,45)	20	19.07	0.04	17.39	0.39	18.38	0.14
(45,67.5)	59	47.20	2.95	52.68	0.76	50.30	1.50
(67.5,85)	37	42.98	0.83	45.30	1.52	45.18	1.48
(85,115)	54	55.33	0.03	52.41	0.05	54.75	0.01
(115,150)	22	22.20	0.00	22.22	0.00	21.35	0.02
χ^2			6.29		3.23		3.67
df			3		2		2
P-value			0.099		0.198		0.159

Table 6.5: Results of the Chi-square goodness-of-fit test with respect to the sighting angles. Obs: Observed counts. Exp: Expected counts, according to respective model. Chi: Chi-square parcel corresponding to a given interval. Density gradients are half-normal (M2), hazard-rate (M3) and normal (M4) based. Summary values are the test statistic (χ^2), corresponding degrees of freedom (*df*) and associated P-value.

Angle		M2		M3		M4	
Cut points	Obs	Exp	Chi	Exp	Chi	Exp	Chi
(0,15)	3	7.12	2.38	6.50	1.88	4.82	0.69
(15,35)	27	21.94	1.17	27.78	0.02	23.25	0.60
(35,55)	43	44.39	0.04	51.59	1.43	48.23	0.57
(55,75)	46	66.51	6.33	62.59	4.40	66.33	6.23
(75,90)	79	58.04	7.57	49.54	17.51	55.36	10.09
χ^2			17.49		25.25		18.19
df			2		1		1
P-value			$< 10^{-3}$		$< 10^{-6}$		$< 10^{-4}$

Table 6.6: Results of the Chi-square goodness-of-fit test with respect to the radial distances and the sighting angles. Obs: Observed counts. Exp: Expected counts, according to respective model. Chi: Chi-square parcel corresponding to a given interval. Density gradients are half-normal (M2), hazard-rate (M3) and normal (M4) based. Summary values are the test statistic (χ^2), corresponding degrees of freedom (*df*) and associated P-value.

Distance	Angle		M2		M3		M4	
Cut points	Cut points	Obs	Exp	Chi	Exp	Chi	Exp	Chi
(0,67.5)		15	14.52	0.02	11.68	0.95	10.91	1.54
(67.5,115)	(0,35)	13	12.04	0.08	16.97	0.93	13.66	0.03
(115,150)		2	2.50	0.10	5.63	2.34	3.50	0.64
(0,67.5)		25	28.85	0.51	30.13	0.87	29.88	0.80
(67.5,115)	(35,65)	33	37.56	0.55	42.91	2.29	40.90	1.52
(115,150)		9	8.98	0.00	9.05	0.00	9.01	0.00
(0,67.5)		45	34.12	3.47	36.27	2.10	35.93	2.29
(67.5,115)	(65,90)	45	48.72	0.28	37.82	1.36	45.37	0.00
(115,150)		11	10.71	0.01	7.54	1.59	8.84	0.53
χ^2				5.02		12.42		7.35
df				6		5		5
P-value				0.541		0.029		0.196

with the unique exception of 45°, and some evidence for angles around 65°-75° being recorded as 80°-90°. Given these results, the most reliable test is likely to be the one based on radial distances only, for which the GOF test statistic indicates that model M4 might be adequate, further reinforcing the results based on AIC.

6.5 Discussion

The methods proposed in this chapter are needed when point transects are placed along a linear feature, such as a road or river, with respect to which the animals of interest might exhibit a density gradient. This commonly occurs in published studies (e.g. Ruette *et al.*, 2003), but most often the possible impacts are not assessed directly.

I focused here on the estimation of abundance in the vicinity of the linear feature, but usually one is interested in making inferences over much wider areas. CDS can be seen as a two-stage process: (1) estimating the probability of detecting an animal given it is in the covered area (model-based) and (2) scaling it up for the wider survey region (usually design-based). Given that under the proposed methods we have a model $D(x)$ describing density as a function of distance from the linear feature, the second stage involves model-based inference beyond $x = w$, although estimation for uncovered areas along the road, within w of the road, is still design-based.

The methods assume that the detection function is independent of θ. If that is not the case, the density gradient and detection function are not separately estimable using the proposed methods. Note that this also poses problems in situations in which say sectors of the point transects closer to the road are more or less visible than sectors away from the road, i.e., in which the probability of given sector being visible is a function of angle (see next chapter for an example and a possible alternative to deal with such a problem).

6.5.1 Simulation exercise

The results presented show that when analyzing distance sampling data with conventional methods, ignoring an existing density gradient can lead to substantial bias, casting doubts over the adequacy of such a procedure.

Based on the simulations, it is apparent that the methods work better in some cases than others. They should work best in situations in which the true underlying P is above around 0.15-0.20, due to inherent positive bias of HTL estimators for small P. (More specifically, the HTL bias is a function of the coefficient of variation in the P estimate; for small values of P the relative precision is usually poor, justifying this advice.) Hence, these should work better for animals that tend to have higher density near the linear structure rather than away from it. If sample size is large, as in the hare survey example, the bias should also be small.

6.5.2 Applying the methods to the hare data

For the hare survey example, despite previous knowledge about the species and the data showing clearly that the conventional methods should not be used, the GOF statistics associated with the conventional analysis were non-significant (at the usual 5% significance level). However, a simple Kolmogorov-Smirnov test of the uniformity of the sighting angles distribution (expected in the absence of a density gradient) would allow identification of the non-uniform distribution of animals with respect to the road (in this case P$< 10^{-4}$), further justifying the suggestion of Buckland *et al.* (2001, p. 275) that sighting angles might have relevant information in point transect surveys. (Note these might also allow identification of responsive movement in the absence of a density gradient.) I therefore recommend that as a precursor to applying CDS approaches, such an inspection of sighting angles is undertaken.

The hare survey data demonstrate the potential difficulties in applying the methods proposed here to real data; alternative models of the density gradient lead to considerable variation in the results. Gathering independent information on $d(x)$ (see Chapter 7) will help to distinguish candidate models, allowing the use of more reliable models for the density gradient and making more reliable inference.

Because detection probabilities seem close to 0 at 150 m, an analysis was implemented for different truncation values, from 99 to 147 m, in 3 m intervals (analysis not shown). As more severe truncation is used, the more similar the estimates from the different models become. Therefore, and although the distinction amongst the contending models does not become easier, such distinction becomes less important. Reducing the truncation distance brings nonetheless other problems, as there is a loss in precision and it becomes less likely that the density at w can be assumed representative of the wider survey region. Hence the choice of an appropriate truncation distance for the proposed methods might be a more influential decision than for conventional methods, for which the choice of truncation distance is usually straightforward.

Given the large differences in estimates considering different models, an approach that incorporates model uncertainty in the estimates might be used, like a bootstrap procedure, in which the choice amongst the competing models is done for each bootstrap resample (e.g. Buckland *et al.*, 1997). This was not implemented here because the focus was on the new methods to deal with non random samplers.

6.5.3 Implications for line transects

The proposed methods consider points laid along lines, but the same problems apply if the linear feature itself is used as a line transect, as described in chapter 5. Hiby and Krishna (2001) showed that, provided animal density is unrelated to paths, distance sampling might provide unbiased estimates along curved paths, under mild assumptions on the path's curvature. However, if density is markedly different in the vicinity of the linear feature, then substantial bias in abundance estimates can be anticipated if line transects are placed along linear features.

The proposed methods are not easily extended to line transects[4] where the sighting angles convey information both on the density gradient and on the search process, and such information can not easily be disentangled. Hence, methods like those of chapter 5 must be used instead.

[4] But see nonetheless the last sentence of chapter 5

Chapter 7

Extending the methods to more complex scenarios

7.1 Introduction

In chapter 6 a hare (*Lepus timidus hibernicus*) data set collected in 2005 in Northern Ireland was used for illustration of the methods. Following on the output of that work, a more extensive survey of the same species, this time in the Republic of Ireland, was carried in 2006, with the objective of providing baseline information on the population status of this species.

In this chapter I analyze the data resulting from that survey. For the purposes of this thesis, the main interest is not on the results *per se*, but to illustrate how the previous methods can be easily extended to accommodate certain particularities of a given data set. The advantages of using methods based on maximum likelihood are again recognized and emphasized: they readily provide a framework for easy generalization of methods and the combination of multiple sources of data.

In the next section the data set is briefly described. This is then followed by the analysis of the data, adding increasing complexity. The chapter finishes with a brief discussion.

7.2 Data description

The motivating data was collected as part of a large survey effort to estimate the population size of hares in the Republic of Ireland, in the spring of 2006, building on the experience obtained through the analysis of the hare data from Northern Ireland in 2004-2005 (part of this analysis is shown in chapter 6).

The data consisted of three separate data (sub-)sets, as described below.

1. The main data set. This comprised 2440 point transects, located on 533 randomly selected squares across Ireland, in which a total of 87 hare clusters were detected. Due to insurmountable logistic difficulties associated with truly random location of points trough the landscape, these points were placed along roads, which motivated the use of the methods of chapter 6, presented in section 7.3.2.

2. A set of repeated points. A random sample of 481 of the 2440 points from the main data set was repeated by another team of observers. For these points, the sections of the point for which visibility was not obstructed by any landscape feature (walls, dense high vegetation, etc.) were recorded. This was done by recording for each of eighteen 20° sectors available in a circle (0°-20°, 20°-40°,...,340°-360°, see Figure 7.3b) whether there was an obstruction in that direction or not. This allowed the evaluation of the effort distribution for each point. Note that because the methods of chapter 6 use the distribution of sighting angles to disentangle the detection function and the density gradient, one of the assumptions of such methods is that the sighting effort is independent of angle – this assumption can be relaxed with this additional data (see section 7.3.3).

3. A secondary transects data set. Based on the results obtained in chapter 6,

there was the notion that obtaining a direct sample from $d(x)$ could be helpful. Hence a total of 93 secondary transects, of variable length, perpendicular to the road, were surveyed, obtaining a direct sample of the density gradient. Unfortunately sample size was very small, with only 9 hare clusters detected (see section 7.3.4).

In section 7.3.5, I consider an analysis that combines the information of the 3 data (sub-)sets in an attempt to provide better density estimates.

7.3 Building complexity in the analysis of the data

In the following sections, the 2006 hare data set is analyzed with increasing degrees of complexity, motivated by considerations about the data and the best way to use the information contained on it. For simplicity I deal exclusively with estimation of the hares density in the vicinity of the road.

7.3.1 Conventional point transect analysis

A naïve analysis of the main data set alone, without the proper knowledge about the survey design being points along roads, could motivate a conventional analysis of such data using software Distance. This analysis is presented because it illustrates well the danger of using software without a profound knowledge of the methods it implements and the data used.

The model chosen for the detection function by minimum AIC was the half-normal (with no adjustment terms, see figure 7.1), and the estimated density was 2.07 hares/km^2. The most interesting feature of this analysis is that, based on it alone, there would be no indication that the results were not sensible, with a good fit of the chosen model to the data (P>0.35 for the standard chi-square, Kolmogorov-Smirnov and Cramér-von Mises GOF tests).

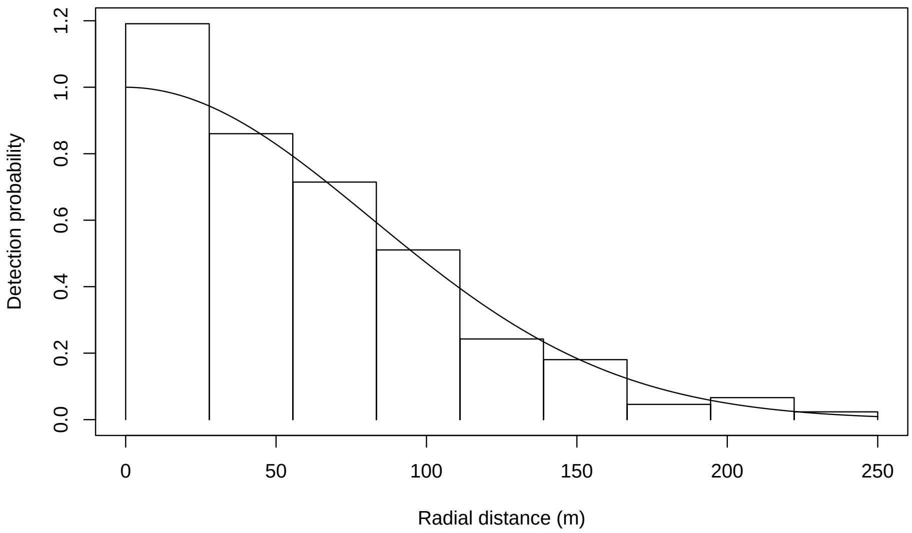

Figure 7.1: Republic of Ireland 2006 hare data and estimated detection function considering conventional distance sampling methods. Note that the bars are rescaled by dividing the actual count by the interval mid point for visual purposes.

The mean observed cluster size was 1.35, while the size bias regression based estimate of mean cluster size was 1.25, suggesting a (not surprising) tendency for larger clusters to be easier to detect. Being a side issue here, the latter value was used as an estimate of mean cluster size for the subsequent estimates.

7.3.2 Using the methods of chapter 6

Given the knowledge that these animals distribution is clearly influenced by roads, and given the survey points were placed along roads, the methods of chapter 6 should be used.

As possible candidates for the density gradient I used the same models as in section 6.4, namely M2, M3 and M4. The estimated density gradient and detection functions are presented in figure 7.2. Unfortunately there is not sufficient information in the data to distinguish clearly between these 3 models, with the ΔAIC being only 2.84 between the lowest (HN density gradient) and the highest (HR density

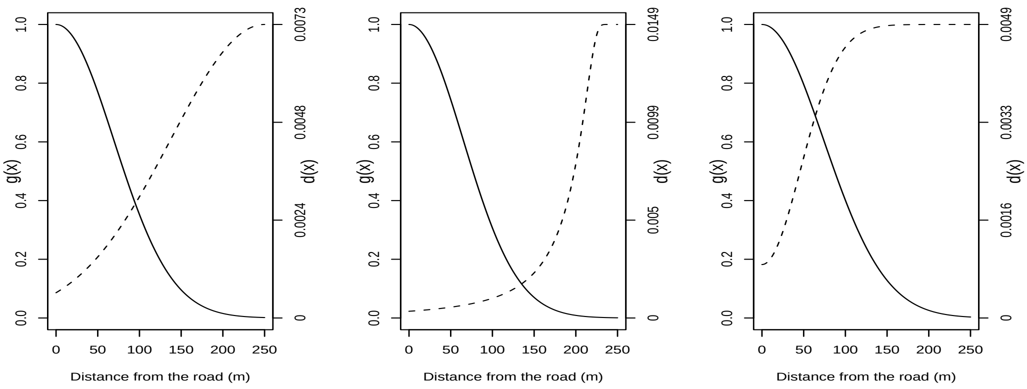

Figure 7.2: Estimated detection function (solid line) and density gradient (dashed line) for the Republic of Ireland 2006 hare data set, considering the methods proposed in chapter 6, according to, from left to right, models M2 (half normal density gradient), M3 (hazard rate density gradient) and M4 (normal based density gradient).

gradient) AIC. This is especially worrying because the 3 models considered lead to very different density estimates. Although all three models indicate consistently that CDS would underestimate hare density, the corresponding point estimates for hare density are considerably different, resulting in estimated densities of 6.35 (M2), 18.91 (M3) and 3.63 (M4) hares/km^2 for the models considered above.

7.3.3 Unequal effort with respect to angles

In conventional distance point transect sampling, if a given sector of a point is not available for detections to be made, the analysis is still straightforward. The area not available for detection is simply subtracted from the total area. For practical purposes this is done by recording the point's effort as the proportion of the point which was available for detection (this was the approach used above in section 7.3.1). The reason one is allowed such a simple solution to the problem is the fact that the sighting angle does not contain relevant information about the processes being modeled (the detection function only).

However, for the methods proposed in chapter 6, the situation is considerably more complicated, because the sighting angle contains relevant information about one of the processes modeled, namely the density gradient. An example illustrates the problem: if the sectors closer to the road (i.e., close to 0 and 180 degrees, if 0 is straight along the road) were available for detections in only half the points, we would interpret the observed data as indication that the animals tended to avoid the road, even if the density gradient was uniform, because there would be considerably less detections near the road than one would expect under such a density gradient.

A closer look at the main survey data revealed that there were many points for which the recorded effort was not 1 (Figure 7.3a), i.e., for which some sectors were not visible (due to walls, fences, dense vegetation, etc) and hence for which no hares could be detected. Provided that the missing sectors were random with respect to sighting angles, the methods should work well. However, based on the indication from the data for the repeated squares, for which there was information not only on the effort for each point but also for which sectors were visible (Figure 7.3b), we see that sectors closer to the road were much more likely to be obstructed by visibility barriers. This meant that the analysis of such data ignoring the unequal availability of angles would not be adequate.

The original idea underlying the proposed approach is due to David Borchers. We can condition the inferences on the angles known to be visible from each point (those at which animal sightings were made). To construct the distribution of sectors available, given sectors visible, I use the repeated points data, for which the sectors available to be detected were recorded. Hence an assumption of this approach is that the effort distribution by sector, for the repeated squares, is representative of the main survey effort.

Consider the following notation:

- $v(\theta)$ is an indicator variable that takes the value 1 if angle θ is visible from the

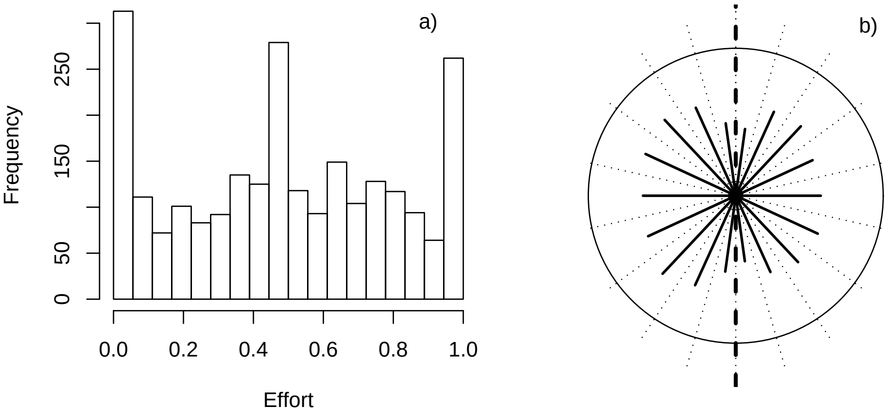

Figure 7.3: Republic of Ireland 2006 hare data set survey effort. a) Number of points with each level of effort for the main survey data set (effort 1 corresponds to the entire point being visible, 0 to a point without any sector visible); b) Solid line length in each sector represents the proportion of points, for the repeated points, for which each of the eighteen 20° sectors were visible. The thicker dotted line represents the road and the thinner dotted lines represent the sector's boundaries.

point and 0 otherwise.

- Θ is the set of angles visible from a point and Θ_{vis} is the set of angles known to be visible from the point, i.e. $\upsilon(\theta)$=1 if $\theta \in \Theta_{vis}$. For survey points angles exist that, although visible, are not contained in Θ_{vis}.

- $p(\upsilon(\theta))$ is the probability of angle θ being visible from the point

- $p(\upsilon(\theta)|\Theta_{vis})$ is the probability that angle θ is seen from the point, conditional on the fact that the angles Θ_{vis} are visible. This is the fundamental quantity that will contribute with new information for the likelihood.

- $g(r, \theta)$ is the detection function, and as before I assume angular symmetry, i.e., $g(r, \theta) \equiv g(r)$.

- $\pi(r, \theta)$ is the *pdf* of animal locations in a point, and $\pi(r, \theta|\Theta_{vis})$ is the corresponding *pdf*, conditional on the angles known to be visible Θ_{vis}.

Assuming that the repeated squares are a representative sample of the main survey effort, that data can be used to estimate both $p(\upsilon(\theta))$ and $p(\upsilon(\theta)|\Theta_{vis})$. Because only points for which at least 1 animal was detected contribute for the likelihood, the main survey data has the information needed about Θ_{vis} for all the relevant points, while for the repeated squares data Θ_{vis} is known for all points. In the case of the main survey, for point k in which n_k detection were made, Θ_{vis} is known for up to n_k angles (if all animals are detected at different angles).

The idea is to write down a *pdf*, which extends equation 6.9, now conditional on the angles known to be visible at each point. This leads to

$$f(r,\theta|\Theta_{vis}) = \frac{p(\upsilon(\theta)|\Theta_{vis})\pi(r,\theta|\Theta_{vis})g(r)}{\int\limits_{R}\int\limits_{\theta} p(\upsilon(\theta)|\Theta_{vis})\pi(r,\theta|\Theta_{vis})g(r)d\theta dr}. \tag{7.1}$$

It is assumed that $\pi(r,\theta|\Theta_{vis})$ is independent of Θ_{vis}, which corresponds to assuming that the hare density in a sector is independent of a sector being visible or not. The extent to which such an assumption might hold is unknown.

Note that while the above formulation assumes $\theta \in (0^{\circ}, 360^{\circ})$, in the hares case we have Θ_{vis} for each point as the set of 20° sectors that were visible, rather than exact angles. For that reason, one needs to assume that if a given angle is known to be seen because an animal was detected at it, the entire 20° sector that contains it was visible (e.g., if a hare was detected in point k at 5° then $\Theta_{vis} = 0^{\circ} - 20^{\circ}$). A further complication related to implementation is due to the fact that methods of chapter 6 assume folding over the two symmetry axis, and after folding the 18 original sectors map onto 5 sectors only.

Using a likelihood based on the conditional *pdf* presented in equation 7.1, I estimated the detection and density gradient, now accounting for the unequal effort by angle. The resulting models are presented in figure 7.4, along with the initial results

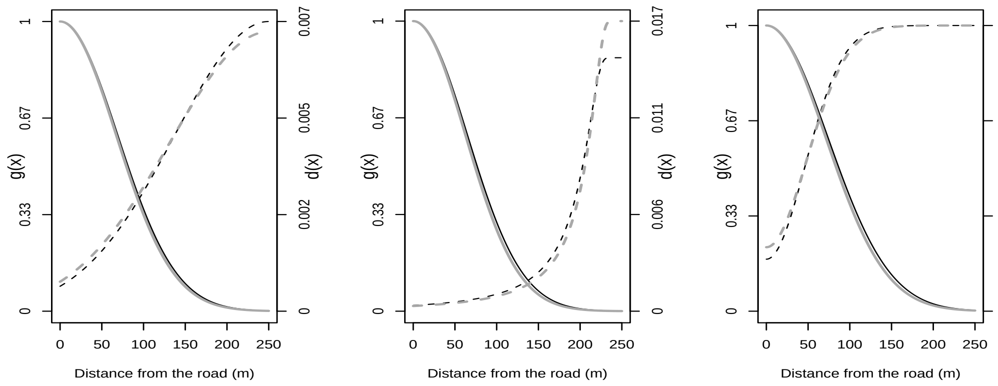

Figure 7.4: Estimated detection function (solid line) and density gradient (dashed line) for the Republic of Ireland 2006 hare data set, considering the methods proposed in chapter 6, according to, from left to right, models M2 (half normal density gradient), M3 (hazard rate density gradient) and M4 (normal based density gradient), accounting for unequal effort by sector (in gray). For comparison the results ignoring accounting for the unequal effort effect are shown in black.

ignoring the effect of the effort by sector for comparison. As expected, for models M2 and M4 the effect of accounting for sectors near the road being less visible is that the density gradient presents relatively more animals near the roads than when ignoring such effect, but the changes are minor. The pattern for M3 is counterintuitive, in the sense that more animals are estimated to be at large distances; this could be due to the inability of the model to adequately describe this data.

In terms of density estimates, that corresponds to a very slight increase in abundances when compared to those obtained in section 7.3.2. The estimates are now respectively 6.37, 23.97 and 3.90 hares/km^2, and based on AIC M2 is the best model, although ΔAIC for models M3 and M4 is less than 2.

7.3.4 Estimating the density gradient using secondary transects

The data on secondary transects is potentially very useful, because it provides direct information about the density gradient, rather than relying on more complicated methods to estimate it. Hence, we can use the distances from the road, for hares seen along the secondary transects, and a likelihood as in equation 5.7, to estimate the parameters of the model assumed for the density gradient. This assumes constant visibility as you go away from the road

However, there is an additional issue regarding the use of this data set to estimate the density gradient. A closer look at the length of the secondary transects surveyed (Figure 7.5) reveals that care needs to be taken in the interpretation of the data, which is not a random sample from the density gradient: even if the transects were a random sample (with respect to the along-road dimension), they are clearly not random in the dimension perpendicular to the road. Larger distances are less represented than smaller ones because more effort was allocated near the roads than far from the roads (note all the transects started at the road).

Assume that T secondary lines, perpendicular to the linear feature, were surveyed, and for line t ($t = 1, 2, ..., T$), n_t animals were seen ($\sum\limits_{t=1}^{T} n_t = n$), at distances x_{tj} from the road ($j = 1, 2, ..., n_t$), such that all the distances seen in transect t are represented by $\underline{x}_t$ and all distances are represented by $\underline{x} = (\underline{x}_1, \underline{x}_2, ..., \underline{x}_j)$. Define the start and end points of line t to be respectively at x_t^s and x_t^e. We could use one of the following two approaches to estimate the density gradient parameters based on this data, which are just extensions to equation 5.7 to accommodate for unequal line lengths.

The first approach is to define an attenuation function as the proportion of lines that cover a given distance x from the road, here represented by $at(x)$, which can be estimated from the data (the start and end points of each line). Then the likelihood

$$\mathcal{L}(\phi_2|\underline{x}) = \prod_{i=1}^{n} \frac{d(x_i)at(x_i)}{\int\limits_{min(x_t^s)}^{max(x_t^e)} d(x)at(x)dx} \tag{7.2}$$

will allow to estimate the parameters of interest accounting for the unequal line

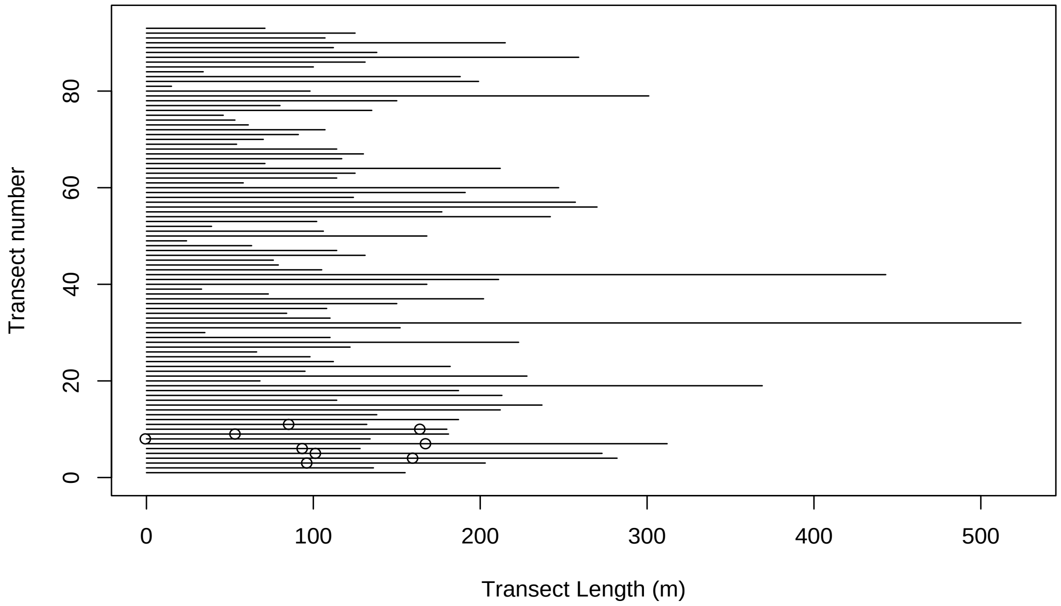

Figure 7.5: Length of each of the 93 secondary transects available for the Republic of Ireland 2006 hare data set. The circles represent distance from the road to the hare clusters detected, along the corresponding transect. Note that transect number was assigned starting with transects with animal sightings, hence the otherwise peculiar pattern, with all the hares detected in the first few transects.

lengths.

A second alternative approach is to consider that the observed data corresponds to censored samples of a common density gradient. Hence we can build a likelihood as the product of the likelihoods associated with each transect (which represent different censored levels) as

$$\mathcal{L}(\phi_2|\underline{x}) = \prod_{t=1}^{T}\prod_{j=1}^{n_t}\frac{d(x_{tj})}{\int\limits_{x_t^s}^{x_t^e} d(x)dx}. \tag{7.3}$$

Despite sharing common points, the two approaches could lead to different results. While for the first approach we assume that all lines could potentially contribute to the likelihood with distances between the global minimum and maximum distance, on the second approach we condition on the actual range of distances covered by each line. Hence, if the density gradient is always the same but animal density may differ by line, the second approach might be preferred over the first one, because the inferences are conditional on the range of distances covered by each line.

Figure 7.6a shows the histogram of secondary line lengths and the corresponding $at(x)$ for the hare data. A half-normal fit to $at(x)$ is also shown (Figure 7.6b).

The density gradient was estimated for the hares data set, using only the data from the secondary transects (i.e., ignoring the point transect data). A normal model for the density gradient was considered, as there was not much evidence supporting the increasing (with distance from the road) density gradient as shown in figures 7.2 and 7.4. The resulting estimated density gradient is shown in figure 7.7, for the two approaches described above. Note that for the first approach, it is shown the result of using both the empirical approximation to $at(x)$, based on the data, and a half-normal approximation to it. The estimated density gradient, ignoring the unequal line length, is also shown for comparison.

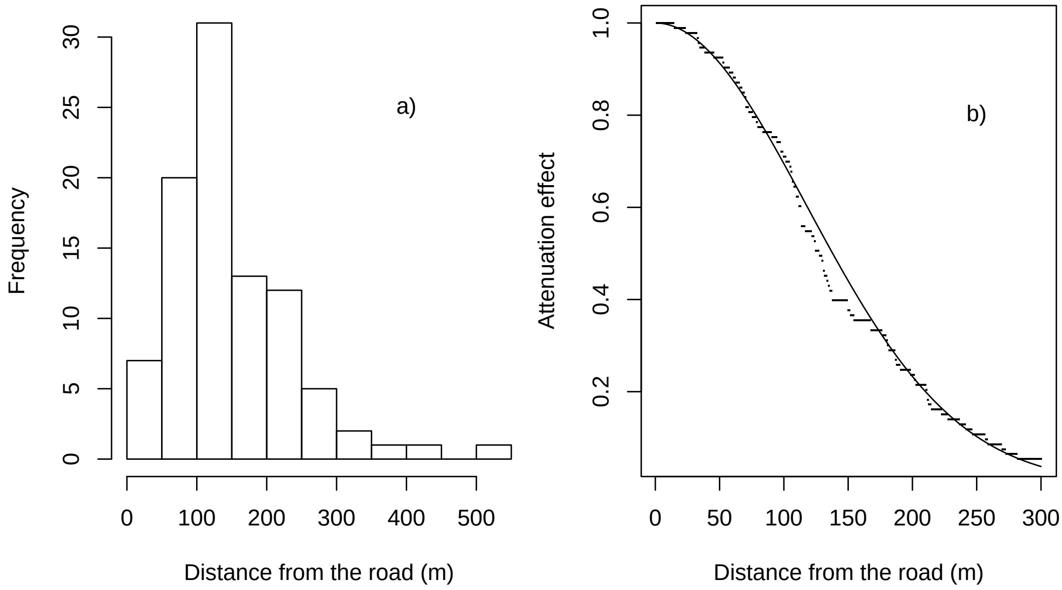

Figure 7.6: Details about the secondary transects effort allocation. a) Histogram of line lengths for the secondary transects; b) The effort attenuation function due to the different line lengths, $at(x)$. The solid line represents a half-normal fit to the data, which can be seen as a continuous approximation to the underlying discrete data.

As expected, because there is considerably less survey effort away from the road, including the unequal effort by distance effect in the likelihood leads to a density gradient with proportionally more animals away from the road, irrespective of the specific method used to estimate the gradient.

7.3.5 Combining the information from points and lines

The density gradient models used for the points likelihood are constrained to have their maximum at w (=250 m). Motivated by the clear differences in the estimated density gradient when considering the information on points or lines separately (cf. figures 7.2 and 7.4 with figure 7.7), a fourth model was used for the density gradient, which would allow a mode for the density gradient. The normal model was used. However, consistent with the previous 3 models considered, the estimated density

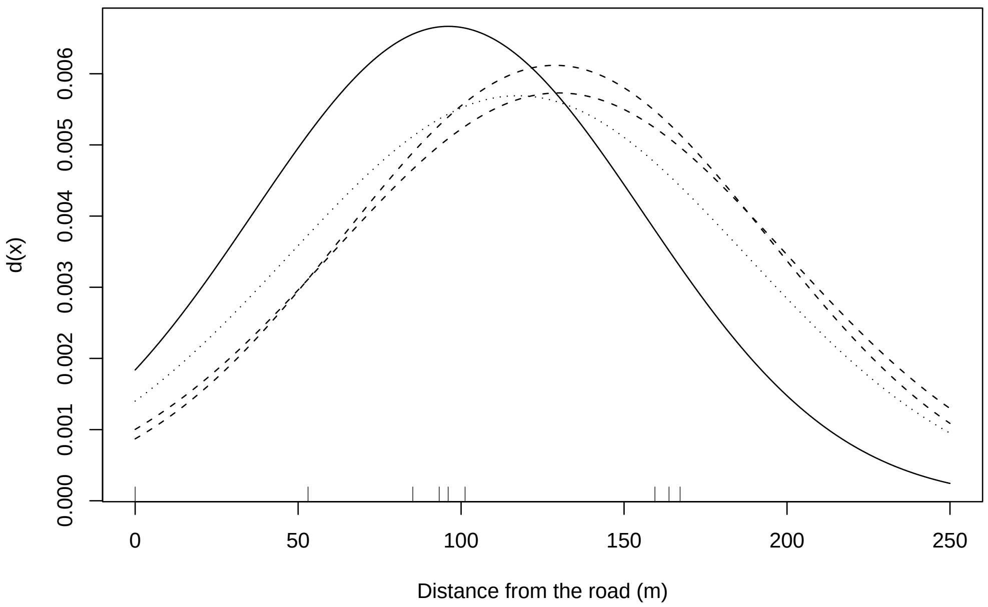

Figure 7.7: Estimated density gradient based on the secondary data alone for the Republic of Ireland 2006 hare data set, considering a normal model for the density gradient. The dashed lines results from using the first approach (based on the attenuation function, considering both an empirical approximation to $at(x)$ and a half-normal approximation). The dotted line results from using the second approach (conditional on the actual line lengths). For comparison the estimated density gradient ignoring the unequal line lengths is also shown (solid line).

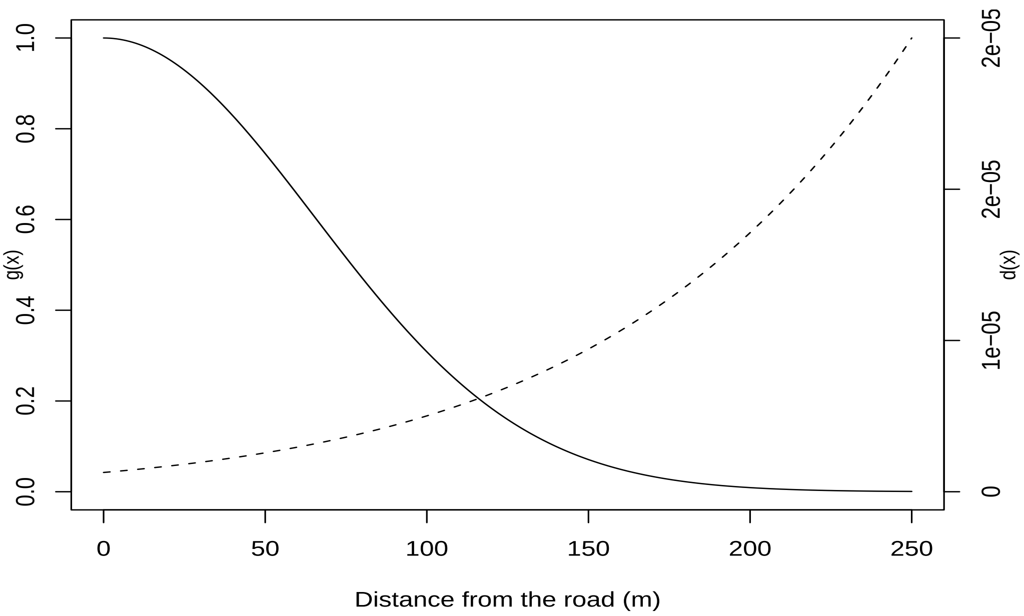

Figure 7.8: Estimated detection function (solid line) and density gradient (dashed line) for the Republic of Ireland 2006 hare data set, using the methods proposed in chapter 6, considering a normal density gradient.

gradient did not present a mode (Figure 7.8).

Note that on the basis of *AIC* alone this model is marginally worse than M2, with ΔAIC=1.1, and a corresponding density of 10.31 hares/km^2. The likelihood surface is flat around the maximum in the density gradient parameter dimensions (results not shown). Note however that density estimates are very insensitive to moderate changes in parameter estimates because for this particular case different density gradient model parameter values translate into to very similar shapes of density gradient at distances less than 250 m.

It seems obvious at this point that one could build a likelihood that integrates the information from points and lines, since there is information about the density gradient in both. As illustrated before in the example of section 6.4, the information about the density gradient contained in the points alone does not allow a clear distinction between different models for the density gradient, and the additional data collected in perpendicular lines with respect to the linear feature at hand might contribute to facilitate such a distinction.

Note that problems can be anticipated for the analysis of the hares data set, because the estimated density gradient is considerably different when we consider the information contained in lines and points separately. I use the data here for illustration purposes, and the reasons for possible inconsistencies are reviewed in the discussion below.

The combined likelihood is given by the product of two distinct likelihoods, namely the points ($\mathcal{L}_p$) and the lines ($\mathcal{L}_l$) likelihoods, as

$$\mathcal{L}_{pl} = \mathcal{L}_p \mathcal{L}_l \tag{7.4}$$

where $\mathcal{L}_p$ is as in equation 6.11, or in the specific case of the hares data, the corresponding unequal angle effort likelihood equation 7.1, and $\mathcal{L}_l$ is as in equation 5.7, or in the case of the hares, the corresponding unequal line length equations 7.2 or 7.3.

The estimated detection function and density gradient model, using this likelihood, with a normal model for the density gradient (such that an estimated mode could be within or beyond 250 m), and considering the unequal line lengths and unequal angle effort options, is shown in figure 7.9.

When we estimate the density gradient from a joint likelihood, combining information from the point data and the secondary transect data, the latter largely determines the model shape (compare Figures 7.2, 7.4 and 7.9). This comes as no surprise, because the likelihood surface for point data alone is almost flat with respect to density gradient model parameters. This means that although we have only 9 observations in the secondary transects, these can change considerably the conclusions drawn from the data. The hare density estimate, assuming this is the correct model, is 3.27 hares/km^2.

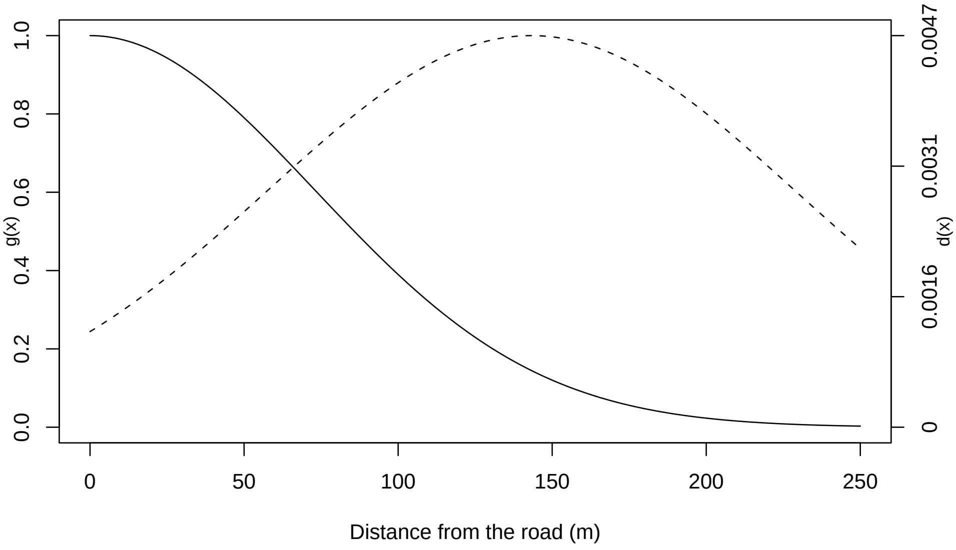

Figure 7.9: Republic of Ireland 2006 hare data estimated detection function and density gradient based on a likelihood combining information on points and lines (equation 7.4), considering a normal density gradient, and using the unequal line lengths and unequal angle effort options.

7.4 Discussion

Given the results presented, it is hard to put forward a single estimate for the density of hares in the Republic of Ireland in 2006, even for the covered area, which in this case corresponds to strips in the vicinity of roads. The task becomes more difficult when one wants to extrapolate for a wider area, which ideally would correspond to the entire country. Being side issues with respect to this thesis, multiple sampling and design issues make such task even harder.

With respect to the material presented in this chapter, the hares data set used had two shortcomings that might have led to results that are hard to interpret: (1) the objective of collecting the repeated points data was not to deal with the unequal effort in the main survey and (2) the secondary transects might have not been a random sample of possible transects.

The repeated points were not surveyed by the same teams responsible for the main survey. Besides that, the objective of these repeated points was not, *a priori*, to use the effort data to correct the main survey data for the sectors not visible. Hence there was no attempt to have clear definitions on what is a visible sector. There are also peculiarities in the point placement that might have been responsible for differences in the effort distribution of the main survey and the repeated squares. Therefore, the small changes in estimated density by incorporating this aspect in the analysis of the data might have been partly due to the inconsistencies across these two data sets. Nonetheless, provided one is aware of these methods, the ideal option will be to record at each point the sectors unavailable for detections to be made. Naturally, if animal density in a sector is dependent on a sector being visible or not, it becomes much harder to deal with such issue.

An alternative to condition inferences in the angles known to be visible could be to condition on the total effort at a point, as there was some evidence that, as one might expect, the smaller the effort at a point, the more such effort tended to be concentrated at sectors away from the road (data not shown).

If the information about the density gradient was consistent for secondary transects and points, we would expect the density gradients estimated to be similar, independently of the information used, and to achieve more consistent and reliable results by combining the methods of chapters 5 and 6. However, comparing all the estimated density gradients using or not the secondary data, we see that accounting for the secondary data leads to considerable changes, and these directly influence the estimated densities. The survey design according to which the secondary transects were selected was not random, and it is likely that this fact impaired the results, because the sampled secondary transects corresponded to areas which were easier to survey and hence not really representative of the whole vicinity of the road. An additional issue that might be partially responsible for the differences in estimated density gradient using the points data versus the secondary transects data is that the censoring in the secondary transects is not independent of the actual density gradient: given that the secondary transects stop at boundaries and presumably density is much lower near boundaries, the two are confounded.

Nonetheless, the data set served well for the main purpose of this chapter: illustrating how, once cast into a likelihood approach, the methods proposed in earlier chapters are readily generalized to more complicated scenarios.

Chapter 8

General discussion and potential new research directions

8.1 Introduction

This thesis deals with two main problems in CDS: (1) measurement error in the distances used for estimation, and (2) situations under which the usual availability of distances proportional to area fails. This failure is usually due to a flawed survey design, with samplers allocated non-independently of animal distribution, but it might also be the consequence of an inadequately small number of samplers. As shown in this work, both of these might lead to considerable bias if ignored, and hence methods to account for them in situations for which they cannot be avoided by adequate survey design and field methods are essential. The abundance of examples, referenced throughout this thesis, in which one or both are potential problems, clearly shows that these are issues worth pursuing further.

Despite being separate issues in practice, the key CDS assumptions (except for $g(0) = 1$) result in a common conceptual problem: the disruption of the underlying distribution of distances available for detection, which is assumed known by design for CDS estimators. In this sense, in terms of their consequences, measurement error, non-random samplers, small number of samplers and even animal movement can be conceptualized as a change in the uniform (for line transects) or triangular (for point transects) distributions of available distances, which would be expected by design. Hence, although the causes might be different, the consequences are very similar. Undetected animal movement might be seen as measurement error, likely with both a systematic and random component, provided animals do not exit or enter the covered strips during the survey. In the presence of random movement or attraction, the resulting distances could potentially be modeled using an underestimation error model, while for the avoidance, an overestimation error model may be adequate. Note, however, if movement is such that animals tend to move in and out of the covered area, density could be greatly overestimated due to the joint effect of the net underestimation of distances and an inflated encounter rate. Similarly, if samplers are laid along a linear feature which the animals avoid, this could be viewed as undetected avoidance movement or a tendency to overestimate distances, while if animals prefer areas near the linear feature, this could be viewed as attraction movement or underestimation of distances. Considering these similarities, it seems likely that methods to deal with measurement error or non uniformity might be also used to deal with say animal movement, with the key difference that the data needed to estimate parameters for a movement model might in practice be harder to collect than the data needed to describe an error model or a density gradient.

In the following, after a section with general comments on the results obtained regarding measurement error (section 8.2) and density gradients (section 8.3), a few possible avenues for further research (section 8.4) and some final conclusions (section 8.5) are presented.

8.2 Measurement error

The effects of measurement error and potential biases in distance sampling estimates were clearly demonstrated. Whenever the methods used to measure distances are sensitive to errors, accounting for their effect should be attempted. Investigators applying distance sampling methods should describe clearly the way in which distances are obtained – something that is sometimes ignored (e.g. Ashenafi *et al.*, 2005). At the very least they should discuss the possible impacts of measurement error for their specific data sets. Training and calibration of observers involved in the survey, as well as evaluating the quality of distance estimation with field trials, should always be considered (as in e.g. Bårdsen and Fox, 2006). The use of better technology should be attempted whenever possible, but even simple alternatives, like pacing (as in e.g. Catt *et al.*, 1998), should provide much better data than pure visual “guesstimates”.

Models can only expect to be representations of truth, but never truth itself. This is particularly true for simple models such as the additive or multiplicative models used for measurement error. The results in Alldredge *et al.* (in press) suggest that using additive or multiplicative measurement error models might be an oversimplification, with evidence for measurement errors in real life studies probably being neither simply additive or multiplicative. Consider, for example, a line transect survey where, rather than measuring perpendicular distances directly, radial distances plus sighting angles are recorded. It seems plausible that the measurement error in the resulting perpendicular distances is neither strictly additive or multiplicative, but rather the joint effect of the error in the two separate components. Nonetheless, the use of models is fundamental for the development and testing of new methods, but not necessarily without problems. As an example, pure additive models are unlikely to hold in most distance sampling scenarios. Following these, a true positive distance could be recorded as (or estimated to be) negative, which is not plausible for points and would require an animal being recorded on the wrong side of the transect for lines, although the latter might happen under some settings where the transect line is not physically marked. This brings additional problems when assessing methods by simulation: *ad-hoc* procedures for dealing with distances changing signs might be needed. This is far from ideal since the smaller distances are the ones more likely to suffer from this problem, but also those that are more influential for distance sampling estimators. Further studies where measurement error is assessed under controlled situations, like Baird and Burkhart (2000), Williams *et al.* (2007) or Alldredge *et al.* (in press), are needed to help clarify which models might be adequate and which should be avoided when dealing with measurement error.

The use of the likelihood approach to deal with measurement errors seems flexible enough to allow its integration in more complex scenarios, as is proposed in Burnham *et al.* (2004, p. 375-376) to incorporate measurement error with $MRDS$.

8.3 Density gradients

If a density gradient is known to be present in a given area then the simplest approach is to avoid the problem by design: place systematically or randomly spaced line transects perpendicular to the linear feature so that animals are distributed uniformly with respect to distance from the transect. An additional advantage of such a procedure is that the encounter rate variance tends to be smaller, because most lines will include both high and low density areas, rather than some lines being exclusively in high density and others exclusively in low density areas (see e.g. Ancrenaz *et al.*, 2004; Thomas *et al.*, 2007). Some additional examples of survey design considerations to avoid problems with non-random samplers can be found in Marsden and Pilgrim (2003) and Oppel (2006), namely trying to avoid potential density gradients promoted by tracks and rivers.

The problem of density gradients with respect to samplers should be taken seriously by investigators using distance sampling, and the implication of these findings is that placing points or lines along any linear feature should be considered only as a last resort.

Because of potential density gradients, non random samplers may lead to severe bias if the gradient is ignored, or strong assumptions if it is to be accounted for during estimation. As an example the influence of trails or roads on animals should not be assessed comparing these linear features as transects versus transects parallel to such features at a given distance from it (e.g. Kuitunen *et al.*, 1998; Miller *et al.*, 1998). The potential for confounding in the results is high, as the non-uniform distribution of animals might promote different bias at different distances from the trail or road.

The approach described here to deal with density gradients should be applicable even in the absence of a density gradient. However, it would be a poor default approach, because the requirement to estimate an extra function (the density gradient) in addition to the detection function leads to increased variance compared with the conventional estimator. Therefore, it should only be used if there is *a priori* reason to expect that density gradients exist.

The use of $\hat{D}(w)$ as an estimate of density in areas not covered, allowing inferences to extend to a wider survey region than what was covered, is only valid if the underlying model is a good representation of reality. This means that at distances greater than w from the samplers the density should be constant (when averaged across samplers). While this seems unrealistic, it is nevertheless more likely to be correct than the assumption that average density is equal to the density at zero distance, required for conventional methods to be valid. A formal test of uniformity of $\pi(x)$, for x large but $< w$, could be a way to assess whether this assumption is likely to be violated.

But even this test will not deal with the fact that this procedure is based on extrapolation of the estimated density gradient beyond the range of distances in the data, and hence extreme care is needed in the interpretation of these results.

As presented here, the methods assume the estimation of a mean density gradient. If the density gradient is thought to change considerably across the study area, say for example if this gradient depends on other covariates, such as ground slope or habitat, a more complex approach might be useful. An option might be the use of a parametric model for the density gradient, where the scale parameter could be a function of such variables. However, this would require strong assumptions about the density shape and the knowledge of these variables for the entire study area, which might be difficult to obtain in real life applications.

The non-random allocation of transects is likely to lead to a non-uniform distribution of animals with respect to samplers. The use of, say, minor roads, is an example of such a situation, quite common although generally criticized, and can lead to severe bias. I stress that this work does not try to justify such cases in any way, but does present options that might reduce bias if such methods are necessary due to practical considerations. However, it is important to point out that a sample of roads, even if these are a random sample of all available roads across a study region, is not a random sample with respect to the available habitat. Any relationship between road location and factors influencing animal distribution like the fact that roads tend to be in flatter areas, or along rivers, etc, will certainly introduce a bias extremely difficult to identify and remove. Tomás *et al.* (2001) give some good examples of why roads should be avoided.

Whether sensible estimates can be obtained by distance sampling, using transects laid along roads, has been addressed by Butler *et al.* (2005) for wild turkeys. This work showed that in this case the turkeys' use of habitat near the roads ($<$ 100 m from it) and away from roads ($>$ 100 m from it) is proportional to the amount of available habitat, at least for some combinations of the factors studied (sex, time of day and time of year). The study results allowed the authors to define optimal time periods for survey, in the sense of times when this proportionality held best. This study is important because it suggests that inferences from roads could not be extrapolated to the wider region during specific time periods, leading to either considerable under or overestimation depending on the period considered (see Butler *et al.* (2005) for further details). I recommend that similar studies are conducted when the use of roads as samplers cannot be avoided. However, it is worth noting that such results legitimize the extrapolation of the obtained estimate to the wider survey region (in the time and place they are carried out), but only provided the estimate in the covered area itself is unbiased. It is still possible to imagine that small scale strong reaction to roads could lead to bias in distance sampling estimates for the covered area (say if animals tended to have a strong preference for the first few meters around the road), yet the amount of survey effort needed to detect such a density gradient in a study like Butler *et al.* (2005) would likely be prohibitive. Because of this, one should also consider the extent to which any observed failure to detect a density gradient is caused by lack of statistical power in the study, rather than lack of a true effect. Reassuringly, however, it stands to reason that the less likely such an effect is to be detected, the less likely it is to influence the actual survey results.

Investigators conducting distance sampling studies in which location of samplers is not random with respect to the animals' locations should routinely assess whether substantial bias results; they should not simply assume no bias. I argue that if non-random samplers are used, this should be clearly stated when reporting results, and the possible implications for the results, given the characteristics of the study species, discussed. See Baldi *et al.* (2001) for such an example.

8.4 Further generalization of methods

With the exception of the preliminary work presented in chapters 3 (section 3.3) and 5 (section 5.2), the methods presented in this thesis were based on specifying parametric models and estimating those models parameters based on maximum likelihood. Once cast in such a framework, the potential for extending and combining methods is almost unlimited, as illustrated in chapter 7.

There are other natural extensions to these methods. A sensible extension to the methods presented is to use semi-parametric models both for the density gradient and the detection function, e.g. using the key+series adjustments approach of Buckland (1992). Extending this rationale further, it seems possible to consider also nonparametric alternatives, considering for example kernels (e.g. Mack and Quang, 1998) or splines (e.g. Rendas and Alpizar-Jara, 2005). Another possibility is to consider additional covariates in models of both the detection function and the density gradient, e.g., following the work of Marques and Buckland (2003). These should not require much in terms of methods development, although code implementation issues might become considerably harder.

An example of potential useful combination of these methods, which would be very interesting to implement in practice, relates to migration counts of whales (see Burnham *et al.*, 2004, p. 359-370 for further details about migration counts). The typical setting consists of having two independent platforms on shore that detect animals passing by. The observations are matched to identify duplicates. Given that the distance to shore is used as a covariate, this puts us in the realm of capture recapture methods (with distance as a covariate), but still not true $MRDS$ because the distribution of animals with respect to shore cannot be assumed uniform. However, one can extend these methods provided it is possible to use additional survey effort (e.g., an aerial survey platform) to survey transects perpendicular to the coast. This data would allow $\pi(x)$ to be estimated directly[1]. This can then be used to implement a true $MRDS$ approach, in the sense that $\pi(x)$ is included in the modelling, but that $\pi(x)$ no longer needs to be uniform, as originally suggested by Burnham *et al.* (2004, p. 364).

Note also that in the case where secondary transects are used to estimate the density gradient, I have assumed that the distance from the road to the animals observed is a random draw from the distribution of the density gradient (even if eventually only after the unequal effort is accounted for). This means that probability of detection along the secondary transects must be independent of the distance from the linear feature. If that is not the case, say because the detection function in the secondary transects is a function of covariates other than distance, which are correlated with distance from the linear feature (e.g. in the migration count example, if wind caused more turbulence as the plane gets further out in the sea, hence decreasing the detection function at larger distances from the coast), one could either (1) consider only the animals detected in the shoulder of the secondary transects detection function, hence avoiding the detection problem or (2) extend the likelihood to account for the effect of the additional covariates, provided these were recorded for each observation.

It seems possible to consider a full generalization of distance sampling methods, paraphrasing Barker and White (2004), a “mother-of-all-models”. This would be a global distance sampling model, potentially addressing most assumption violations, with $g(0) < 1$, multiple covariates, measurement error, non-uniform distribution of animals (and eventually movement and spatial models). This model would ideally

[1] Note I use $\pi(x)$ here, but $d(x)$ would be equivalent in this case, because despite the observations being made from a point, the observations are treated as line transect data, in the sense that all the observation are collapsed onto the line perpendicular to the coast passing by the observation platform, hence becoming perpendicular rather than radial distances.

be factorized such that each factor would correspond to a different component of the data. Then, given the situation at hand, one would simplify the model accordingly, dealing only with the model components that were deemed necessary.

8.5 Concluding remarks

One obvious conclusion is that for most problems, and most certainly for measurement error and non-uniform transects, it is best to avoid the problems by using adequate field methods and survey design than to attempt to deal with the problems at the analysis stage. Hence, the need for adequate pilot surveys is strongly stressed, with the explicit objective of minimizing assumption violation, allowing fine tuning of both survey design and field methods.

The development of statistical models is important, but researchers need to focus also on ways to make their research accessible to practitioners, under the risk of methods becoming useless in practice. It seems symptomatic that methods to deal with measurement error, like Chen (1998), Chen and Cowling (2001) or Marques (2004), have not, to the best of my knowledge, ever been implemented by others. These references have not been cited unless to strengthen general comments on the need to address the problem. Practitioner-orientated publications and free software development should be a priority for applied statisticians whenever methods are developed and believed to be of general use.

As it tends to happen with statistical methods applied by a wide range of practitioners, the failure of assumptions of distance sampling methods is sometimes ignored in practice. The view often seems to be that once the data have been collected, they should be analyzed and conclusions drawn even if there is good evidence that problems could arise due to assumption failure. It is usually more difficult to get away with this negligent approach to assumption failure if (1) the effect of assumption violation is well understood from a theoretical point of view, and the resulting bias known, (2) illustrative real life examples exist showing how poor estimates might be when assumption failure is ignored, and (3) methods to explicitly account for the failure of assumptions are available. This thesis is an attempt to provide all three of these in the case of measurement error and the use of non-random samplers in distance sampling. In the future it will be harder to claim that the dangers of doing so are unclear or unstated, or that no methods exist to deal with them.

Bibliography

Alldredge, M. W., Simons, T. R., and Pollock, K. H. (in press). An experimental evaluation of distance measurement error in avian point count surveys. *Journal of Wildlife Management*.

Alpizar-Jara, R. (1997). *Assessing assumption violation in line transect sampling*. Ph.D. thesis, North Carolina State University, Raleigh.

Ancrenaz, M., Goossens, B., Gimenez, O., Sawang, A., and Lackman-Ancrenaz, I. (2004). Determination of ape distribution and population size using ground and aerial surveys: a case study with orang-utans in lower Kinabatangan, Sabah, Malaysia. *Animal Conservation*, **7**, 375–385.

Anderson, D. R. (2001). The need to get the basics right in wildlife field studies. *Wildlife Society Bulletin*, **29**, 1294–1297.

Anderson, D. R. (2003). Response to Engeman: index values rarely constitute reliable information. *Wildlife Society Bulletin*, **31**, 288–291.

Anderson, D. R. and Pospahala, R. S. (1970). Correction of bias in belt transects of immotile objects. *Journal of Wildlife Management*, **34**, 141–146.

Anderson, D. R., Burnham, K. P., Lubow, B. C., Thomas, L., Corn, P. S., Medica, P. A., and Marlow, R. W. (2001). Field trials of line transect methods applied to estimation of desert tortoise abundance. *Journal of Wildlife Management*, **65**, 583–597.

Ashenafi, Z. T., Coulson, T., Sillero-Zubiri, C., and Leader-Williams, N. (2005). Behaviour and ecology of the Ethiopian wolf (*Canis simensis*) in a human-dominated landscape outside protected areas. *Animal Conservation*, **8**, 113–121.

Baird, R. W. and Burkhart, S. M. (2000). Bias and variability in distance estimation on the water: implications for the management of whale watching. IWC meeting document SC/52/WW1, IWC.

Baldi, R., Albon, S. D., and Elston, D. A. (2001). Guanacos and sheep: Evidence for continuing competition in arid Patagonia. *Oecologia*, **129**, 561–570.

Bårdsen, B.-J. and Fox, J. (2006). Evaluation of line transect sampling for density estimates of chiru *Pantholops hodgsoni* in the Aru Basin, Tibet. *Wildlife Biology*, **12**, 89–100.

Barker, R. J. and White, G. C. (2004). Towards the mother-of-all-models : customised construction of the mark-recapture likelihood function. *Animal Biodiversity and Conservation*, **27**, 177–185.

Bart, J., Droege, S., Geissler, P., Peterjohn, B., and Ralph, C. J. (2004). Density estimation in wildlife surveys. *Wildlife Society Bulletin*, **32**, 1242–1247.

Beavers, S. C. and Ramsey, F. L. (1998). Detectability analysis in transect surveys. *Journal of Wildlife Management*, **62**, 948–957.

Becker, B. H., Beissinger, S. R., and Carter, H. R. (1997). At-sea density monitoring of marbled murrelets in central California: methodological considerations. *Condor*, **99**, 743–755.

Biswas, S. and Sankar, K. (2002). Prey abundance and food habit of tigers (*Panthera tigris tigris*) in Pench National Park, Madhya Pradesh, India. *Journal of Zoology*, **256**, 411–420.

Boano, G. and Toffoli, R. (2002). A line transect survey of wintering raptors in the western Po plain of northern Italy. *Journal of Raptor Research*, **36**, 128–135.

Borchers, D. L. (1996). *Line Transect Estimation with Uncertain Detection on the Trackline*. Ph.D. thesis, University of Cape Town, Cape Town.

Borchers, D. L., Buckland, S. T., Goedhart, P. W., Clarke, E. D., and Hedley, S. L. (1998a). Horvitz-Thompson estimators for double-platform line transect surveys. *Biometrics*, **54**, 1221–1237.

Borchers, D. L., Zucchini, W., and Fewster, R. (1998b). Mark-recapture models for line transect surveys. *Biometrics*, **54**, 1207–1220.

Borchers, D. L., Buckland, S. T., and Zucchini, W. (2002). *Estimating Animal Abundance*. Springer, London.

Borchers, D. L., Laake, J. L., Southwell, C., and Paxton, C. G. M. (2006). Accommodating unmodelled heterogeneity in double-observer distance sampling surveys. *Biometrics*, **62**, 372–378.

Borchers, D. L., Marques, T. A., and Gunnlaugsson, T. (in prep a). Distance sampling with measurement errors.

Borchers, D. L., Pike, D., Gunnlaugsson, T., and Vikingson, G. A. (in review). Minke whale abundance estimation from the NASS 1987 and 2001 cue counting surveys taking account of distance estimation errors. *North Atlantic Marine Mammal Commission Special Issue*.

Borralho, R., Rego, F., and Pinto, P. V. (1996). Is driven transect sampling suitable for estimating red-legged partridge *Alectoris rufa* densities? *Wildlife Biology*, **2**, 259–268.

Brown, J. A. and Boyce, M. S. (1998). Line transect sampling of Karner blue butterflies (*Lycaeides melissa samuelis*). *Environmental and Ecological Statistics*, **5**, 81–91. ID263.

Buckland, S. T. (1992). Fitting density functions with polynomials. *Applied Statistics*, **41**, 63–76.

Buckland, S. T. and Anganuzzi, A. A. (1988). Comparison of smearing methods in the analysis of minke sightings data from IWC/IDCR Antarctic cruises. *Report of the International Whaling Commission*, **38**, 257–263.

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

Buckland, S. T., Anderson, D. R., Burnham, K. P., and Laake, J. L. (1993a). *Distance Sampling - Estimating abundance of biological populations*. Chapman and Hall, London.

Buckland, S. T., Breiwick, J. M., Cattanach, K. L., and Laake, J. L. (1993b). Estimated population size of the California gray whale. *Marine Mammal Science*, **9**, 235–249.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, **53**, 603–618.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., and Thomas, L. (2001). *Introduction to distance sampling - Estimating abundance of biological populations*. Oxford University Press, Oxford.

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D., and Thomas, L. (2004). *Advanced Distance Sampling*. Oxford University Press, Oxford.

Buckland, S. T., Borchers, D. L., Johnston, A., Henrys, P. A., and Marques, T. A. (2007). Line transect methods for plant surveys. *Biometrics*.

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

Burnham, K. P., Buckland, S. T., Laake, J. L., Borchers, D. L., Marques, T. A., Bishop, J. R. B., and Thomas, L. (2004). Further topics in distance sampling. In S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas, editors, *Advanced Distance Sampling*, pages 307–392. Oxford University Press, Oxford.

Butler, M. J., Wallace, M. C., Ballard, W. B., DeMaso, S. J., and Applegate, R. D. (2005). From the field: The relation of Rio Grande wild turkey distributions to roads. *Wildlife Society Bulletin*, **33**, 745–748.

Butterworth, D. S. (1982). On the functional form used for $g(y)$ for minke whale sightings, and bias in its estimation due to measurement inaccuracies. *Report of the International Whaling Commission*, **32**, 883–888.

Butterworth, D. S., Best, P. B., and Hembree, D. (1984). Analysis of experiments carried out during the 1981/82 IWC/IDCR Antarctic minke whale assessment cruise in Area II. *Report of the International Whaling Commission*, **34**, 365–392.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995). *Measurement error in nonlinear models*. Chapman & Hall London.

Catt, D. C., Baines, D., Picozzi, N., Moss, R., and Summers, R. W. (1998). Abundance and distribution of capercaillie *Tetrao urogallus* in Scotland 1992-1994. *Biological Conservation*, **85**, 257–267.

Chen, S. X. (1996). A kernel estimate for the density of a biological population by using line transect sampling. *Applied Statistics*, **45**, 135–150.

Chen, S. X. (1998). Measurement errors in line transect surveys. *Biometrics*, **54**, 899–908.

Chen, S. X. and Cowling, A. (2001). Measurement errors in line transect surveys where detectability varies with distance and size. *Biometrics*, **57**, 732–742.

Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, **89**, 1314–1328.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge.

DeJong, M. J. and Emlen, J. T. (1985). The shape of the auditory detection function and its implication for songbird censusing. *Journal of Field Ornithology*, **56**, 213–223.

Diefenbach, D. R., Brauning, D. W., and Mattice, J. A. (2003). Variability in grassland bird counts related to observer differences and species detection rates. *The Auk*, **120**, 1168–1179.

Diefenbach, D. R., Finley, J. C., Luloff, A. E., Stedman, R., Swope, C. B., Zinn, H. C., and San Julian, G. J. (2005). Bear and deer hunter density and distribution on public land in Pennsylvania. *Human Dimensions of Wildlife*, **10**, 201–212.

Dörgeloh, W. G. (2005). Density estimates of francolin in a *Sporobolus iocladosAcacia tortilis* Savanna using distance sampling. *South African Journal of Wildlife Research*, **35**(1), 89–94.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman Hall, New York.

Ellingson, A. R. and Lukacs, P. M. (2003). Improving methods for regional landbird monitoring: a reply to Hutto and Young. *Wildlife Society Bulletin*, **31**, 896–902.

Fewster, R. M. and Buckland, S. T. (2004). Assessment of distance sampling estimators. In S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas, editors, *Advanced Distance Sampling*, pages 281–306. Oxford University Press, Oxford.

Fewster, R. M., Laake, J. L., and Buckland, S. T. (2005). Line transect sampling in small and large regions. *Biometrics*, **61**, 856–859.

Fewster, R. M., Buckland, S. T., Burnham, K. P., Borchers, D. L., Jupp, P. E., Laake, J. L., and Thomas., L. (in review). Estimating the encounter rate variance in distance sampling. *Biometrics*.

Fletcher, J. R. and Hutto, L. R. (2006). Estimating detection probabilities of river birds using double surveys. *The Auk*, **123**, 695–707.

Gordon, J. (2001). Measuring the range to animals at sea from boats using photographic and video images. *Journal of Applied Ecology*, **38**, 879–887.

Hammond, P. S. (1984). An investigation into the effects of different techniques of smearing the IWC/IDCR minke whale sighting data and the use of different models to estimate density of schools. *Report of the International Whaling Commission*, **34**, 301–307.

Harmata, A. R., Podruzny, K. M., Zelenak, J. R., and Morrison, M. L. (1999). Using marine surveillance radar to study bird movements and impact assessment. *Wildlife Society Bulletin*, **27**, 44–52.

Hedley, S. L. (2000). *Modelling heterogeneity in cetacean surveys*. Ph.D. thesis, Univerity of St. Andrews, St Andrews.

Hedley, S. L., Buckland, S. T., and Borchers, D. L. (1999). Spatial modelling from line transect data. *Journal of Cetacean Research and Management*, **1**, 255–264.

Heydon, M. J., Reynolds, J. C., and Short, M. J. (2000). Variation in abundance of foxes (*Vulpes vulpes*) between three regions of rural Britain, in relation to landscape and other variables. *Journal of Zoology*, **251**, 253–264.

Hiby, L. and Krishna, M. B. (2001). Line transect sampling from a curved path. *Biometrics*, **57**, 727–731.

Hiby, L., Ward, A., and Lovell, P. (1989). Analysis of the North Atlantic sightings survey 1987: Aerial survey results. *Report of the International Whaling Commission*, **39**, 447–455.

Hounsome, T. D., Young, R. P., Davison, J., Yarnell, R. W., Trewby, I. D., Garnettand, B. T., Delahay, R. J., and Wilson, G. J. (2005). An evaluation of distance sampling to estimate badger (*Meles meles*) abundance. *Journal of Zoology, London*, **266**, 81–87.

Hutto, R. L. and Young, J. S. (2002). Regional landbird monitoring: perspectives from the Northern Rocky Mountains. *Wildlife Society Bulletin*, **30**, 738–750.

Hutto, R. L. and Young, J. S. (2003). On the design of monitoring programs and the use of population indices: a reply to Ellingson and Lukacs. *Wildlife Society Bulletin*, **31**, 903–910.

Kuitunen, M., Rossi, E., and Stenroos, A. (1998). Do highways influence density of land birds? *Environmental Management*, **22**, 297–302.

Kulbicki, M. and Sarramegna, S. (1999). Comparison of density estimates derived from strip transect and distance sampling for underwater visual censuses: a case study of Chaetodontidae and Pomacanthidae. *Aquatic Living Resources*, **12**, 315–325.

Laake, J. L. (1978). *Line transect sampling estimators robust to animal movement*. Master's thesis, Utah State University, Logan.

Laake, J. L. and Borchers, D. L. (2004). Methods for incomplete detection at distance zero. In S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas, editors, *Advanced Distance Sampling*, pages 108–189. Oxford University Press, Oxford.

Mack, Y. and Quang, P. X. (1998). Kernel methods in line and point transect sampling. *Biometrics*, **54**, 606–619.

Marques, F. F. C. (2001). *Estimating wildlife distribution and abundance from line transect surveys conducted from platforms of opportunity*. Ph.D. thesis, University of St. Andrews, St Andrews.

Marques, F. F. C. and Buckland, S. T. (2003). Incorporating covariates into standard line transect analyses. *Biometrics*, **59**, 924–935.

Marques, F. F. C. and Buckland, S. T. (2004). Covariate models for the detection function. In S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas, editors, *Advanced Distance Sampling*, pages 31–47. Oxford University Press, Oxford.

Marques, T. A. (2004). Predicting and correcting bias caused by error measurement in line transect sampling using multiplicative error models. *Biometrics*, **60**, 757–763.

Marques, T. A. and Buckland, S. T. (2005). Transectos lineares em situações de não uniformidade das distâncias disponíveis para detecção. In C. A. Braumann,

P. Infante, M. M. Oliveira, R. Alpízar-Jara, and F. Rosado, editors, *Estatística Jubilar*, pages 445–454. SPE, Evora.

Marques, T. A., Andersen, M., Christensen-Dalsgaard, S., Belikov, S., Boltunov, A., Wiig, O., Buckland, S. T., and Aars, J. (2006). The use of global positioning systems to record distances in a helicopter line-transect survey. *Wildlife Society Bulletin*, **34**, 759–763.

Marques, T. A., Thomas, L., Fancy, S. G., and Buckland, S. T. (in press). Improving estimates of bird density using multiple covariate distance sampling. *The Auk*, **124**.

Marsden, S. J. and Pilgrim, J. D. (2003). Factors influencing the abundance of parrots and hornbills in pristine and disturbed forests on New Britain, PNG. *Ibis*, **145**, 45–53.

Melville, G. and Welsh, A. H. (2001). Line transect sampling in small regions. *Biometrics*, **57**, 1130–1137.

Miller, S. G., Knight, R. L., and Miller, C. K. (1998). Influence of recreational trails on breeding birds communities. *Ecological Applications*, **8**, 162–169.

Norvell, R. E., Howe, F. P., and Parrish, J. R. (2003). A seven-year comparison of relative-abundance and distance-sampling methods. *The Auk*, **120**, 1013–1028.

Ogutu, J., Bhola, N., Piepho, H.-P., and Reid, R. (2006). Efficiency of strip- and line-transect surveys of African savanna mammals. *Journal of Zoology*, **269**, 149–160.

Øien, N. and Schweder, T. (1992). Estimates of bias and variability in visual distance measurements made by observers during shipboard surveys of Northeastern Atlantic minke whales. *Report of the International Whaling Commission*, **42**, 407–412.

Oppel, S. (2006). Using distance sampling to quantify Odonata density in tropical rainforests. *International Journal of Odonatology*, **9**, 81–88.

Otto, M. C. and Pollock, K. H. (1990). Size bias in line transect sampling: a field test. *Biometrics*, **46**, 239–245.

Palka, D. and Hammond, P. S. (2001). Accounting for responsive movement in line transect estimates of abundance. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 777–787.

Pollock, K. H., Nichols, J. D., Simons, T. R., Farnsworth, G. L., Bailey, L. L., and Sauer, J. R. (2002). Large scale wildlife monitoring studies: statistical methods for design and analysis. *Environmetrics*, **13**, 105–119.

R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rendas, L. and Alpizar-Jara, R. (2005). O modelo logspline aplicado aos transectos lineares. In C. A. Braumann, P. Infante, M. M. Oliveira, R. Alpízar-Jara, and F. Rosado, editors, *Estatística Jubilar*, pages 629–640. SPE, Evora.

Rosenstock, S. S., Anderson, D. R., Giesen, K. M., Leukering, T., and Carter, M. F. (2002). Landbird counting techniques: current practices and an alternative. *The Auk*, **119**, 46–53.

Ruette, S., Stahl, P., and Albaret, M. (2003). Applying distance-sampling methods to spotlight counts of red foxes. *Journal of Applied Ecology*, **40**, 32–43.

Schwarz, C. J. and Seber, G. A. F. (1999). Estimating animal abundance: review III. *Statistical Science*, **14**, 427–56.

Schweder, T. (1996). A note on a buoy-sighting experiment in the North Sea in 1990. *Report of the International Whaling Commission*, **46**, 383–385.

Schweder, T. (1997). Measurement error models for the Norwegian minke whale survey in 1995. *Report of the International Whaling Commission*, **47**, 485–488.

Schweder, T., Skaug, H. J., Langaas, M., and Dimakos, X. K. (1999). Simulated likelihood methods for complex double-platform line transect surveys. *Biometrics*, **55**, 678–687.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance, 2nd Ed.* Griffin, London.

Seber, G. A. F. (1986). A review of estimating animal abundance. *Biometrics*, **42**, 267–292.

Seber, G. A. F. (1992). A review of estimating animal abundance II. *International Statistical Review*, **60**, 129–166.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

Simons, T. R., Shriner, S. A., and Farnsworth, G. L. (2006). Comparison of breeding bird and vegetation communities in primary and secondary forests of Great Smoky Mountains National Park. *Biological Conservation*, **129**, 302–311.

Skaug, H. J. and Schweder, T. (1999). Hazard models for line transect surveys with independent observers. *Biometrics*, **55**, 29–36.

Skaug, H. J., Øien, N., Schweder, T., and Bøthun, G. (2004). Abundance of minke whales (*Balaenoptera acutorostrata*) in the Northeast Atlantic: variability in time and space. *Canadian Journal of Fisheries and Aquatic Sciences*, **61**, 870–886.

Smith, G. E. J. (1979). Some aspects of line transect sampling when the target population moves. *Biometrics*, **35**, 323–329.

Southwell, C., de la Mare, B., Underwood, M., Quartararo, F., and Cope, K. (2002). An automated system to log and process distance sight-resight aerial survey data. *Wildlife Society Bulletin*, **30**, 394–404.

Strindberg, S., Buckland, S. T., and Thomas, L. (2004). Design of distance sampling surveys and geographic information systems. In S. T. Buckland, D. R. Anderson, K. P. Burnham, J. L. Laake, D. L. Borchers, and L. Thomas, editors, *Advanced Distance Sampling*, pages 190–228. Oxford University Press, Oxford.

Thomas, L., Laake, J. L., Strindberg, S., Marques, F. F. C., Buckland, S. T., Borchers, D. L., Burnham, K. P., Hedley, S. L., and Pollard, J. H. (2002). Distance 4.0. release 2. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK. http://www.ruwpa.st-and.ac.uk/distance/.

Thomas, L., Laake, J. L., Strindberg, S., Marques, F. F. C., Buckland, S. T., Borchers, D. L., Anderson, D. R., Burnham, K. P., Hedley, S. L., Pollard, J. H., Bishop, J. R. B., and Marques, T. A. (2005). Distance 5.0. beta 4. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK. http://www.ruwpa.st-and.ac.uk/distance/.

Thompson, D. and Hiby, A. R. (1985). The use of scale binoculars for distance estimation and a time lapse camera for angle estimation during the 1983/84 IDCR minke whale assessment cruise. *Reports of the International Whaling Commission*, **35**, 309–314.

Thompson, S. K. (1992). *Sampling*. John Wiley & Sons, Inc.

Thompson, W. L. (2002). Towards reliable bird surveys: accounting for individuals present but not detected. *The Auk*, **119**, 18–25.

Tobias, J. A. and Seddon, N. (2002). Estimating population size in the subdesert mesite (*Monias benschi*): new methods and implications for conservation. *Biological Conservation*, **108**, 199–212.

Tomás, W. M., McShea, W., de Miranda, G. H. B., Moreira, J. R., Mourão, G., and Lima Borges, P. A. (2001). A survey of a pampas deer, *Ozotoceros bezoarticus leucogaster* (Arctiodactyla, Cervidae), population in the Pantanal wetland, Brazil, using the distance sampling technique. *Animal Biodiversity and Conservation*, **24**, 101–106.

Turnock, B. J. and Quinn, II, T. J. (1991). The effect of responsive movement on abundance estimation using line transect sampling. *Biometrics*, **47**, 701–716.

Ward, A. I., White, P. C. L., and Critchley, C. H. (2004). Roe deer *Capreolus capreolus* behaviour affects density estimates from distance sampling surveys. *Mammal Review*, **34**, 315–319.

Welsh, A. H. (2002). Incomplete detection in enumeration surveys: whither distance sampling. *Australian & New Zealand Journal of Statistics*, **44**, 13–22.

Williams, R., Leaper, R., Zerbini, A. N., and Hammond, P. S. (2007). Methods for investigating measurement error in cetacean line-transect surveys. *Journal of the Marine Biological Association of the United Kingdom*, **87**, 313–320.

Appendix A

Developing the methods of section 6.2 starting from $\pi(x, r)$

The methods in section 6.2, dealing with points along linear features to which animals respond to, were originally developed starting from the distribution of the perpendicular distances (x) and the radial distances (r). This was intuitive in the sense that we have two processes involved, the detection function, which I assume a function of r alone, and the density gradient, which I assume a function of x alone.

Here is presented the original derivation under those settings, which was actually the approach implemented in the R code used. As can be seen below, and not surprisingly, the end result of implementing one or the other is the same (cf. equations A.11 and 6.10). Some additional results are shown to be the same under either approach.

A.1 Deriving the likelihood

We want to obtain $\pi(x, r) = \pi(r|x)\pi(x)$. Note that the $\pi(x)$ is obtained just as before (see equation 6.4), hence only the conditional distribution is slightly different. We can see that, given $X = x$, the value of r can take as a minimum value x, and as a maximum value w (Figure 6.1b). And also that given $X = x$

$$R|_{X=x} = \sqrt{x^2 + U^2} \tag{A.1}$$

where $U \sim Uniform(0, \sqrt{w^2 - x^2})$. This uniform arises based on the fact that if points are randomly placed along the transect, any value for U is equally likely, where U is a distance measured from the point $(x, 0)$ to the point $(x, \sqrt{r^2 - x^2})$, in a coordinate system where the origin (0,0) is at the center of the circle representing the point transect (Figure 6.1c). Therefore,

$$F_{R|X}(r) = Pr(R|X \leq r) = Pr(\sqrt{x^2 + U^2} \leq r) = F_U(\sqrt{r^2 - x^2}) \tag{A.2}$$

and hence it follows that

$$\begin{aligned}
\pi_{R|X}(r) &= \frac{dF_{R|X}(r)}{dr} && \text{(A.3)}\\
&= \frac{dF_U(\sqrt{r^2 - x^2})}{dr} && \text{(A.4)}\\
&= \pi_u(\sqrt{r^2 - x^2})\frac{d(\sqrt{r^2 - x^2})}{dr} && \text{(A.5)}\\
&= \frac{r}{\sqrt{w^2 - x^2}\sqrt{r^2 - x^2}} && \text{(A.6)}
\end{aligned}$$

which leads to the joint distribution being

$$\pi(x, r) = \pi(r|x)\pi(x) = \frac{r}{\sqrt{w^2 - x^2}\sqrt{r^2 - x^2}} \frac{d(x)\varphi(x)}{\int\limits_0^w d(x)\varphi(x)dx}. \tag{A.7}$$

Note that as with the approach presented in the thesis main chapter, we can now build a *pdf* of detected distances that is used to estimate the parameters of the processes involved, namely

$$f(x, r) = \frac{\pi(x, r)g(r)}{\int\limits_R \int\limits_X \pi(x, r)g(r)dxdr}. \tag{A.8}$$

Using the appropriate expressions leads to

$$f(x,r) = \frac{\frac{r}{\sqrt{w^2-x^2}\sqrt{r^2-x^2}} \frac{d(x)\varphi(x)}{\int_0^w d(x)\varphi(x)dx} g(r)}{\int_R \int_X \frac{r}{\sqrt{w^2-x^2}\sqrt{r^2-x^2}} \frac{d(x)\varphi(x)}{\int_0^w d(x)\varphi(x)dx} g(r)dxdr} \tag{A.9}$$

and this expression simplifies considerably because the integral in $\pi(x)$ denominator is a constant after integration, leading to

$$f(x,r) = \frac{\frac{r}{\sqrt{r^2-x^2}} d(x)g(r)}{\int_0^w \int_0^r \frac{r}{\sqrt{r^2-x^2}} d(x)g(r)dxdr}. \tag{A.10}$$

The integral involved in the denominator turned out to be hard to implement. By transforming $x = r\sin\theta$ ($dx = r\cos\theta \; d\theta$) it simplifies to

$$f(x,r) = \frac{\frac{1}{\sqrt{1-\frac{x^2}{r^2}}} d(x)g(r)}{\int_0^w \int_0^{\frac{\pi}{2}} r \; d(r\sin\theta)g(r)d\theta dr}. \tag{A.11}$$

After logs, simplifying and discarding constants not dependent on the parameters this corresponds to maximizing the following log-likelihood

$$l(\sigma_1, \sigma_2|\underline{x}, \underline{r}) = [n \; ln \int_0^w \int_0^{\frac{\pi}{2}} r \; d(r\sin\theta)g(r)d\theta dr]^{-1} + \sum_{i=1}^n log[d(x)g(r)] \tag{A.12}$$

which was the actual likelihood implemented. Apart from a constant which is not a function of the parameters, equation A.11 is the same as equation 6.10, and hence the same parameter estimates are obtained by considering the likelihood based on either of these.

A.2 Distribution of radial distances given the density gradient is uniform

Under this setting, the distribution of r, the radial distances in a circle of radius w, is given by

$$\pi(r) = \int\limits_X \pi_{R|X}(r)\pi(x)dx. \tag{A.13}$$

On the other hand, under the conventional setting, the distribution $\pi(r)$ is given by equation 2.27,

$$\pi(r) = \frac{2r}{w^2}. \tag{A.14}$$

Here I show that, as expected, given $\pi(x)$ is uniform, A.13 leads to A.14. Starting from A.13, we have

$$\pi(r) = \int\limits_0^r \pi_{R|X}(r)\pi(x)\ dx \tag{A.15}$$

$$= \int\limits_X \frac{r}{\sqrt{w^2 - x^2}\sqrt{r^2 - x^2}} \frac{d(x)\varphi(x)}{\int\limits_0^w d(x)\varphi(x)\ dx}\ dx \tag{A.16}$$

which considering an uniform $d(x)$ leads to

$$u(r) = \int_0^r \frac{r}{\sqrt{w^2-x^2}\sqrt{r^2-x^2}} \frac{\frac{1}{w}\frac{\sqrt{w^2-x^2}}{w}}{\int_0^w \frac{1}{w}\frac{\sqrt{w^2-x^2}}{w}\ dx}\ dx \tag{A.17}$$

$$= \int_0^r \frac{r}{\sqrt{r^2-x^2}\int_0^w \sqrt{w^2-x^2}\ dx}\ dx \tag{A.18}$$

$$= \frac{r}{\int_0^w \sqrt{w^2-x^2}dx} \int_0^r \frac{1}{\sqrt{r^2-x^2}}\ dx. \tag{A.19}$$

Given a constant a,

$$\int \frac{1}{\sqrt{a^2-x^2}}\ dx = sin^{-1}(\frac{x}{a}) + C \tag{A.20}$$

and

$$\int \sqrt{a^2-x^2}\ dx = \frac{x\sqrt{a^2-x^2}}{2} + \frac{a^2}{2}sin^{-1}(\frac{x}{a}) + C \tag{A.21}$$

are standard integration results (C is a constant), which applied to the integrals in equation A.19 lead to

$$\int_0^w \sqrt{w^2-x^2}\ dx = \frac{w^2\pi}{2} \tag{A.22}$$

$$\int_0^r \frac{1}{\sqrt{r^2-x^2}}\ dx = \pi \tag{A.23}$$

hence

$$u(r) = \frac{r}{\frac{w^2\pi}{2}}\pi \tag{A.24}$$

$$= \frac{2r}{w^2} \tag{A.25}$$

which concludes the demonstration.

A.3 Calculating the probability of detection

As in the conventional methods, the probability of detection is the mean value of the detection function, with respect to the available distances, given by

$$P = \int_0^w g(r)\pi(r)dr. \tag{A.26}$$

The only added complication is the way the distribution of available distances is obtained. Under the methods of chapter 6, the above expression becomes

$$P = \int_0^w g(r) \int_X \pi(x,r)dxdr = \int_0^w g(r) \int_X \pi(r|x)\pi(x)dxdr. \tag{A.27}$$

Using the appropriate expressions, this leads to

$$P = \int_0^w g(r) \int_0^r \frac{r}{\sqrt{w^2-x^2}\sqrt{r^2-x^2}} \frac{d(x)\varphi(x)}{\int_0^w d(x)\varphi(x)dx} dxdr \tag{A.28}$$

which simplifies to

$$\begin{aligned} P &= \int_0^w g(r) \int_0^r \frac{r}{\sqrt{r^2-x^2}} \frac{d(x)}{\int_0^w d(x)\varphi(x)dx} dxdr & \text{(A.29)} \\ &= \int_0^w \frac{rg(r)}{\int_0^w d(x)\varphi(x)dx} \int_0^r \frac{d(x)}{\sqrt{r^2-x^2}} dxdr. & \text{(A.30)} \end{aligned}$$

Using as before the transformation $x = r\sin\theta$ ($dx = r\cos\theta \; d\theta$) this further simplifies to

$$\frac{1}{\int\limits_0^w d(x)\varphi(x)dx} \int\limits_0^w \int\limits_0^{\frac{\pi}{2}} rg(r)d(r\sin\theta)d\theta dr. \tag{A.31}$$

This expression can be evaluated numerically, as an example using the multi-dimensional integration routine *adapt* in R. Note that, not surprisingly, this is the same as equation 6.16.