bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Distance Sampling in R

David L Miller
Integrated Statistics and University of St Andrews

Eric Rexstad
University of St Andrews

Len Thomas
University of St Andrews

Laura Marshall
University of St Andrews

Jeffrey L Laake
Marine Mammal Laboratory

Abstract

Estimating the abundance and spatial distribution of animal and plant populations is essential for conservation and management. We introduce the R package **Distance** that implements distance sampling methods to estimate abundance. We describe how users can obtain estimates of abundance (and density) using the package as well documenting the links it provides with other more specialized R packages. We also demonstrate how **Distance** provides a migration pathway from previous software, thereby allowing us to deliver cutting-edge methods to the users more quickly.

Keywords: distance sampling, abundance estimation, line transect, point transect, detection function, Horvitz-Thompson, R, Distance.

1. Introduction

Distance sampling (Buckland, Anderson, Burnham, Borchers, and Thomas 2001; Buckland, Anderson, Burnham, Laake, Borchers, and Thomas 2004; Buckland, Rexstad, Marques, and Oedekoven 2015) encompasses a suite of methods used to estimate the density and/or abundance of biological populations. Distance sampling can be thought of as an extension of plot sampling. Plot sampling involves selecting a number of plots (small areas) at random within the study area and counting the objects of interest that are contained within each plot. By selecting the plots at random we can assume that the density of objects in the plots is representative of the study area as a whole. One of the key assumptions of plot sampling is that all objects within each of the plots are counted. Distance sampling relaxes this assumption in that observers are no longer required to detect (either by eye, video/audio recording etc) and count everything within selected plots. While plot sampling techniques are adequate for static populations occurring at high density they are inefficient for more sparsely distributed populations. Distance sampling provides a more efficient solution in such circumstances.

Conventional distance sampling assumes the observer is located either at a point or moving along a line and will observe all objects that occur at the point or on the line. The further

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

away an object is from the point or line (also known as the sampler or transect) the less likely it is that the observer will see it. We can use the distances to each of the detected objects from the line or point to build a model of the probability of detection given distance from the sampler — the *detection function*. The detection function can be used to infer how many objects were missed and thereby produce estimates of density and/or abundance.

The Windows program Distance (or "DISTANCE"; for clarity henceforth "Distance for Windows") can be used to fit detection functions to distance sampling data. It was first released (versions 1.0 - 3.0; principally programmed by Jeff Laake while working at the National Marine Mammal Laboratory) as a console-based application (this in turn was based on earlier software TRANSECT, Burnham, Anderson, and Laake 1980 and algorithms developed in Buckland 1992), before the first graphical interface (Distance for Windows 3.5) was released in November 1998. Since this time it has evolved to include various design and analysis features (Thomas, Buckland, Rexstad, Laake, Strindberg, Hedley, Bishop, Marques, and Burnham 2010). Distance for Windows versions 5 onwards have included R (R Core Team 2015) packages as the analysis engines providing additional, more complex analysis options than those offered by the original (Fortran) code.

As Distance for Windows becomes increasingly reliant on analyses performed in R and many new methods are being developed, we are encouraging the use of our R packages directly. R provides a huge variety of functionality for data exploration and reproducible research, much more than is possible in Distance for Windows.

Until now those wishing to use our R packages for straight forward distance sampling analyses would have had to negotiate the package **mrds** (Laake, Borchers, Thomas, Miller, and Bishop 2015) designed for mark-recapture distance sampling (Burt, Borchers, Jenkins, and Marques 2014), requiring a complex data structure to perform analyses. **Distance** is a wrapper package around **mrds** making it easier to get started with basic distance sampling analyses in R. The most basic detection function estimation only requires a numeric vector of distances. Here we demonstrate how to use **Distance** to fit detection functions, perform model checking and selection, and estimate abundance.

1.1. Distance sampling

The distribution of the observed distances is a product of detectability (sometimes referred to as "perception bias"; Marsh and Sinclair 1989) and the distribution of the animals with respect to the line or point. Our survey design allows us to assume a distribution for the animals with respect to the sampler.

For line transect studies we assume that objects are uniformly distributed with respect to the line (i.e., the number of animals available for detection is the same at all distances). For point transect surveys, area increases linearly with radial distance, implying a triangular distribution with respect to the point. Figure 1 shows how these distributions, when combined with a detection function, give rise to the observed distribution of recorded distances.

Figure 2 shows simulated sampling of a population of 500 objects using line and point transects and their corresponding histograms of observed detection distances. Note that for the purposes of distance sampling an "object" may either refer to an individual in a population or a cluster of individuals.

Good survey design is essential to produce reliable density and abundance estimates from distance sampling data. Survey design is beyond the scope of this article but we refer readers

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

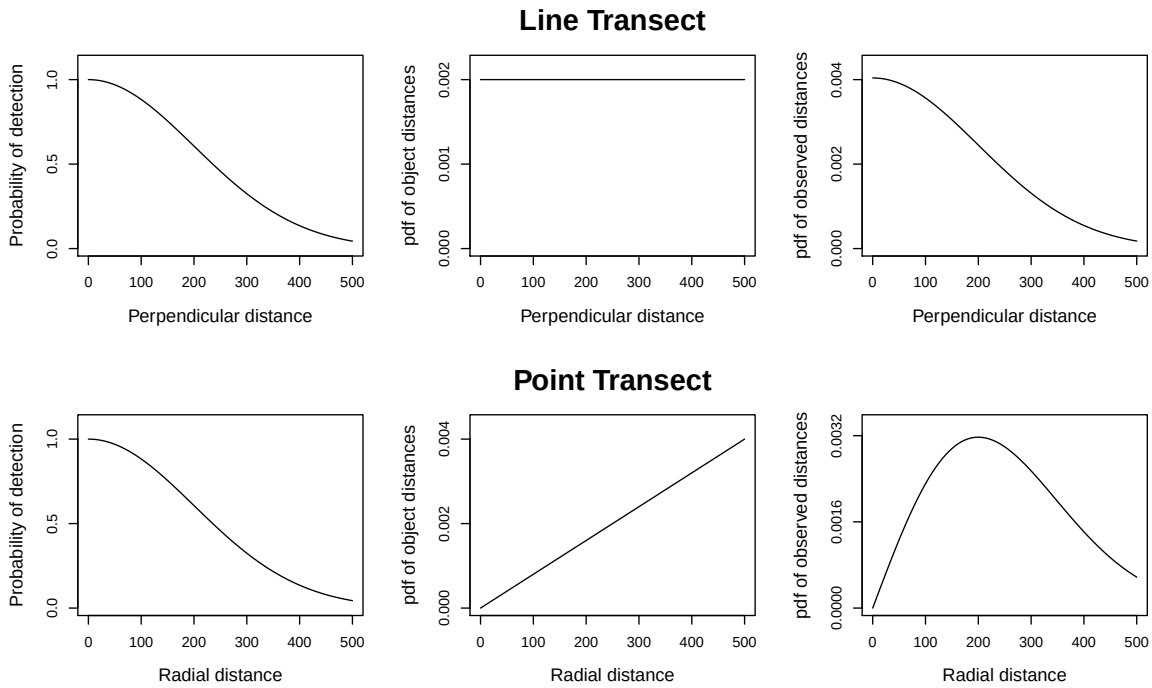

Figure 1: Panels show an example detection function (left), the pdf of object distances (middle) and the resulting pdf of detection distances (right) for line transects (top row) and point transects (bottom row). The pdf of observed detection distances in the right hand plots are obtained by multiplying the detection function by the pdf of object distances and rescaling. In this example, detection probability becomes effectively zero at 500 (distances shown on x-axis are arbitrary).

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

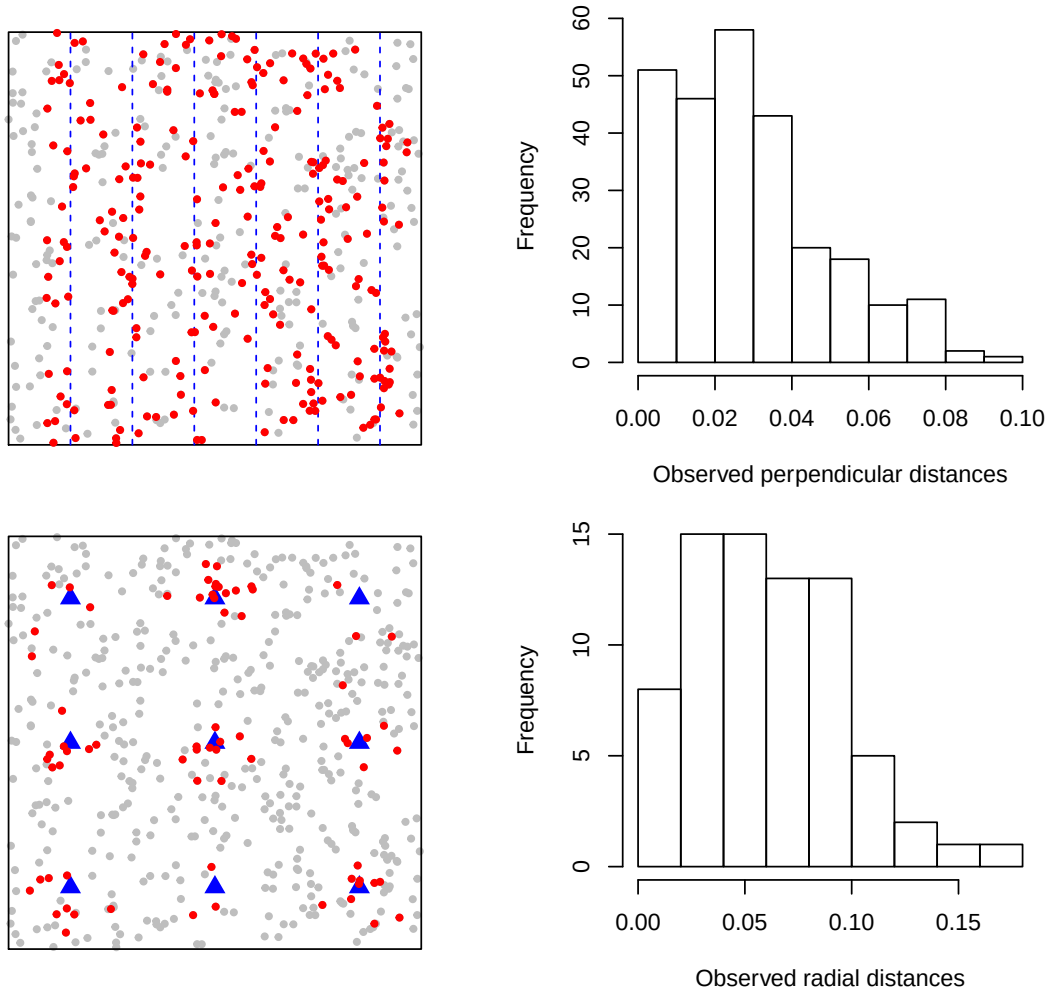

Figure 2: Left side plots show an example of a survey of a (1x1) area containing a population of 500 objects; blue dashed lines (top plot) and triangles (bottom plot) indicate sampler placement, red dots indicate detected individuals and grey dots give the locations of unobserved individuals. The right side of the figure shows histograms of observed distances (again, lines top and points bottom).

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

to (Buckland *et al.* 2001, Chapter 7) and (Buckland *et al.* 2015, Chapter 2) for introductory information; Strindberg, Buckland, and Thomas (2004) contains information on automated survey design using geographical information systems (GIS); Thomas, Williams, and Sandilands (2007) gives an example of designing a distance sampling survey in a complex region. We also note that Distance for Windows includes a GIS survey design tool (Thomas *et al.* 2010).

Distance provides a selection of candidate functions to describe the probability of detection and estimates the associated parameters using maximum likelihood estimation. The probability of detecting an object may not only depend on how far it is from the observer but also on other factors such as weather conditions, ground cover, cluster size etc. The **Distance** package also allows the incorporation of such covariates into the detection function allowing the detection function scale parameter to vary based on these covariates.

Having estimated the detection function's parameters, one can then integrate out distance from the function (as the detection function describes the probability of detection *given* distance) to get an "average" probability of detection, which can be used to correct the observed counts. Summing the corrected counts gives an estimate of abundance in the area covered by surveys, which can be multiplied up to the total study area.

In addition to randomly placed samplers distance sampling relies on three other main assumptions. Firstly, all objects directly on the transect line or point (i.e., those at zero distance) are detected (see "Extensions" for methods to deal with the situation when this is not possible). Secondly, objects are stationary or detected prior to any movement. Thirdly, distance to the object must be measured accurately, or the observation allocated to the correct distance bin for grouped data. Depending on the survey species some of these assumptions may be more difficult to meet than others. Further information on field methods to help meet these assumptions can be found in Buckland *et al.* (2001) and Buckland *et al.* (2015).

The rest of the paper has the following structure: we describe data formatting for **Distance**; candidate detection function models are described in terms of formulation and fitted examples in R. We then show how to perform model checking, goodness of fit and model selection. We go on to show how to estimate abundance, including stratified estimates of abundance. The final two sections of the article look at extensions (both in terms of methodology and software) and put the package in a broader context amongst other R packages used for estimating the abundance of biological populations from distance sampling data.

2. Data

We introduce two example analyses performed in **Distance**: one line transect and one point transect. These data sets have been chosen as they represent typical data seen in practice. The below example analyses are not intended to serve as guidelines, but to demonstrate features of the software. Practical advice on approaches to analysis is given in Thomas *et al.* (2010).

2.1. Minke whales

The line transect data have been simulated from models fitted to Antarctic minke whale (*Balaenoptera bonaerensis*) data. These data were collected as part of the International Whaling Commission's International Decade of Cetacean Research Southern Ocean Whale and Ecosys-

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tem Research (IWC IDCR-SOWER) programme 1992-1993 austral summer surveys. These data consist of 99 observations that are stratified based on location (near or distant from ice edge) and effort data (transect lengths). Further details on the survey are available in Branch and Butterworth (2001) (data is simulated based on the design used for "1992/93 Area III" therein).

2.2. Amakihi

The point transect data set consists of 1485 observations of Amakihi (*Hemignathus virens*; a Hawaiian songbird), collected at 41 points between 1992 and 1995. The data include distances and two covariates collected during the survey: observer (a three level factor), time after sunrise (transformed to minutes (continuous) or hours (factor) covariates). Data are analysed comprehensively in Marques, Thomas, Fancy, and Buckland (2007).

2.3. Data setup

Generally, data collected in the field will require some formatting before use with **Distance**, though there are a range of possible formats, dependent on the model specification and the output required:

- In the simplest case, where the objective is to estimate a detection function and exact distances are collected, all that is required is a numeric vector of distances.

- To include additional covariates into the detection function (see "Detection functions") a `data.frame` is required. Each row of the `data.frame` contains the data on one observation. The `data.frame` must contain a column named `distance` (containing the observed distances) and additional named columns for any covariates that may affect detectability (for example `observer` or `seastate`). The column name `size` is reserved for the cluster sizes (sometimes referred to as group sizes) in this case each row represents an observation of a cluster rather than individual (see Buckland *et al.* 2001, Section 1.4.3 for more on defining clusters and "Extensions" below for one approach to dealing with uncertain cluster size). Additional reserved names include `object` and `detected`, these are not required for conventional distance sampling and should be avoided (see "Extensions" for an explanation of their use).

- To estimate density or to estimate abundance beyond the sampled area, additional information is required. Additional columns should be included in the `data.frame` specifying: `Sample.Label`, the ID of the transect; `Effort`, transect effort (for lines their length and for points the number of times that point was visited); `Region.Label`, the stratum containing the transect (which may be from pre- or post-survey stratification, see "Estimating abundance and variance"); `Area`, the area of the strata. Transects which were surveyed but have no observations must be included in the data set with `NA` for `distance` and any other covariates. We refer to this data format (where all information is contained in one table) as "flatfile" as it can be easily created in a single spreadsheet.

As we will see in "Extensions", further information is also required for fitting more complex models.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

It is also possible to use distances collected in intervals ("binned" or "grouped" data, as opposed to "exact" distances) to model the detection function. In this case the column `distance` is replaced by two columns `distbegin` and `distend` referring to the distance interval start and end cutpoints. More information on binned data is included in Buckland *et al.* (2001) sections 4.5 and 7.4.1.2.

The columns `distance`, `Area` and (in the case of line transects) `Effort` have associated units (though these are not explicitly included in a **Distance** analysis). We recommend that data in these columns are converted to SI units before starting any analysis to ensure that resulting abundance and density estimates have sensible units. For example, if distances from a line transect survey are recorded in metres, the `Effort` columns should contain line lengths also in metres and the `Area` column gives the stratum areas in square metres. This would lead to density estimates of animals per square metre.

The minke whale data follows the "flatfile" format given in the last bullet point:

```
R> library("Distance")
R> data("minke")
R> head(minke)

  Region.Label  Area Sample.Label Effort distance
1        South 84734            1  86.75     0.10
2        South 84734            1  86.75     0.22
3        South 84734            1  86.75     0.16
4        South 84734            1  86.75     0.78
5        South 84734            1  86.75     0.21
6        South 84734            1  86.75     0.95
```

Whereas the amakihi data lacks effort and stratum data:

```
R> data("amakihi")
R> head(amakihi)

   survey object distance obs mas has detected
1 July 92      1       40 TJS  50   1        1
2 July 92      2       60 TJS  50   1        1
3 July 92      3       45 TJS  50   1        1
4 July 92      4      100 TJS  50   1        1
5 July 92      5      125 TJS  50   1        1
6 July 92      6      120 TJS  50   1        1
```

We will explore the consequences of including effort and stratum data in the analysis below.

3. Detection functions

The detection function models the probability $\mathbb{P}(\text{object detected} \mid \text{object at distance } y)$ and is usually denoted $g(y; \boldsymbol{\theta})$ where y is distance (from a line or point) and $\boldsymbol{\theta}$ is a vector of parameters to be estimated. Our goal is to estimate an *average probability of detection* (p,

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

average in the sense of an average over distance from 0 to truncation distance w), so we must integrate out distance (y) from the detection function:

$$p = \int_0^w \pi(y) g(y; \boldsymbol{\theta}) dy$$

where $\pi(y)$ describes the distribution of objects with respect to the sampler; $\pi(x) = 1/w$ for line transects and $\pi(r) = \frac{2r}{w^2}$ for point transects, taking into account the geometry of the sampler (usually referred to as the *probability density function of distances* Buckland *et al.* 2001, Chapter 3), . When considering a particular transect type we let x denote a perpendicular distance from a line and r denote radial distance from a point (rather than using y).

It is crucial that the detection function accurately models detectability at small distances; we are less worried by its behaviour further away from 0. To ensure that the model is not overly influenced by distances far from zero, we truncate the distances beyond a given distance w (this can be done during analysis or while collecting data in the field).

Models for the detection function are expected to have the following properties (Buckland *et al.* 2015, Chapter 5):

- *Shoulder*: we expect observers to be able to see objects near them, not just those directly in front of them. For this reason, we expect the detection function to be flat near the line or point.
- *Non-increasing*: we do not think that observers should be more likely to see distant objects than those nearer the transect. If this occurs, it usually indicates an issue with survey design or field procedures (for example that the distribution of objects with respect to the line, $\pi(y)$ is not what we expect), so we do not want the detection function to model this (Marques, Buckland, Tosh, McDonald, and Borchers 2010; Marques, Buckland, Bispo, and Howland 2012; Miller and Thomas 2015).
- *Model robust*: models should be flexible enough to fit many different shapes.
- *Pooling robust*: many factors can affect the probability of detection and it is not possible to measure all of these. We would like models to produce unbiased results without inclusion of these factors.
- *Estimator efficiency*: we would like models to have low variances, given they satisfy the other properties above (which, if satisfied, would give low bias).

Given these criteria, we can formulate models for g.

3.1. Formulations

There is a wide literature on possible formulations for the detection function (Buckland 1992; Eidous 2005; Becker and Quang 2009; Giammarino and Quatto 2014; Miller and Thomas 2015; Becker and Christ 2015). `Distance` includes the most popular of these models. Here we detail the most popular detection function approach: "key function plus adjustments" (K+A).

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

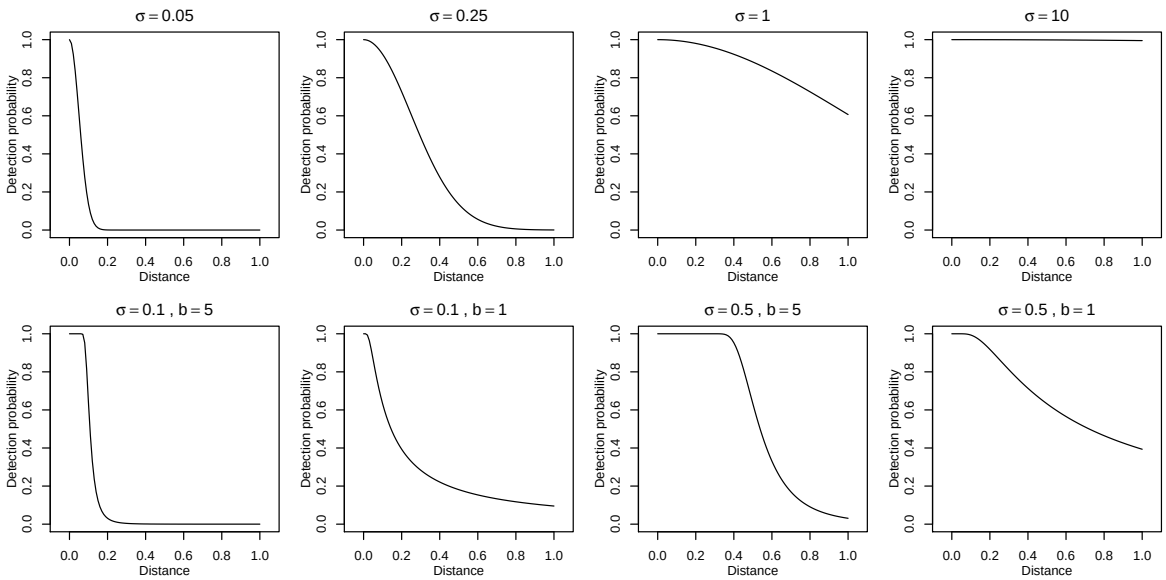

Figure 3: Half-normal (top row) and hazard-rate (bottom row) detection functions without adjustments, varying scale (σ) and (for hazard-rate) shape (b) parameters (values are given above the plots). On the top row from left to right, the study species becomes more detectable (higher probability of detection at larger distances). The bottom row shows the hazard-rate model's more pronounced shoulder.

Key function plus adjustments (K+A)

Key function plus adjustment terms (or adjustment series) models are formulated by taking a "key" function and optionally adding "adjustments" to it to improve the fit (Buckland 1992). Mathematically we formulate this as:

$$g(y;\boldsymbol{\theta}) = k(y;\boldsymbol{\theta}_{\text{key}})\left(1 + \alpha_O(y;\boldsymbol{\theta}_{\text{adjust}})\right),$$

where k is the key function and α_O is sum series of functions (given in Table 1), described as an *adjustment of order* O. Subscripts on the parameter vector indicate those parameters belonging to each part of the model (i.e., $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\text{key}}, \boldsymbol{\theta}_{\text{adjust}})$).

Available models for the key are as follows:

$$k(y) = \begin{cases} \exp\left(-\frac{y^2}{2\sigma^2}\right) & \text{half-normal,} \\ 1 - \exp\left(\left(-\frac{y}{\sigma}\right)^{-b}\right) & \text{hazard-rate,} \\ 1/w & \text{uniform.} \end{cases}$$

Possible modelling options for key and adjustments are given in Table 1 and illustrated in Figures 3 and 4. We select the number of adjustment terms (K) by AIC (further details in "Model checking and model selection").

When adjustment terms are used it is necessary to standardise the results to ensure that

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Key function	Form	Adjustment series	Form
Uniform	$1/w$	cosine	$\sum_{o=1}^{O} a_o \cos(o\pi y/w)$
		Simple polynomial	$\sum_{o=1}^{O} a_o (y/w)^{2o}$
Half-normal	$\exp\left(-\frac{y^2}{2\sigma^2}\right)$	cosine	$\sum_{o=2}^{O} a_o \cos(o\pi y/w)$
		Hermite polynomial	$\sum_{o=2}^{O} a_o H_{2o}(y/\sigma)$
Hazard-rate	$1-\exp\left[-\left(\frac{y}{\sigma}\right)^{-b}\right]$	cosine	$\sum_{o=2}^{O} a_o \cos(o\pi y/w)$
		Simple polynomial	$\sum_{o=2}^{O} a_o (y/w)^{2o}$

Table 1: Modelling options for key plus adjustment series models for the detection function. For each key function the default adjustments are cosine in **Distance**. Note that in the adjustments functions distance is divided by the width or the scale parameter to ensure the shape of adjustment independent of the units of y (Marques *et al.* 2007); defaults are shown here, though either can be selected to rescale the distances.

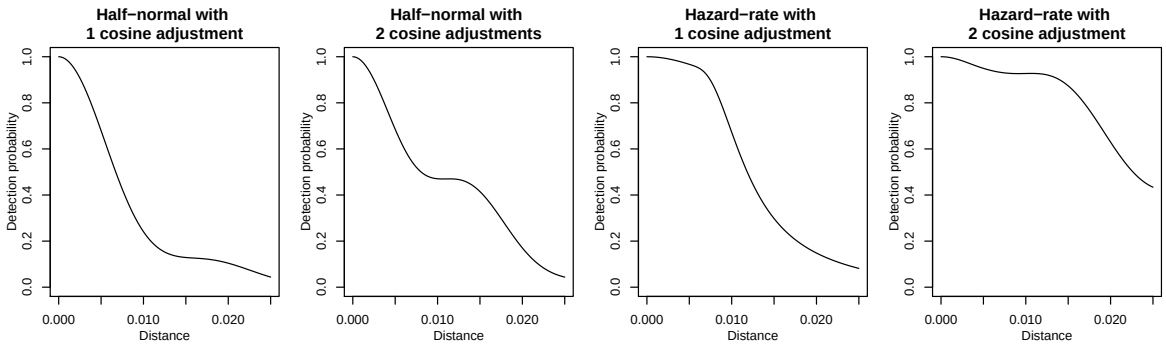

Figure 4: Possible shapes for the detection function when cosine adjustments are included for half-normal and hazard-rate models.

$g(0) = 1$, so we can redefine the detection function as:

$$g(y;\boldsymbol{\theta}) = \frac{k(y;\boldsymbol{\theta}_{\text{key}})\,(1+\alpha_O(y;\boldsymbol{\theta}_{\text{adjust}}))}{k(0;\boldsymbol{\theta}_{\text{key}})\,(1+\alpha_O(0;\boldsymbol{\theta}_{\text{adjust}}))}.$$

A disadvantage of K+A models is that we must resort to constrained optimisation (via the **Rsolnp** package; Ghalanos and Theussl 2014) to ensure that the resulting detection function is monotonic non-increasing over its range.

It is not always necessary to include adjustments (except in the case of the uniform key) and in such cases we refer to these as "key only" models (see the next section and "Model checking and model selection").

Covariates

There are many factors that can affect the probability of detecting an object: observer skill, cluster size (if objects occur in clusters), the vessel or platform used, sea state, other weather conditions, time of day and more. In **Distance** we assume that these variables affect detection

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

only via the scale of the detection function (and do not affect the shape).

Covariates can be included in this formulation by considering the scale parameter from the half-normal or hazard-rate detection functions as a(n exponentiated) linear model of the (J) covariates ($\mathbf{z}$; a vector of length J for each observation):

$$\sigma(\mathbf{z}) = \exp(\beta_0 + \sum_{j=1}^{J} \beta_j z_j).$$

Including covariates has an important implication for our calculation of detectability. We do not know the true distribution of the covariates, we therefore must either: (i) model the distribution of the covariates and integrate the covariates out of the joint density (thus making strong assumptions about their distribution), or (ii) calculate the probability of detection conditional on the observed values of the covariates (Marques and Buckland 2003). We opt for the latter:

$$p(\mathbf{z_i}) = \int_0^w \pi(y) g(y, \mathbf{z_i}; \boldsymbol{\theta}) dy,$$

where $\mathbf{z_i}$ is the vector of J covariates associated with observation i. For covariate models, we calculate a value of "average" probability of detection (average in the sense of distance being integrated out) per observation. There are as many unique values of $p(\mathbf{z_i})$ as there are unique covariate combinations in our data.

K+A models that include covariates and one or more adjustments cannot be guaranteed to be monotonic non-increasing for all covariate combinations. Without a model for the distribution of the covariates, it is not possible to know what the behaviour of the detection function will be across the ranges of the covariates. As such we cannot set meaningful constraints on monotonicity. For this reason, we advise caution when using both adjustments and covariates in a detection function (see Miller and Thomas 2015, for an example of when this can be problematic).

3.2. Fitting detection functions in R

A detection function can be fitted in **Distance** using the `ds` function. Here we apply some of the possible formulations for the detection function we have seen above to the minke whale and amakihi data.

Minke whale

First we fit a model to the minke whale data, setting the truncation at 1.5km and using the default options in `ds` very simply:

```
R> minke_hn <- ds(minke, truncation = 1.5)

Starting AIC adjustment term selection.

Fitting half-normal key function

Key only model: not constraining for monotonicity.
```

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


```
AIC= 46.872

Fitting half-normal key function with cosine(2) adjustments

AIC= 48.872


Half-normal key function selected.
```

Note that when there are no covariates in the model, `ds` will add adjustment terms to the model until there is no improvement in AIC.

Figure 5 (left panel) shows the result of calling `plot` on the resulting model object. We can also call `summary` on the model object to get summary information about the fitted model (we postpone this until the next section).

A different form for the detection function can be specified via the `key=` argument to `ds`. For example, a hazard rate model can be fitted as:

```
R> minke_hrcos <- ds(minke, truncation = 1.5, key = "hr")

Starting AIC adjustment term selection.

Fitting hazard-rate key function

Key only model: not constraining for monotonicity.

AIC= 48.637

Fitting hazard-rate key function with cosine(2) adjustments

AIC= 50.386


Hazard-rate key function selected.
```

Here `ds` also fits the hazard-rate model then hazard-rate with a cosine adjustment but the AIC improvement is insufficient to select the adjustment, so the hazard-rate key-only model is returned.

Other adjustment series can be selected using the `adjustment=` argument and specific orders of adjustments can be set using `order=`. For example, to specify a uniform model with cosine adjustments of order 1 and 2 we can write:

```
R> minke_unifcos <- ds(minke, truncation = 1.5, key = "unif",
+                      adjustment = "cos", order = c(1, 2))

Fitting uniform key function with cosine(1,2) adjustments
```

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


```
AIC= 48.268
```

Hermite polynomial adjustments use the code `"herm"` and simple polynomials `"poly"`, adjustment order should be in line with Table 1.

Amakihi

`ds` assumes the data given to it has been collected as line transects, but we can switch to point transects using the argument `transect="point"`. We can include covariates in the scale parameter via the `formula=~...` argument to `ds`. A hazard-rate model for the amakihi that includes observer as a covariate and a truncation distance of 82.5m (Marques *et al.* 2007) can be specified using :

```
R> amakihi_hr_obs <- ds(amakihi, truncation = 82.5, transect = "point",
+                       key = "hr", formula = ~obs)

Model contains covariate term(s): no adjustment terms will be included.

Fitting hazard-rate key function

AIC= 10778.448

No survey area information supplied, only estimating detection function.
```

Note that here, unlike with the minke whale data, `ds` warns us that we have only supplied enough information to estimate the detection function (not density or abundance).

While automatic AIC selection is performed on adjustment terms, model selection for covariates must be performed manually. Here we add a second covariate: minutes after sunrise. We will compare these two models further in the following section.

```
R> amakihi_hr_obs_mas <- ds(amakihi, truncation = 82.5, transect = "point",
+                           key = "hr", formula = ~obs + mas)

Model contains covariate term(s): no adjustment terms will be included.

Fitting hazard-rate key function

AIC= 10777.376

No survey area information supplied, only estimating detection function.
```

As with the minke whale model, we can plot the resulting models (Figure 5, middle and right panels). However, for point transect studies, probability density function plots give a better sense of model fit than the detection function plots. This is because when plotting the detection function for point transect data, the histogram must be rescaled to account for the geometry of the point sampler. The amakihi models included covariates, so the plots show the detection function averaged over levels/values of the covariate. Points on the plot indicate

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

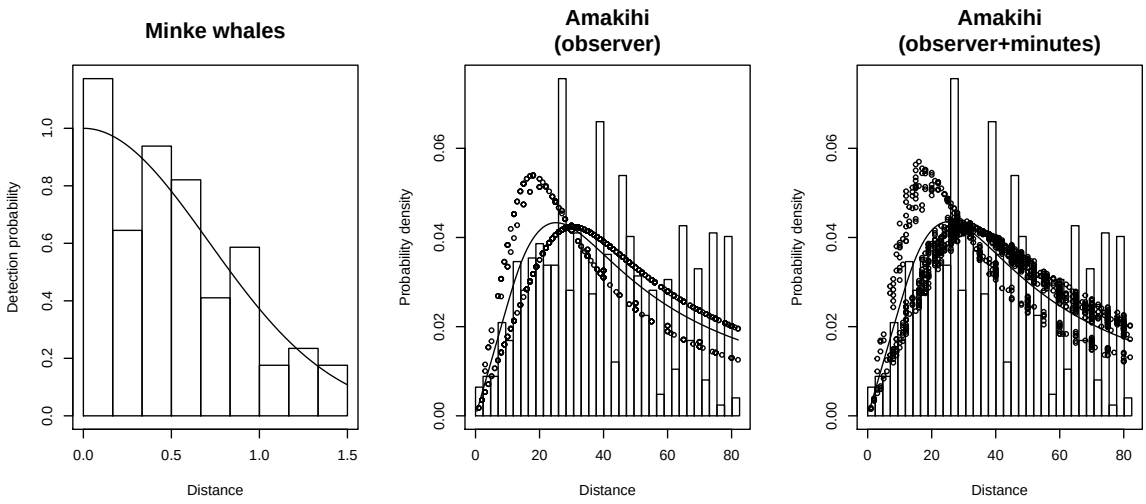

Figure 5: Left: fitted detection function overlayed on the histogram of observed distances for the minke whale data using half-normal model. Centre and right: plots of the probability density function for the amakihi models. Centre, hazard-rate with observer as a covariate; right, hazard-rate model with observer and minutes after sunrise as covariates. Points indicate probability of detection for a given observation (given that observation's covariate values) and lines indicate the average detection function.

probability of detection for each observation. For the `amakihi_hr_obs` model we see fairly clear levels of the observer covariate in the points. Looking at the right panel of Figure 5, this is less clear when adding minutes after sunrise (a continuous covariates) to the model.

4. Model checking and model selection

As with models fitted using `lm` or `glm` in R, we can use `summary` to give useful information about our fitted model. For our hazard-rate model of the amakihi data, with observer as a covariate:

```
R> summary(amakihi_hr_obs)


Summary for distance analysis
Number of observations :  1243
Distance range         :  0  -  82.5

Model : Hazard-rate key function
AIC   : 10778.45

Detection function parameters
Scale coefficient(s):
```

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


```
              estimate         se
(Intercept) 3.06441705 0.10878121
obsTJS      0.53017364 0.09956539
obsTKP      0.08885471 0.18071851

Shape coefficient(s):
             estimate         se
(Intercept) 0.8690009 0.06261764

                        Estimate          SE         CV
Average p              0.3142723   0.0204413 0.06504326
N in covered region 3955.1685709 274.2284029 0.06933419
```

This summary information includes details of the data and model specification, as well as the values of the coefficients (β_j) and their uncertainties, an "average" value for the detectability (see "Estimating abundance and variance" for details on how this is calculated) and its uncertainty. The final line gives an estimate of abundance in the area covered by the survey (see the next section; though note this estimate does not take into account cluster size).

4.1. Goodness of fit

We use a quantile-quantile plot (Q-Q plot) to visually assess how well a detection functions fits the data when we have exact distances. Here the Q-Q plot compares the cumulative distribution function (CDF) of the fitted detection function to the distribution of the data (empirical distribution function or EDF). The Q-Q plots in **Distance** plot a point for every observation. The EDF is the proportion of points that have been observed at a distance equal to or less than the distance of that observation. The CDF is calculated from the fitted detection function as the probability of observing an object at a distance less than or equal to that of the given observation. This can be interpreted as assessing whether the number of observations up to a given distance is in line with what the model says they should be. As usual for Q-Q plots, "good" models will have values close to the line $y = x$, poor models will show greater deviations from that line.

Q-Q plots can be inspected visually, though this is prone to subjective judgments. Therefore, we also quantify the Q-Q plot's information using a Kolmogorov-Smirnov or Cramér-von Mises test (Burnham, Buckland, Laake, Borchers, Bishop, and Thomas 2004). Both test whether points from the EDF and CDF are from the same distribution. The Kolmogorov-Smirnov uses the test statistic of the largest difference between a point on the Q-Q plot and the line $y = x$, whereas the Cramér-von Mises test uses the sum of all the differences. As it takes into account more information and is therefore more powerful, the Cramér-von Mises is generally preferred. A significant result from either test gives evidence against the null hypothesis (that the data arose from the fitted model), suggesting that the model does not fit the data well.

We can generate a Q-Q plot and test results using the `gof_ds` function. Figure 6 shows the goodness of fit tests for two models for the amakihi data. We first fit a half-normal model without covariates or adjustments (setting `adjustment=NULL` will force `ds` to fit a model with no adjustments):

```
R> amakihi_hn <- ds(amakihi, truncation = 82.5, transect = "point",
```

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.


```
+                    key = "hn", adjustment = NULL)

Fitting half-normal key function

Key only model: not constraining for monotonicity.

AIC= 10833.841

No survey area information supplied, only estimating detection function.

R> gof_ds(amakihi_hn)


Goodness of fit results for ddf object

Distance sampling Kolmogorov-Smirnov test
Test statistic =  0.059345  P =  0.00031527

Distance sampling Cramer-von Mises test (unweighted)
Test statistic =  0.93083  P =  0.003578

R> gof_ds(amakihi_hr_obs_mas)


Goodness of fit results for ddf object

Distance sampling Kolmogorov-Smirnov test
Test statistic =  0.036251  P =  0.076237

Distance sampling Cramer-von Mises test (unweighted)
Test statistic =  0.15016  P =  0.38908
```

We can conclude that the half-normal model does not pass our goodness of fit tests and should be discarded. Both hazard-rate models (output only shown for hazard-rate model with observer and minutes after sunrise) had non-significant goodness of fit test statistics and are both therefore deemed plausible models. The corresponding Q-Q plots are shown in Figure 6, comparing the half-normal model with the hazard-rate model with observer and minutes after sunrise included.

For non-exact distance data, a χ^2-test can be used to assess goodness of fit (see Buckland *et al.* 2001, Section 3.4.4). χ^2-test results are produced by `gof_ds` when binned/grouped distances are provided.

4.2. Model selection

Once we have a set of plausible models, we can use Akaike's Information Criterion (AIC) to select between models (see e.g. Burnham and Anderson 2003). **Distance** includes a function

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

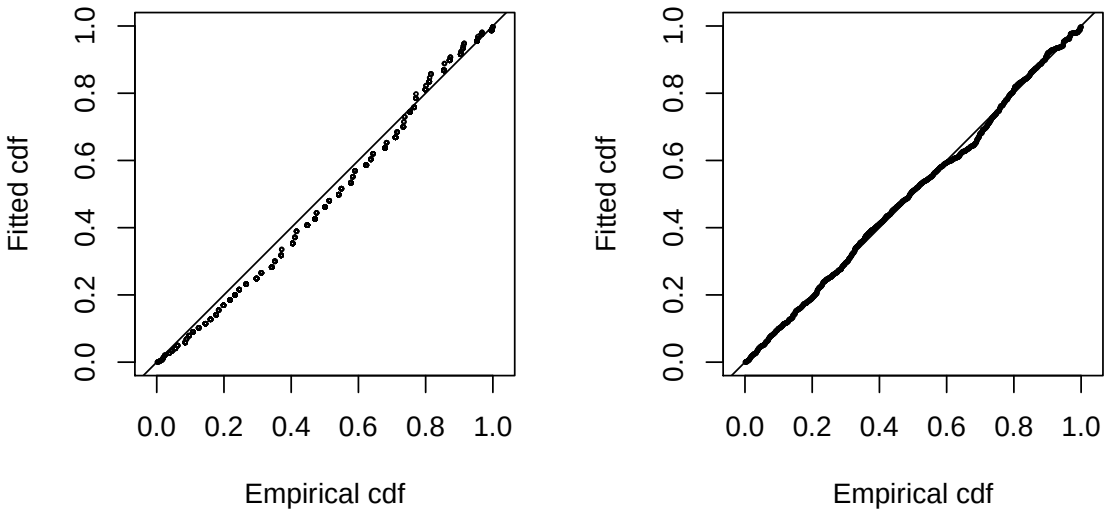

Figure 6: Comparison of quantile-quantile plots for a half-normal model (no adjustments, no covariates; left) and hazard-rate model with observer and minutes after sunrise (right) for the amakihi data.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Model	Key function	Formula	C-vM p-value	$\hat{P}_a$	se($\hat{P}_a$)	ΔAIC
`amakihi_hr_obs_mas`	Hazard-rate	obs + mas	0.389	0.319	0.020	0.000
`amakihi_hr_obs`	Hazard-rate	obs	0.271	0.314	0.020	1.073
`amakihi_hn`	Half-normal	1	0.004	0.351	0.011	56.465

Table 2: Summary for the detection function models fitted to the amakihi data. "C-vM" stands for Cramér-von Mises, P_a is average detectability (see "Estimating abundance and variance"), se is standard error. Models are sorted according to AIC.

to create table of summary information for fitted models, making it easy to get an overview of a large number of models. The `summarize_ds_models` function takes models as input and can be especially useful when paired with **knitr**'s `kable` function to create summary tables for publication (Xie 2015). An example of this output (with all models included) is shown in Table 2 and was generated by the following call to `summarize_ds_models`:

```
summarize_ds_models(amakihi_hn, amakihi_hr_obs, amakihi_hr_obs_mas)
```

In this case we may be skeptical about the top model as selected by AIC being truly better than the second best model, as there is only a very small difference in AICs. Generally, if the difference between AICs is less than 2, we may investigate multiple "best" models, potentially resorting to the simplest of these models. In the authors' experience, it is often the case that models with similar AICs will have similar estimates probabilities of detection, so in practice there is little difference in selecting between these models. It is important to note that comparing AICs between models with different truncations is not appropriate, as models with different truncation use different data.

5. Estimating abundance and variance

Though fitting the detection function is the primary modelling step in distance sampling, we are really interested in estimating density or abundance. It is also important to calculate our uncertainty for these estimates. This section addresses these issues mathematically before showing how to estimate abundance and its variance in R.

5.1. Abundance

We wish to obtain the abundance in a study region, of which we have sampled a random subset. To do this we first calculate the abundance in the area we have surveyed (the *covered area*) to obtain $\hat{N}_{\mathrm{C}}$, we can then scale this up to the full study area by multiplying it by the ratio of covered area to study area. We discuss other methods for spatially explicit abundance estimation in "Extensions".

First, to estimate abundance in the covered area ($\hat{N}_{\mathrm{C}}$), we use the estimates of detection probability (the $\{\hat{p}(\mathbf{z}_i); i = 1, \ldots, n\}$, above) in a Horvitz-Thompson-like estimator:

$$\hat{N}_{\mathrm{C}} = \sum_{i=1}^{n} \frac{s_i}{\hat{p}(\mathbf{z}_i)}, \tag{1}$$

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

where s_i are the sizes of the observed clusters of objects, which are all equal to 1 if objects only occur singly (Borchers and Burnham 2004). Thompson (2002) is the canonical reference to this type of estimator. Intuitively, we can think of the estimates of detectability ($\hat{p}(\mathbf{z}_i)$) as "inflating" the cluster sizes (s_i), we then sum over the detections (i) to obtain the abundance estimate. For models that do not include covariates, $\hat{p}(\mathbf{z}_i)$ is equal for all i, so this is equivalent to summing the clusters and inflating that sum by dividing through by the corresponding $\hat{p}(= \hat{p}(\mathbf{z}_i) \forall i)$.

Having obtained the abundance in the covered area, we can then scale-up to the study area:

$$\hat{N} = \frac{A}{a}\hat{N}_{\mathrm{C}},$$

where A is the area of the study region to which to extrapolate the abundance estimate and a is the covered area. For line transects $a = 2wL$ (twice the truncation distance multiplied by the total length of transects surveyed, L) and for points $a = \pi w^2 T$ (where πw^2 is the area of a single surveyed circle and T is the sum of the number of visits to the sampled points).

We can use the Horvitz-Thompson-like estimator to calculate the "average" detectability for models which include covariates. We can consider what single detectability value would give the estimated $\hat{N}_{\mathrm{C}}$ and therefore calculate:

$$\hat{P}_a = n/\hat{N}_{\mathrm{C}}.$$

This can be a useful summary statistic, giving us an idea of how detectable our n observed animals would have to be to estimate the same $\hat{N}$ if there were no observed covariates. It can also be compared to similar estimates in mark-recapture studies. P_a is included in the `summary` output and the table produced by `summarize_ds_models`.

Stratification

We may wish to calculate abundance estimates for some sub-regions of the study region, we call these areas *strata*. Stratification can be used to increase the precision of estimates if we know *a priori* that density varies between different parts of the study area. For example, strata may be defined by habitat types or animal sex (or some combination) which may be of interest for biological or management reasons. To calculate estimates for a given stratification each observation must occur in a stratum which must be labelled with a `Region.Label` and have a corresponding `Area` (if we are using an animal characteristic like sex, we would have the areas be the same but if we were using say forested vs. wetland habitat the areas of those strata would be different). Finally, we must also know the stratum in which each observation occurs.

As an example, the minke whale data contain two strata: `North` and `South` relating to strata further away from and nearer to the Antarctic ice edge, respectively.

5.2. Variance

We take an intuitive approach to uncertainty estimation, for a full derivation consult Marques and Buckland (2003). Uncertainty in $\hat{N}$ comes from two sources:

1. *Detection function*: Uncertainty in parameter ($\boldsymbol{\theta}$) estimation.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2. *Encounter rate*: Sampling variability arising from differences in the number of observations per transect.

We can see this by looking at the Horvitz-Thompson-like estimation in (1) and consider the terms which are random. These are: the detectability $\hat{p}(\mathbf{z}_i)$ (and hence the parameters of the detection function it is derived from) and n, the number of observations.

Model parameter uncertainty can be addressed using standard maximum likelihood theory. We can invert the Hessian matrix of the likelihood to obtain a variance-covariance matrix. We can then pre- and post-multiply this by the derivatives of $\hat{N}_{\mathrm{C}}$ with respect to the parameters of the detection function

$$\widehat{\mathrm{Var}}_{\mathrm{model}}\left(\hat{N}_{\mathrm{C}}\right) = \left(\frac{\partial \hat{N}_{\mathrm{C}}}{\partial \hat{\boldsymbol{\theta}}}\right)^{\top} \left(\hat{\mathbf{H}}(\hat{\boldsymbol{\theta}})^{-1}\right) \frac{\partial \hat{N}_{\mathrm{C}}}{\partial \hat{\boldsymbol{\theta}}}$$

where the partial derivatives of $\hat{N}_{\mathrm{C}}$ are evaluated at the MLE ($\hat{\theta}$) and $\mathbf{H}$ is the first partial Hessian (outer product of first derivatives of the log likelihood) for numerical stability (Buckland *et al.* 2001, p 62). Note that although we calculate uncertainty in $\hat{N}_{\mathrm{C}}$, we can trivially scale-up to variance of $\hat{N}$ (by noting that $\hat{N} = \frac{A}{a}\hat{N}_{\mathrm{C}}$ and therefore $\widehat{\mathrm{Var}}_{\mathrm{model}}\left(\hat{N}\right) = \left(\frac{A}{a}\right)^2 \widehat{\mathrm{Var}}_{\mathrm{model}}\left(\hat{N}_{\mathrm{C}}\right)$).

Encounter rate is the number of objects per unit transect (rather than just n). When covariates are not included in the detection function we can define the encounter rate as n/L for line transects (where L is the total line length) or n/T for point transects (where T is the total number of visits summed over all points). When covariates are included in the detection function, it is recommended that we substitute the n in the encounter rate with the estimated abundance $\hat{N}_{\mathrm{C}}$ as this will take into account the effects of the covariates (Innes, Heide-Jørgensen, Laake, Laidre, Cleator, Richard, and Stewart 2002).

For line transects, by default, **Distance** uses a variation of the estimator "R2" from Fewster, Buckland, Burnham, Borchers, Jupp, Laake, and Thomas (2009) replacing number of observations per sample with the estimated abundance per sample, thus taking into account cluster size if this is recorded (Innes *et al.* 2002; Marques and Buckland 2003):

$$\widehat{\mathrm{Var}}_{\mathrm{encounter},R2}\left(\hat{N}_{\mathrm{C}}\right) = \frac{K}{L^2(K-1)} \sum_{k=1}^{K} l_k^2 \left(\frac{\hat{N_{\mathrm{C},k}}}{l_k} - \frac{\hat{N}_{\mathrm{C}}}{L}\right)^2,$$

where l_k are the lengths of the K transects (such that $L = \sum_{k=1}^{K} l_k$) and $\hat{N_{C,k}}$ is the abundance in the covered area for transect k. For point transects we use the estimator "P3" from Fewster *et al.* (2009) but again replace n by $\hat{N}_{\mathrm{C}}$ in the encounter rate definition, to obtain the following estimator:

$$\widehat{\mathrm{Var}}_{\mathrm{encounter},P3}\left(\hat{N}_{\mathrm{C}}\right) = \frac{1}{T(K-1)} \sum_{k=1}^{K} t_k \left(\frac{\hat{N_{\mathrm{C},k}}}{t_k} - \frac{\hat{N}_{\mathrm{C}}}{T}\right)^2,$$

where t_k is the number of visits to point k and $T = \sum_{k=1}^{K} t_k$ (the total number of visits to all points is the sum of the visits to each point). Again, it is straightforward to calculate the encounter rate variance for $\hat{N}$ from the encounter rate variance for $\hat{N}_{\mathrm{C}}$.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Other formulations for the encounter rate variance are discussed in detail in Fewster *et al.* (2009). **Distance** implements all of the estimators of encounter rate variance given in that article. The `varn` manual page gives further advice and technical detail on encounter rate variance. For example for systematic survey designs, estimators S1, S2 and O1, O2 and O3 will typically provide smaller estimates of the variance.

We combine these two variances by noting that squared coefficients of variation (approximately) add (often referred to as "the delta method"; Seber 1982).

5.3. Estimating abundance and variance in R

Returning to the minke whale data, we have the necessary information to calculate A and a above, so we can estimate abundance and its variance. When we supply data to `ds` in the "flatfile" format given above, `ds` will automatically calculate abundance estimates based on the survey information in the data.

Having already fitted a model to the minke whale data, we can see the results of the abundance estimation by viewing the model summary:

```
R> summary(minke_hn)


Summary for distance analysis
Number of observations :  88
Distance range         :  0  -  1.5

Model : Half-normal key function
AIC   : 46.87216

Detection function parameters
Scale coefficient(s):
              estimate        se
(Intercept) -0.3411766 0.1070304

                       Estimate          SE         CV
Average p             0.5733038  0.04980421 0.08687229
N in covered region 153.4962706 17.08959835 0.11133559

Summary statistics:
  Region   Area CoveredArea  Effort  n  k         ER      se.ER     cv.ER
1  North 630582     4075.14 1358.38 49 12 0.03607238 0.01317937 0.3653591
2  South  84734     1453.23  484.41 39 13 0.08051031 0.01809954 0.2248102
3  Total 715316     5528.37 1842.79 88 25 0.04775368 0.01129627 0.2365529

Abundance:
  Label Estimate        se        cv      lcl      ucl       df
1 North 13225.44 4966.7495 0.3755450 6005.590 29124.93 12.27398
2 South  3966.46  955.9616 0.2410113 2395.606  6567.36 15.80275
3 Total 17191.90 5135.5862 0.2987212 9183.475 32184.07 14.00459
```

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Stratum	$\hat{N}$	se($\hat{N}$)	CV($\hat{N}$)
North	13225.44	4966.750	0.376
South	3966.46	955.962	0.241
Total	17191.90	5135.586	0.299

Table 3: Summary of abundance estimation for the half-normal model for the minke whale data.

```
Density:
  Label   Estimate          se        cv         lcl        ucl       df
1 North 0.02097339 0.007876453 0.3755450 0.009523884 0.04618738 12.27398
2 South 0.04681073 0.011281913 0.2410113 0.028272077 0.07750560 15.80275
3 Total 0.02403400 0.007179465 0.2987212 0.012838347 0.04499280 14.00459
```

This prints a rather large amount of information: first the detection function summary, then three tables:

1. `Summary statistics`: giving the areas, covered areas, effort, number of observations, number of transects, encounter rate, its standard error and coefficient of variation for each stratum. 2. `Abundance`: giving estimates, standard errors, coefficients of variation, lower and upper confidence intervals and finally the degrees of freedom for each stratum's abundance estimate. 3. `Density`: lists the same statistics as `Abundance` but for a density estimate.

In each table the bottom row gives a total for the whole study area.

The summary can be more concisely expressed by extracting information from the summary object. This object is a `list` of `data.frames`, so we can use the `kable` function from **knitr** to create summary tables of abundance estimates and measures of precision, such as Table 3. We prepare the `data.frame` as follows before using `kable`:

```
minke_table <- summary(minke_hn)$dht$individuals$N
minke_table$lcl <- minke_table$ucl <- minke_table$df <- NULL
colnames(minke_table) <- c("Stratum", "$\\hat{N}$", "$\\text{se}(\\hat{N}$)",
                           "$\\text{CV}(\\hat{N}$)")
```

6. Extensions

Distance sampling has been applied in a wide variety of situations. Objects being detected need not be animals; plants (Buckland, Borchers, Johnston, Henrys, and Marques 2007), beer cans (Otto and Pollock 1990), elephant dung (Nchanji and Plumptre 2001) and bricks on a lake bottom (Bergstedt and Anderson 1990) have been subjects of distance sampling investigations. The method of detection need not be visual sightings. Detections of objects can be through auditory as well as visual means (Marques, Thomas, Martin, Mellinger, Ward, Moretti, Harris, and Tyack 2013). Songs of birds (Buckland 2006) and whale vocalisations (Borchers, Pike, Gunnlaugsson, and Víkingsson 2009) are just two examples. Blows made by whales are examples of processes that can be modelled using distance sampling. Songs and

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

blows are indirect sampling methods producing estimates of cue density. Cue densities can be converted to animal densities with additional data needed to estimate the rate at which cues are produced and the rate at which they disappear. These are but a few examples of the applications to which distance sampling has been applied (a larger, but incomplete list of references is given at `http://distancesampling.org/dbib.html`).

The features of **Distance** are deliberately limited to provide a simplified interface for users. For more complex analyses of distance sampling data, we provide additional packages for modelling in R.

We noted at the start of the article that **Distance** is a simple-to-use wrapper around the package **mrds**. Additional features available in **mrds** include models that relax the assumption that detection is certain at zero distance from the transect (by including data from additional observers). This is done using mark-recapture type methods which require additional survey methodology, known as double observer surveys (see Burt *et al.* 2014, for an introduction).

Distance can provide us with estimates of abundance or density for each strata as a whole but tells us nothing about the distribution of animals within strata. One option is to divide the study area into smaller and smaller strata to try to detect patterns in spatial distribution, however, a more rigorous approach is to build a spatial model. Such models incorporate spatially-referenced environmental data (for example derived from GIS products). **Distance** interfaces with one such package used to perform this type of analysis: **dsm** (Miller, Rexstad, Burt, Bravington, and Hedley 2015). So-called "density surface modelling" uses the generalized additive model framework (e.g. Wood 2006) to build models of abundance (adjusting counts for imperfect detectability) as a function of environmental covariates, as part of a two stage model (Hedley and Buckland 2004; Miller, Burt, Rexstad, and Thomas 2013).

Uncertainty in measured covariates (e.g. cluster size) and model uncertainty (when two models have similar fit but substantially different estimates) can be incorporated using the multi-analysis distance sampling package **mads** (Marshall 2015b). In addition, **mads** can also incorporate sightings with unknown species identification. This is done by estimating the abundance of these 'unidentified' sightings and pro-rating them to the known species (Gerrodette and Forcada 2005).

As mentioned above, survey design is critical to ensuring that resulting distance sampling data can be reliably analysed. **DSsim** allows users to test out different designs in their study region and tailor population attributes to reflect the species they are working with. **DSsim** (Marshall 2015a) allows users to more easily identify challenges unique to their study and select a survey design which is more likely to yield accurate and precise estimates.

Distance for Windows has many users (over 45,000 downloads since 2002) and they may be overwhelmed by the prospect of switching existing analyses to R. For that reason we have created the **readdst** (Miller 2015) package to interface with projects created by Distance for Windows. The package can take analyses created using the CDS, MCDS and MRDS engines in Distance for Windows, extract the data and create equivalent models in R. **readdst** can also run these analyses and test the resulting statistics (for example, $\hat{N}$ or $\hat{P}_a$) calculated in R against those calculated by Distance for Windows. We hope that **readdst** will provide a useful transition to R for interested users. **readdst** is currently available on GitHub at `https://github.com/distancedevelopment/readdst`.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

7. Conclusion

We have given an introduction as to how to perform a distance sampling analysis in R. We have covered the possible models for detectability, model checking and selection and finally abundance and variance estimation.

In combination with tools such as **knitr** and **rmarkdown** (Allaire, Cheng, Xie, McPherson, Chang, Allen, Wickham, Atkins, and Hyndman 2015), the helper functions in **Distance** provide a useful set of tools to perform reproducible analyses of wildlife abundance for both managers and ecologists. We also direct readers' attention to **DsShiny** (Laake 2014), a package that builds on **Shiny** (Chang, Cheng, Allaire, Xie, and McPherson 2016) and **mrds** to allow users to build and fit models in a graphical interface, with output to RMarkdown.

R and its extension packages provide many tools exploratory data analysis that can be useful for a distance sampling analysis. We hope that this paper provides useful examples for those wishing to pursue distance sampling in R. More information on distance sampling can be found at `http://distancesampling.org` and a mailing list is maintained at `https://groups.google.com/forum/#!forum/distance-sampling`.

We note that there are other packages available for performing distance sampling analyses in R but believe that **Distance** is the most flexible and feature-complete, and provides pathways to a range of more complicated analyses. Appendix A gives a feature comparison between **Distance** and other R packages for analysis of distance sampling data.

8. Acknowledgements

The authors would like to thank the many users of **Distance**, **mrds** and Distance for Windows who have contributed bug reports and suggestions for improvements over the years. We would particularly like to thank Steve Buckland, David Borchers, Tiago Marques, Jon Bishop and Lorenzo Milazzo for their contributions. We thank Tiago Marques a second time for helpful comments on an early version of the paper and Colin Beale who suggested the "flatfile" data format. We also thank Ken Burnham and David Anderson for fundamental contributions to the early development of these methods.

Appendix A: Feature comparison

There are four packages available for analysis of distance sampling data in R that we are aware of. All are available on CRAN. On top of **Distance** and **mrds** (Laake *et al.* 2015) described above, they are **Rdistance** (McDonald, Nielson, and Carlisle 2015) and **unmarked** (Fiske and Chandler 2011). Table 4 provides a feature comparison of these packages.

Abundance estimation

Distance and **mrds** models can be used as part of a density surface model using **dsm**, which allows abundance to be modelled as a function of spatially varying covariates (such as location, sea depth, altitude etc). See Miller *et al.* (2013) for more information.

unmarked also allows abundance to vary according to covariates, via the abundance part of the likelihood. See Fiske and Chandler (2011) for more information on the package and Royle,

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Feature	**Distance**	**Rdistance**	**unmarked**	**mrds**
Line transects	x	x	x	x
Point transets	x		x	x
Interval (binned) distances	x		x	x
Exact distances	x	x		x
Continuous individual level covariates	x			x
Factor individual level covariates	x	x		x
Transect level covariates	x		x	x
Objects in clusters	x	x	x	x
Left truncation	x	x		x
Half-normal key	x	x	x	x
Hazard-rate key	x	x	x	x
Uniform key	x	x	x	x
Gamma key		x		x
Negative exponential key			x	
Adjustment terms	x	x		x
AIC adjustment selection	x	x		
Monotonicity constraints	x			x
Availability bias model			x	
Perception bias model				x
Abundance estimation	x	x	x	x
Density estimation	x	x	x	x
User-defined likelihood functions		x		

Table 4: Feature comparison of the available packages to perform distance sampling analyses.

Dawson, and Bates (2004) for more information on methodology.

Results **Rdistance** models can be use in combination with R modelling functions such as `lm`, `glm` etc to build abundance estimates which vary according to covariates. More information is available on the project's wiki.

References

Allaire JJ, Cheng J, Xie Y, McPherson J, Chang W, Allen J, Wickham H, Atkins A, Hyndman R (2015). ***rmarkdown**: Dynamic Documents for R*. R package version 0.8.1.

Becker EF, Christ AM (2015). "A Unimodal Model for Double Observer Distance Sampling Surveys." *PLoS ONE*, **10**(8), e0136403–18.

Becker EF, Quang PX (2009). "A Gamma-Shaped Detection Function for Line-Transect Surveys with Mark-Recapture and Covariate Data." *Journal of Agricultural, Biological, and Environmental Statistics*, **14**(2), 207–223.

Bergstedt RA, Anderson DR (1990). "Evaluation of line transect sampling based on remotely sensed data from underwater video." *Transactions of the American Fisheries Society*, **119**, 86–91.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Borchers DL, Burnham KP (2004). "General Formulation for Distance Sampling." In *Advanced Distance Sampling*, pp. 6–30. Oxford University Press, Oxford, UK.

Borchers DL, Pike D, Gunnlaugsson T, Víkingsson GA (2009). "Minke Whale Abundance Estimation from the NASS 1987 and 2001 Aerial CueâĂŞCounting Surveys taking Appropriate Account of Distance Estimation Errors." *NAMMCO Scientific Publications*, **7**, 95–110.

Branch TA, Butterworth DS (2001). "Southern Hemisphere Minke Whales: Standardised Abundance Estimates from the 1978/79 to 1997/98 IDCR-SOWER Surveys." *Journal of Cetacean Research and Management*.

Buckland ST (1992). "Fitting Density Functions with Polynomials." *Applied Statistics*, **41**(1), 63.

Buckland ST (2006). "Point transect surveys for songbirds: robust methodologies." *The Auk*, **123**(2), 345–357.

Buckland ST, Anderson DR, Burnham KP, Borchers DL, Thomas L (2001). *Introduction to Distance Sampling*. Oxford University Press, Oxford, UK.

Buckland ST, Anderson DR, Burnham KP, Laake JL, Borchers DL, Thomas L (2004). *Advanced Distance Sampling*. Estimating abundance of biological populations. Oxford University Press, Oxford, UK.

Buckland ST, Borchers DL, Johnston A, Henrys PA, Marques TA (2007). "Line Transect Methods for Plant Surveys." *Biometrics*, **63**(4), 989–998.

Buckland ST, Rexstad EA, Marques TA, Oedekoven CS (2015). *Distance Sampling: Methods and Applications*. Methods in Statistical Ecology. Springer International Publishing.

Burnham KP, Anderson DR (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Burnham KP, Anderson DR, Laake JL (1980). "Estimation of density from line transect sampling of biological populations."

Burnham KP, Buckland ST, Laake JL, Borchers DL, Bishop JRB, Thomas L (2004). "Further Topics in Distance Sampling." In *Advanced Distance Sampling*, pp. 385–389. Oxford University Press, Oxford, UK.

Burt ML, Borchers DL, Jenkins KJ, Marques TA (2014). "Using mark-recapture distance sampling methods on line transect surveys." *Methods in Ecology and Evolution*, **5**(11), 1180–1191.

Chang W, Cheng J, Allaire J, Xie Y, McPherson J (2016). ***shiny**: Web Application Framework for R*. R package version 0.13.1, URL `https://CRAN.R-project.org/package=shiny`.

Eidous OM (2005). "On Improving Kernel Estimators Using Line Transect Sampling." *Communications in Statistics - Theory and Methods*, **34**(4), 931–941.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fewster RM, Buckland ST, Burnham KP, Borchers DL, Jupp PE, Laake JL, Thomas L (2009). "Estimating the Encounter Rate Variance in Distance Sampling." *Biometrics*, **65**(1), 225–236.

Fiske I, Chandler R (2011). "unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance." *Journal of Statistical Software*, **43**(10).

Gerrodette T, Forcada J (2005). "Non-Recovery of Two Spotted and Spinner Dolphin Populations in the Eastern Tropical Pacific Ocean." *Marine Ecology Progress Series*, **291**, 1–21.

Ghalanos A, Theussl S (2014). ***Rsolnp**: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.15, URL https://CRAN.R-project.org/package=Rsolnp.

Giammarino M, Quatto P (2014). "On Estimating Hooded Crow Density from Line Transect Data through Exponential Mixture Models." *Environmental and Ecological Statistics*, **21**(4), 689–696.

Hedley SL, Buckland ST (2004). "Spatial Models for Line Transect Sampling." *Journal of Agricultural, Biological, and Environmental Statistics*, **9**(2), 181–199.

Innes S, Heide-Jørgensen MP, Laake JL, Laidre KL, Cleator HJ, Richard P, Stewart REA (2002). "Surveys of Belugas and Narwhals in the Canadian High Arctic in 1996." *NAMMCO Scientific Publications*, **4**(0), 169–190.

Laake JL (2014). ***DsShiny**: Shiny/knitr Application for Mark-Recapture Distance Sampling (MRDS)*. R package version 0.0.1, URL https://github.com/jlaake/DsShiny/.

Laake JL, Borchers DL, Thomas L, Miller DL, Bishop JRB (2015). ***mrds**: Mark-Recapture Distance Sampling*. R package version 2.1.15, URL http://CRAN.R-project.org/package=mrds.

Marques FFC, Buckland ST (2003). "Incorporating Covariates into Standard Line Transect Analyses." *Biometrics*, **59**(4), 924–935.

Marques TA, Buckland ST, Bispo R, Howland B (2012). "Accounting for Animal Density Gradients using Independent Information in Distance Sampling Surveys." *Statistical Methods & Applications*, **22**(1), 67–80.

Marques TA, Buckland ST, Tosh D, McDonald RA, Borchers DL (2010). "Point Transect Sampling Along Linear Features." *Biometrics*, **66**(4), 1247–1255.

Marques TA, Thomas L, Fancy SG, Buckland ST (2007). "Improving Estimates of Bird Density Using Multiple-Covariate Distance Sampling." *The Auk*, **124**(4), 1229.

Marques TA, Thomas L, Martin SW, Mellinger DK, Ward JA, Moretti DJ, Harris D, Tyack PL (2013). "Estimating animal population density using passive acoustics." *Biological Reviews*, **88**, 287–309. doi:10.1111/brv.12001.

Marsh H, Sinclair DF (1989). "Correcting for Visibility Bias in Strip Transect Aerial Surveys of Aquatic Fauna." *The Journal of Wildlife Management*, **53**(4), 1017.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Marshall L (2015a). ***DSsim**: Distance Sampling Simulations*. R package version 1.0.4, URL http://CRAN.R-project.org/package=DSsim.

Marshall L (2015b). ***mads**: Multi-Analysis Distance Sampling*. R package version 0.1.3, URL http://CRAN.R-project.org/package=mads.

McDonald T, Nielson R, Carlisle J (2015). *Rdistance: Distance Sampling Analyses*. R package version 1.3.2, URL https://github.com/tmcd82070/Rdistance/wiki.

Miller DL (2015). ***readdst**: Convert Distance for Windows Projects to R Analyses*. R package version 0.0.3, URL https://github.com/distancedevelopment/readdst.

Miller DL, Burt ML, Rexstad EA, Thomas L (2013). "Spatial models for distance sampling data: recent developments and future directions." *Methods in Ecology and Evolution*, **4**(11), 1001–1010.

Miller DL, Rexstad EA, Burt LM, Bravington MV, Hedley S (2015). ***dsm**: Density Surface Modelling of Distance Sampling Data*. R package version 2.2.11, URL http://CRAN.R-project.org/package=dsm.

Miller DL, Thomas L (2015). "Mixture models for distance sampling detection functions." *PLoS ONE*.

Nchanji AC, Plumptre A (2001). "Seasonality in elephant dung decay and implications for censusing and population monitoring in south-western Cameroon." *African Journal of Ecology*, **39**, 24–32.

Otto MC, Pollock KH (1990). "Size bias in line transect sampling: a field test." *Biometrics*, **46**(1), 239.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Royle JA, Dawson DK, Bates S (2004). "Modeling Abundance Effects in Distance Sampling." *Ecology*, **85**(6), 1591–1597.

Seber GAF (1982). *The Estimation of Animal Abundance and Related Parameters*. Macmillan.

Strindberg S, Buckland ST, Thomas L (2004). "Design of distance sampling surveys and Geographic Information Systems." In *Advanced Distance Sampling*. Oxford University Press, Oxford, UK.

Thomas L, Buckland ST, Rexstad EA, Laake JL, Strindberg S, Hedley SL, Bishop JRB, Marques TA, Burnham KP (2010). "Distance software: design and analysis of distance sampling surveys for estimating population size." *Journal of Applied Ecology*, **47**(1), 5–14.

Thomas L, Williams R, Sandilands D (2007). "Designing line transect surveys for complex survey regions." *Journal of Cetacean Research and Management*, **9**(1), 1–13.

Thompson SK (2002). *Sampling*. 2 edition. John Wiley & Sons.

Wood SN (2006). *Generalized Additive Models*. An Introduction with R. CRC Press.

bioRxiv preprint doi: https://doi.org/10.1101/063891; this version posted July 14, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Xie Y (2015). ***knitr**: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.11.

Affiliation:

David L Miller
Integrated Statistics and University of St Andrews
16 Sumner Street, Woods Hole, MA 02543, USA and
Centre for Research into Ecological and Environmental Modelling,
The Observatory,
St Andrews, Fife KY16 9LZ, Scotland
E-mail: dave@ninepointeightone.net
URL: http://converged.yt

Eric Rexstad
University of St Andrews
Centre for Research into Ecological and Environmental Modelling,
The Observatory,
St Andrews, Fife KY16 9LZ, Scotland
E-mail: Eric.Rexstad@st-andrews.ac.uk

Len Thomas
University of St Andrews
Centre for Research into Ecological and Environmental Modelling,
The Observatory,
St Andrews, Fife KY16 9LZ, Scotland
E-mail: len.thomas@st-andrews.ac.uk
URL: http://lenthomas.org

Laura Marshall
University of St Andrews
Centre for Research into Ecological and Environmental Modelling,
The Observatory,
St Andrews, Fife KY16 9LZ, Scotland
E-mail: lhm@st-andrews.ac.uk

Jeffrey L Laake
Marine Mammal Laboratory
Alaska Fisheries Science Center
National Oceanic and Atmospheric Administration
7600 Sand Point Way N.E., Seattle, WA 98115, USA
E-mail: Jeff.Laake@noaa.gov