A Method for Analysing Heterogeneous Data with 2D Detection Function Distance Sampling

Calliste Fagard-Jenkin

January 2018

1 Abstract

This project focuses on line-transect distance sampling; a hybrid method which uses both design and model based inference to produce estimates on the sizes of animal populations. An extension of distance sampling by (Borchers & Cox, 2017) allows for the use of 2D detection functions, describing the probability of detection of an animal in both forward and perpendicular directions. This methodology produces estimates of abundance which are far less biased when compared to estimates produced by conventional distance sampling, when the animals in the population of interest display responsive movement. The raison d'être of this project is to further extend the codebase of the (Borchers & Cox, 2017) methodology's R package (called LT2D) to include the possibility of covariate inclusion to cope with data that are heterogeneous with respect to detection probability, and to provide a number of illustrative analyses on a range of simulated datasets to demonstrate the improved functionality.

2 Introduction

2.1 Context and background

2.1.1 What is distance sampling?

Distance sampling is a methodology within ecological statistics which allows us to produce unbiased estimates of animal population sizes even when the detection of any individual in the survey area is uncertain. First, a quick description of its simpler alternative, strip sampling (in which detection of individuals is certain), will be given in order to illustrate the fundamental principles of greatest importance.

Consider a survey area within which the population of interest is closed (there is no emigration or immigration of individuals). We further assume that births and deaths are considered to occur at a negligible rate throughout the duration of the survey. In this way, the survey is considered to be a snapshot of the population at some given moment. In our example, we choose a square survey area, which we subdivide into five strips of regular width. We must further assume that animals are uniformly distributed throughout the survey area, in both the perpendicular and the horizontal axis (in fact, we must simply assume that the distribution of animals with respect to the transect is known. Uniformity arises as a natural consequence of independently distributed transect lines in conventional distance sampling. In the case of point transect distance sampling, the assumed distribution is triangular). The below figure illustrates uniformly distributed animals within a survey area.

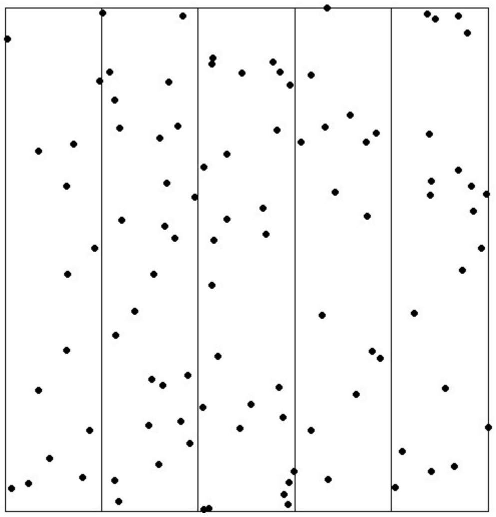

Figure 1: Example distribution of animals in a square survey area divided into 5 equal strips

Given this rather simplistic design, obtaining an estimate of abundance in the survey area ($\hat{N}$) is extremely simple, so long as we are able to sample all of the individuals within whole strips. As an example we consider strips 2 and 4 (from the left) as the randomly selected strips which form our survey sample. Strip 2 contains 25 individuals, and strip 4 contains 15, giving us a total of $n = 40$ observed individuals. If we denote the total survey area by A and the sampled area by a, then we observe that our survey effort of $\frac{a}{A} = \frac{2}{5}$. Hence a reasonable intuitive estimator for the abundance in the whole survey area is

$$\hat{N} = \frac{n}{\frac{a}{A}} = \frac{40}{0.4} = 100 \tag{1}$$

which is a simple rescaling of the number of observed animals which accounts for the portion of the survey area which has remained unseen. This example forms a basic overview of an easy methodology we can use when detection of every individual is guaranteed, however, we would now like to find a way of relaxing this rather unrealistic assumption. If we assume all animals in the strip have some fixed probability p of being detected, then we can modify our estimator to account for the animals we expect we have missed:

$$\hat{N} = \frac{n}{\frac{a}{A}p} \tag{2}$$

This again is nothing more than rescaling the estimate by an appropriate constant, to account for animals that we believe were present, but not detected by any observers.

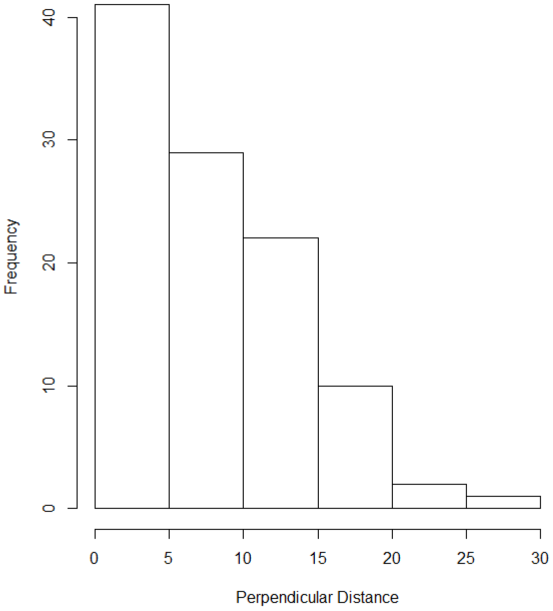

Figure 2: Histogram of perpendicular distances of detected animals

The real hard work begins when we start to ask ourselves how we can estimate this value of average detection probability, which turns out not to be so simple. However, distance sampling comes to the rescue! In line transect distance sampling, one or multiple observers travel along a transect line which has been pre-determined by the experimental design. Observers look to the left and right of the line as they travel along its path (either on foot, by plane, boat, car etc) and record the perpendicular distance at which each observed animal was detected. Once we pool together all these perpendicular distances for all detected animals, we expect to see a histogram similar to figure 2, which clearly displays that the frequency of detection drops as perpendicular distance increases. This trend has a rather obvious explanation, owing to the fact that animals become harder and harder to see the farther away they find themselves from the observer(s).

A model-based approach which involves the fitting of a detection function, allows us to estimate the way in which distance affects detection probability. Figure 3 illustrates this by overlaying a (scaled) half-normal detection function over simulated detection data. Over the course of an analysis, multiple detection functions will be fitted to the data and the most appropriate one selected. In the case of the Distance package written and maintained by the Centre for Research into Ecological and Environmental Modelling, Saint Andrews, this is done by the numerical computation of maximum likelihood estimates of the detection function parameters, and model selection is determined by goodness-of-fit tests as well as information criteria such as Akaike's Information Criterion.

If we momentarily return to the case of perfect detection, we notice that we would expect to see a perfectly flat histogram of perpendicular detection distances (as a result of uniform distribution of animals about the transect line). In this case the bars of our histogram would perfectly cover both the shaded, and un-shaded areas of Figure 3. Hence, the ratio of our detections with respect to this 'perfect detection rectangle' gives us some indication of the average detection probability of the population of interest. In practice, the 'perfect area rectangle' is extremely easy to calculate, and the area covered by our own detections is seen as the integral of the detection function over an interval from 0 to a chosen perpendicular truncation distance w. In order to obtain the sampled area (a, in the previous example) in the case of a line transect distance sampling survey, we simply consider $a = 2Lw$, where L is the total length of all covered transect lines and w is the previously mentioned perpendicular truncation distance. Thus, with these tools at our disposal, we can modify (2) to obtain:

$$\hat{N} = \frac{n}{\frac{2Lw}{A}\hat{p}} \tag{3}$$

where $\hat{p}$ is the estimate of average detection probability we obtain by integrating the detection function and considering the area of the 'perfect detection rectangle'. It should be noted that this description of distance sampling is extremely rudimentary, and is far more intuitive than it is technical. For a deeper understanding of the likelihood functions involved, the design-based, and model-based aspects of the method, reference should be made to both classic and modern literature, such as (Borchers, Buckland, & Zucchini, 2002) and (Buckland, Rexstad, Marques, & Oedekoven, 2004).

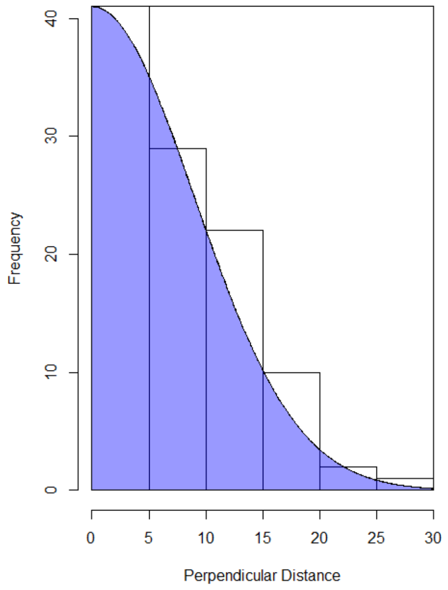

Figure 3: Histogram of detection distances with overlayed detection function

2.1.2 The significance of 2D detection functions

One of the assumptions of distance sampling is that the distribution of animals about the transect line is known. In the previous case of line-transects, it was uniform. Unfortunately, there are many situations in which this assumption is strongly violated, leading to significant bias in estimates of animal abundance and density. One reason for this could be poor survey design, whereby insufficient randomisation of transect line locations leads to trends in the distribution of animals between them. There is no effective analytical solution for poor survey design, and so this case is not considered here. A more interesting cause for this phenomenon is responsive movement of animals in the population of interest. Animals can be scared by observers and hence move away, causing a dip in observed detection frequencies on and close to the transect line, as well a spike in observed detection frequency some relatively small distance away from the transect line. This behaviour is typically referred to as avoidance, and is extremely common in primates, with gibbons being so notoriously prone to the behaviour that they are the subject of regular conventions between ecologists and statisticians. Conversely, animals may also display attraction, whereby an attraction to the transect line causes a spike in detections on the transect line and a dip further out. Dolphins typically display this behaviour when data is collected via shipboard surveys, where the wake of the vessel draws them in.

By considering time-to-detection models, and observing that when an observer is moving along the transect at a constant speed, the forward distance of a detected animal is directly proportional to the time to detection (assuming the animal is stationary), Borchers and Cox (2017) formulate a method by which the relationship between distance and detection probability can be estimated in two-dimensions. This extra information provided in the forward distances of detections also allows the estimation of the perpendicular distributions about the transect line of the detected animals. Once this distribution is known, the issues of responsive movement are largely remedied. It should be noted that the vast majority of studies which collect data which are suitable for the fitting of 2D detection functions often do so by measuring the radial distance to the animal, as well as the angle of incidence of the animal to the observer. Thus the values of perpendicular and forward difference are easily calculated using basic trigonometry. Many of the implementations and intricacies involved with handling edge cases arise directly from this convention and the ways it affects the data and their collection.

2.2 Previous contributions

From June - August 2017 I undertook work on the code-base as a summer student at the University of Saint Andrews. The notable changes I made to the code-base over this period will be clearly outlined in this section to distinguish them from the work which was undertaken from September 2017 onward as part of this project, however these will be not be described in depth, in order to avoid removing focus from the more relevant implementations. If more detail on the work conducted in this summer studentship is of interest, it can be found in the slides of a presentation conducted at the 2017 St Andrews Mathematics Undergraduate Research Conference, which have been included in the appendix.

2.2.1 Object Oriented Additions

The R programming language allows for packages to include their own classes to produce objects with behaviours and properties tailored to the task at hand.

These typically come in the form of S3 or S4 objects. As part of my summer work I wrote methods for S3 generic functions which allow the LT2D package to be more streamlined and provide the user with a cleaner and more straight forward interface with the functionality provided under the hood. The first set of these methods are dedicated to the R plot function, allowing the user to plot LT2D fitted model objects with a simple call to the generic plot function, with the aforementioned model object as the first argument.

2.2.2 Adding in Abundance Estimation

Previous versions of the code-base only produced estimates for average detection probability, $\hat{p}$. Code supplied by Prof. Borchers was modified and integrated with existing LT2D code to allow the package to produce estimates of abundance as well as average detection probability and effective strip half width. This work was done towards the end of the project and hence was not fully integrated and tested with all other pieces of functionality available within the software, therefore modifications of abundance estimation code and its continual testing and iterative improvement is necessary for the completion of this project, however the details of these updates are of far less interest than the more ambitious aspects of the project and so they shall not be mentioned further in any significant amount of detail. Any addition or modification of code or theoretical technicalities related to the LT2D package mentioned out-with of section 1.2 are not a result of previous work and form the improvements and extensions which the project aims to produce.

2.2.3 Rounded Angles

When measurements of small detection angles are rounded down to 0 in the field, the magnitude of the bias of estimated average detection probability, $\hat{p}$, can drastically increase. An appropriate modification to the method's likelihood function was theorised by prof. Borchers and then I implemented the new functionality into the source code. I conducted multiple simulations to confirm that bias was significantly reduced, all of which clearly demonstrated the new estimator performed far better.

2.2.4 New Perpendicular Density

A reparameterisation of an existing perpendicular density was required to allow for a half-normal perpendicular density function with a non-zero expected frequency of animals on the transect line. Prof. Borchers and I decided on a suitable formulation and I implemented it within the LT2D package.

2.2.5 Concatenation of GoF Functions

Goodness of fit tests (Cramer Von Mises and Kolmogorov-Smirnov) are conducted separately in the x and y (perpendicular and forward) directions, respectively, by the LT2D. These were split into distinct functions, making their use tedious to the end user. I rectified this issue by creating a wrapper function which was easier for the user to interact with and acted as an interface with the pre-existing code.

2.2.6 Model Collecting and Summarising

As with many model-based statistical methodologies, it is often good practice to fit a range of varied models to the available data, and to use selection criteria (such as Akaike's Information Criterion) to select which models are most reasonable. This leads to the user's R global environment being peppered with objects created by the LT2D package, making it difficult to keep track of the models of interest. Inspired by another piece of software used for animal abundance (RMARK - An interface to the popular MARK windows software used as a tool for analysis of mark-recapture data) - model-collecting and comparison-table-generating functions were implemented within the LT2D package.

2.3 Project objectives and outline

The project aims to include the ability for users to specify covariate inclusion for the two-dimensional detection functions, making them not only a function of perpendicular and forward distance, but also of user-defined relationships between other variables. Upon discussion with Prof. Borchers, it was decided this functionality should be kept as general as possible, following the following main criteria:

1. Users should be able to choose whether the covariate affects the behaviour of the detection function in the perpendicular direction, the forward direction or at the intercept, or any combination of the above.

2. Users should be able to define the formula which describes the linear predictor for the relevant parameter and the covariate(s) of interest themselves, rather than be restricted to predefined options

3. Users should be able to specify different formulas for each parameter (forward, perpendicular and intercept) wherever the detection function allows these to be different.

4. There should be some level of inbuilt error-checking to ensure that users choices for formulas and starting parameter values are consistent between themselves and their choice of detection function

All code calculating estimates of abundance ($\hat{N}$), average detection probability ($\hat{p}$) and AIC should be modified to work with all reasonable choices of covariates and related linear predictors. S3 methods should be written for the plot function to allow it to handle fitted models which contain cases of both discrete and continuous covariate inclusion. The goodness of fit function should also be generalised to use the Kolmogorov-Smirnov and Cramer Von Mises tests when the fitted model has included covariates. The inclusion of code which produces (non-parametric) bootstrap confidence intervals of estimated abundance would also be desirable, for both covariate-including and covariate-excluding models.

3 Implementation

3.1 Covariates; why include them, and how?

There are many reasons for the desirability of covariate inclusion within the LT2D package. Firstly, when dealing with animal populations which are heterogeneous with respect to detection probability (that is, individuals in the population of interest do not all share a common detection function due to their varying behaviours or characteristics), failing to take this into account often leads to biased estimates of abundance and can even lead to an inability to distinguish amongst reasonable models (Link, 2003), this can be especially prominent in Mark-Recapture analyses, which are the subject of the aforementioned citation. Although distance sampling benefits from the property of pooling robustness, and hence can cope with reasonable levels of heterogeneity within a population (Buckland et al., 2015), 2D detection sampling is based more fundamentally on time-to-detection models and survival analysis, meaning there is no theoretical guarantee that it also shares this robustness with respect to heterogeneous populations. Secondly, we may be interested in performing an analysis on data for which we have associated environmental factors of particular interest. By fitting various models which include different combinations of these covariates, analysing these models' estimated parameters, and by comparing them by standard means of model selection and goodness-of-fit, we may gain insight into the relationship between these covariates and the detection probabilities of individuals within the survey population. Finally, we hope to be able to reduce variance in estimates of abundance and density, by providing more information with which to assess detection probability than the coordinates of detected individuals alone.

We examine the detection functions outlined in (Borchers & Cox, 2017) and available as part of the associated code-base in order to gain an appreciation for their general form and to ascertain which parameters are most appropriate for covariate inclusion. It should be noted that these detection functions are formulated as hazards of detection, given the location of an animal. Therefore, the terms 'detection function' and 'hazard function' will be used interchangeably henceforth, to allow us to remain consistent with the previous code-base of the LT2D package. Although other, less used hazard functions are present within this code-base, they shall not be examined nor considered for covariate implementation.

$$h_1(y, x; \boldsymbol{\theta}) = \theta_1 \left(y^2 + x^2\right)^{-\frac{\theta_2}{2}} \tag{4}$$

$$ip_1(y, x; \boldsymbol{\theta}) = \theta_1 \left\{ \frac{1}{\sqrt{1 + \left(\frac{x}{\theta_2}\right)^2 + \left(\frac{y}{\theta_2}\right)^2}} \right\}^{\theta_3 + 1} \tag{5}$$

$$ip_2(y, x; \boldsymbol{\theta}) = \theta_1 \left\{ \frac{1}{\sqrt{1 + \left(\frac{x}{\theta_2}\right)^2 + \left(\frac{y}{\theta_4}\right)^2}} \right\}^{\theta_3 + 1} \tag{6}$$

$$ep_1(y, x; \boldsymbol{\theta}) = \theta_1 \left\{ \left(\frac{x}{\theta_2}\right)^{\theta_3} + \left(\frac{y}{\theta_2}\right)^{\theta_3} \right\} \tag{7}$$

$$ep_2\left(y, x; \boldsymbol{\theta}\right) = \theta_1 \left\{ \left(\frac{x}{\theta_2}\right)^{\theta_3} + \left(\frac{y}{\theta_4}\right)^{\theta_3} \right\} \tag{8}$$

It should be noted that we introduce a log link function to all θ parameters, in order to allow numerical optimisation routines tasked with finding the maximum likelihood estimates of these parameters the possibility to perform their search anywhere on the real axis. This is with the exception of ep_1 and ep_2, which use a logit link function for their θ_1 parameters instead.

We notice that all of the above hazard functions share a similar parameter structure. Firstly, they all contain a scale parameter which multiplies the whole function. This is θ_1, which can be interpreted as an 'intercept' for detection probability, given an increase in its value will increase the detection function's value for a given x and y, and a reduction in its value will cause the opposite. Secondly, all hazard functions contain parameters which affect the behaviour of the function in the x and y directions. Sometimes the behaviour in both these directions is determined by a single parameter (such as h_1, ip_1 and ep_1, which have both their x and y direction dependent on θ_2), however, in the cases of ip_2 and ep_2, the inclusion of a θ_4 parameter allows the behaviour in the y direction to be different to that in the x direction.

In the case of θ_1, modifying this parameter to include covariates would enable us to allow for heterogeneity which we believe affects an individual's detection probability, but not the shape of its detection function. In the case of θ_2, covariate inclusion would cause changes in detection function shape in either the x direction alone, or in the x and y directions in the same way, depending on the user's selection of hazard within their model criteria. Any change in θ_4 would cause a change in detection function shape in the y direction alone. We propose the following reformulations of θ_1, θ_2 and θ_4 which allow them to be modelled as the combination of an intercept and a user-specified formula:

$$\mathcal{G}_m\left(\theta_1\right) = \lambda_1 + \mathcal{C}_i\left(\boldsymbol{\gamma}_i\right) \tag{9}$$

$$ln\left(\theta_2\right) = \lambda_2 + \mathcal{C}_x\left(\boldsymbol{\gamma}_x\right) \tag{10}$$

$$ln\left(\theta_4\right) = \lambda_4 + \mathcal{C}_y\left(\boldsymbol{\gamma}_y\right) \tag{11}$$

Where $\mathcal{G}_m$ is the link function appropriate to θ_1 given the detection function given by model specification m the user has selected, λ_1, λ_2 and λ_4 are intercept parameters for the θ parameter with the corresponding subscript, $\mathcal{C}_i$ is the user specified formula for covariate inclusion in the detection function's intercept, $\mathcal{C}_x$ is the user specified formula for covariate inclusion in the x direction, and $\mathcal{C}_y$ is the user specified formula for covariate inclusion in the y direction. The $\boldsymbol{\gamma}$ are vectors of parameters of the appropriate length required for the corresponding covariate relationship, $\mathcal{C}$. We must have $\mathcal{C}_x = \mathcal{C}_y$ if the detection function specified within model m is h_1, ip_1 or $ep1$, since θ_2 will affect the detection function's shape in both the x and y directions.

Before considering an example model specification, we introduce the four perpendicular density functions within the LT2D package which allow the methodology to account for (and describe) the nature of the observed responsive movement:

$$\pi_U\left(x; \boldsymbol{\phi}\right) = e^{-\frac{x^2}{2\phi_1^2}} \left\{ \int_0^w e^{-\frac{x^2}{2\phi_1^2}}\,\mathrm{d}x \right\}^{-1} \tag{12}$$

$$\pi_N\left(x; \boldsymbol{\phi}\right) = e^{-\frac{(x-\phi_2)^2}{2\phi_1^2}} \left\{ \int_0^w e^{-\frac{(x-\phi_2)^2}{2\phi_1^2}}\mathrm{d}x \right\}^{-1} \tag{13}$$

$$\pi_{CN}\left(x; \boldsymbol{\phi}\right) = \left\{ 1 - e^{-\frac{x^2}{2\phi_1^2}} \right\} \tag{14}$$

We consider an example model m specified by a user of the LT2D package, which selects an ep_2 detection function. Data collected by the user includes all of the standard information required to perform an estimate of abundance through conventional distance sampling (CDS), as well as the forward distance of each detection, a factor covariate *species*, a continuous covariate $forest.cover$ and another factor covariate $size$, which for each detection refers to the number of individuals in the detected group. The user has specified that the i direction (the intercept) should include the relationship $i \sim forest.cover$

It becomes clear that allowing the user a selection of choices for covariate inclusion quickly becomes an issue of not only statistical theory, but also a problem rooted in software design. We must construct a general methodology which allows users to specify

3.2 Understanding the LT2D software and its architecture

In order to implement covariate inclusion it is vital to first understand the inner workings of the LT2D package, so that we may ascertain both the complexity of the task at hand, and also the most efficient method with which to proceed. A simplification of the software's top-level fitting function call hierarchy (as of August 2017, before the project began), which retains the features of greatest importance and relevance is illustrated by figure 4.

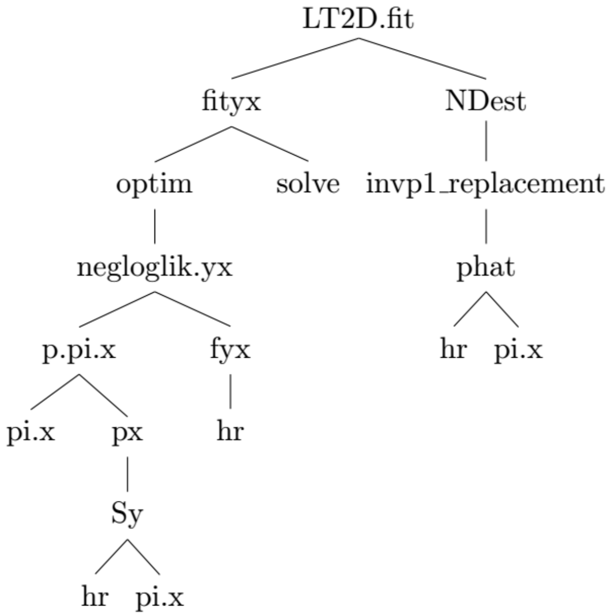

Figure 4: Simplified call hierarchy for the top-level fitting function in the LT2D software

This diagram illustrates in more detail the inner workings of the LT2D package, compared to the specifications of the methodology laid out by (Borchers & Cox, 2017). Below a brief outline and description of each of the functions in figure 4 will be given, to give the reader a general understanding of how the LT2D package used to fit models to data before covariate implementation, and how it evaluates the likelihood function outlined in (Borchers & Cox, 2017).

- LT2D.fit ; A function which given an appropriate input R data.frame and model specification settings, produces a two-dimensional detection function distance sampling maximum likelihood estimation of model parameters and abundance.

- fityx ; The function responsible for the model-fitting provided by LT2D.fit.

- NDest ; The function tasked with calculating estimates of abundance and density, when provided with the output of the fityx function.

- optim ; The linear optimisation routine which provides maximum likelihood estimation of the model parameters.

- solve ; The function which solves the hessian matrix provided by optim in order to produce a variance-covariance matrix for the estimated parameters.

- invp1_replacement ; The function which produces an estimate of inverse effective half strip width, given the chosen model settings and their estimated parameters.

- negloglik.yx ; The function which evaluates the negative log likelihood - the probability of observing the observations we saw, given the model specifications and their estimated parameters.

- phat ; The function tasked with calculating effective half strip width, given model specifications and their estimated parameters.

- p.pi.x ; The function which evaluates the product of the probability of detection at a given perpendicular distance and the probability of an animal's presence at that perpendicular distance, given model specifications and their estimated parameters.

- fyx ; Calculates the probability density function of the 'waiting distance', given model specifications and their estimated parameters.

- hr ; Evaluates the two-dimensional detection function specified by the user as part of the model specification. Returns the hazard of detection given estimated parameters.

- pi.x ; Evaluates the assumed perpendicular density distribution of animals about the transect line, selected by the user. Returns the probability of the animal's location at a given perpendicular distance given estimated parameters.

- px ; Numerically calculates the perpendicular detection function from a specification of a hazard function (in the form of hr, above) and its estimated parameters.

- Sy ; Calculates the survivor function to a given forward distance. Evaluates the probability of an animal remaining undetected until a given forward distance ahead of the observer, and then being detected at that distance, given model specifications and their estimated parameters.

3.3 Covariate Inclusion into the Likelihood

3.4 Covariate Inclusion into Abundance Estimation

3.5 Modification of Goodness of Fit Functions

3.6 Modification of Plot Functions

4 Usage of the new software

4.1 Model specification for the user

4.2 Simulation studies

5 Going forwards

5.1 Additional desired features

5.2 A proposed approach for mixture model incorporation

References

Borchers, D. L., Buckland, S. T., & Zucchini, W. (2002). *Estimating animal abundance.* Springer, London.

Borchers, D. L., & Cox, M. J. (2017). Distance sampling detection functions: 2d or not 2d? *Biometrics*, *73*(2), 593-602. doi: 10.1111/biom.12581

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (2015). *Advanced distance sampling.* Oxford University Press, Oxford.

Buckland, S. T., Rexstad, E. A., Marques, T. A., & Oedekoven, C. S. (2004). *Distance sampling.* Springer.

Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, *59*(4), 1123-1130. doi: 10.1111/j.0006-341X.2003.00129.x