



Distance Sampling in R

David L Miller Eric Rexstad Len Thomas Laura Marshall Jeffrey L Laake

University of St Andrews University of St Andrews University of St Andrews University of St Andrews NOAA

Abstract

The study of plants or animals often involves the estimation of population size and/or their spatial distribution. These characteristics are important considerations in wildlife management. We introduce the R package **Distance** that uses distance sampling data to estimate populations size (abundance) and can be used in conjunction with another R package **dsm** to estimate spatial distribution. We describe example analyses using this package as well demonstrating the links with other R packages offering specialist solutions to specific distance sampling challenges. We demonstrate how this package 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 transects, point transects, detection functions, Horvitz-Thompson, R, Distance.

1. Introduction

(LT: I wonder who the intended audience is for this paper? (LM: - Distance for Windows users and other people looking do distance sampling in R?) The way it is currently written, it “kinda” introduces distance sampling as well as the relevant commands. Is this standard for a paper in this journal – to introduce the methods as well as the software? Can we assume a reasonable level of analytical sophistication on the part of the reader, and so start with a few more formulae? Can we assume they know the methods, more or less, or point them at somewhere where this is already introduced. (Perhaps I should go back and have a look at the EcolApps paper to see what approach we took there... (LM: The paper on Distance in JAppEcol? There is quite a big intro on distance sampling there.)))

Distance sampling (Buckland et al. 2001; Buckland et al. 2004; Buckland et al. 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 see and count everything.

Conventional distance sampling assumes the observer is either stood at a point or moving along a line and will only see everything that occurs exactly at that point or on the line. It is then expected that the further an object is from the sampler (the point or line) the less likely it is that the observer will see it. By recording the distance to each of the detected objects we can fit a model to describe the probability of detection given distance from the sampler, which we refer to as 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. Further information on distance sampling, including its application, various extensions to deal with more complex situations and recent methodological developments can be found at <http://distancesampling.org/dbib.html>.

The Windows program Distance (or “DISTANCE”; for clarity henceforth “Distance for Windows”) was first released in November 1998 as Distance for Windows version 3.5 and was initially based on its predecessor “TRANSECT”. Since this time it has evolved to include various design and analysis features. Distance for Windows versions 5 onwards have included R packages as the analysis engines providing additional more complex analysis options than those offered by the original FROTRAN engines. (LM Add part here to say what this paper describes... pathway to sustainability using R?) Further details of the history of Distance for Windows are included in Thomas et al. (2010).

(LM: still need to incorporate this) Here we show how to fit detection functions, perform model checking and selection, and estimate abundance in R (R Core Team 2015) using the package **Distance** (Miller 2015a). **Distance** is a wrapper package around the more complex (and more powerful) **mrds** package (Laake et al. 2015) and offers a subset of the analyses possible with that package.

(LT: Possibly edit to add the actual years.) The Windows program Distance (or “DISTANCE”; for clarity henceforth “Distance for Windows”) has been available for many years to perform distance sampling analyses (and before that program “TRANSECT”). From version 5 of Distance for Windows, R packages have been included to perform some parts of an analysis (further details of the history of Distance for Windows are included in Thomas et al. 2010).

(LT: This won’t mean much unless we’ve defined detection function first – possibly need to do a little more work introducing the method and being explicit about where to go to find out more – e.g., encyclopaedia articles, free online stuff, etc. The concepts mostly are explained a little further down here, so possibly just some re-ordering needed.)

1.1. Distance sampling

(LT: a census is a complete count; I don’t really think that sample surveys fall under this. So, a bit of editing needed here.) Census-type surveys (quadrat or strip transects assuming perfect detectability within strips or quadrats) are inefficient (requiring considerable field effort) and we should expect that not all objects (animals, plants, dung, etc) can be observed. Accounting for imperfect detectability is an important consideration when obtaining accurate estimates of abundance (Lahoz-Monfort, Guillera-Aroita, and Wintle 2013). Using the extra

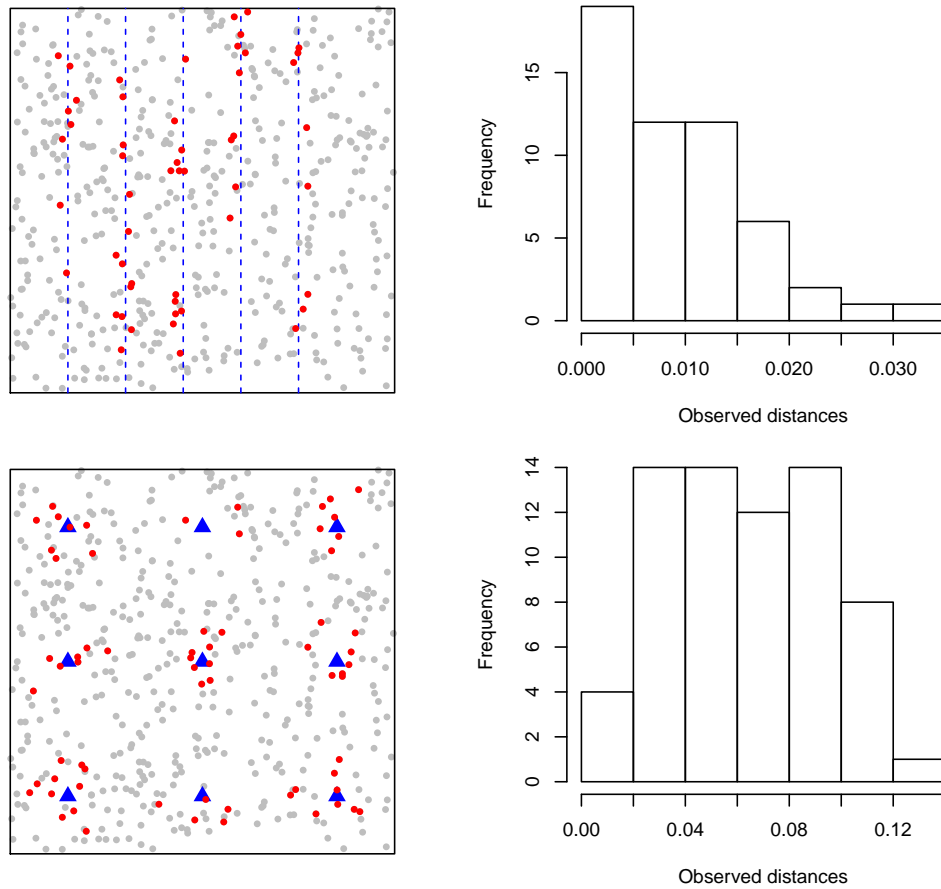


Figure 1: Left side plots show an example of a survey of an area containing a population of 500 objects; blue lines (top plot) and crosses (bottom plot) indicate sampler placement, red dots indicate detected individuals and grey dots give the locations of unobserved individuals. The right side of the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

information gained by recording distance from the sampler to the observation, it is possible to model detectability. Because we expect detectability to decrease with increasing distance from the sampler, we model detectability as a function of distance (plus perhaps other covariates, see below). (LT: The word detection function is used again without it having been defined explicitly.) We use the probabilities of detection calculated from the detection function to estimate abundance in the area covered by the survey, which we then scale-up to the study region of interest.

Distance sampling surveys are conducted using two transect types: line and point transects. In line transect sampling observers walk (or fly, sail, etc) lines, observing objects and recording their perpendicular distance to the line; whereas in point transect sampling observers remain stationary at a location and record distances to objects from that point. Field methods are chosen to be suitable to species and habitat constraints (Buckland et al. 2015).

(LT: What is said about the decreasing number of observations is incorrect, really, and the “taking account of” just makes it confusing, I think. Needs a bit more editing - possibly

re-order to explain what we see first.) For both points and lines, given the geometry of the sampler (see “Detection functions”), the histogram of distances should show a decreasing number of observations with increasing distance from the sampler (having taken into account sample geometry, see below). For line transects we expect objects to be distributed uniformly with respect to distance from the line (assuming that lines and points are placed independently of animal distribution) with detectability causing the number of detections to decrease (as distance from the sampler increases). For point transects we recognise that as distance from the point increases, the area of the circle with a corresponding radius increases with distance squared, hence the number of objects available to be detected is a linearly increasing function of distance from the point.

(LT: A detection function is not a line; it is also scaled differently. This needs editing.) Plotting a histogram of observed distances, we can crudely estimate the drop-off in detectability by eye if we trace a line that approximates the tops of the histogram bars: this is a detection function. **Distance** estimates the parameters for a fixed-form detection function using maximum likelihood estimation. We address possible models below.

(LT: stranded one-sentence paragraph – needs integrating.) Figure 1 shows sampling of a population of 500 individuals using line and point transects and their corresponding histograms.

The rest of the paper has the following structure: we describe data organisation for use with **Distance**; models for the detection function are described in terms of formulation and examples of fitting 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 software packages used for estimating the abundance of biological populations. (LT: Do we mean software packages in general (e.g., Mark) or R packages? Do we actually do this – i.e., talk about how Distance fits in among all this, or is this now an appendix, and does it cover all methods or just distance-based methods.)

(LT: Only got this far: will do more later...)

2. Data

We show examples of analyses in **Distance** using two data sets: one line transect and one point transect. (LT: I think it might be better to delete this next sentence – after all, two datasets can’t possibly represent all the possibilities there are in the real world – CDS, MCDS, MRDS, with and without covariates, problems such as spiked data, animal movement, measurement error, etc etc.) These data sets have been chosen to be representative of data seen in practice.

Minke whales

The line transect data is based on a survey of Antarctic minke whales (*Balaenoptera bonaerensis*). The data are simulated from models fitted to data from the International Whaling Commission’s International Decade of Cetacean Research Southern Ocean Whale and Ecosystem Research (IWC IDCR-SOWER) programme 1992-1993 austral summer surveys (Branch and Butterworth 2001). They consist of 99 observations and include a stratification based on location (near or distant from ice edge) and effort data (transect lengths). The survey is shown in Figure 2.

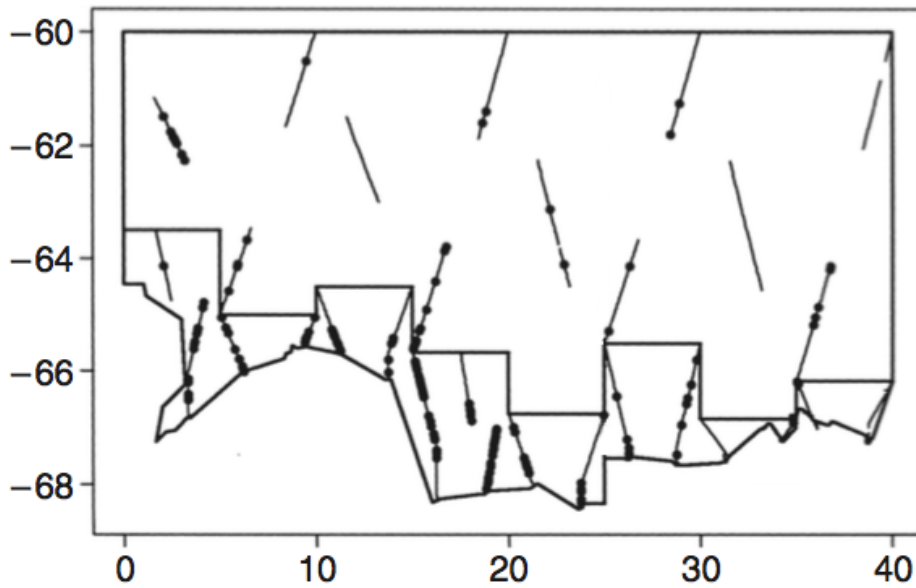


Figure 2: Strata used for the minke whale data adapted from Hedley and Buckland (2004). Points show the (simulated) locations of observations along transect lines. The stepped line shows the boundary between North and South strata. Further details on the survey are available in Branch and Butterworth (2001) (simulated data is based on “1992/93 Area III” therein).

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 et al. (2007).

2.1. Data setup

The two example data sets used here are distributed with **Distance** so readers can reproduce our results. However, generally data will be collected in the field and will need to be formatted correctly for use with **Distance**. The package allows for a flexible format for data input ranging from very simple to complex:

- In the simplest case, where the objective is to estimate a detection function, all that is needed is a vector of distances.
- To include additional covariates into the detection function (see “Detection functions”) we use a `data.frame`. The `data.frame` contains a column called `distance` (containing the observed distances) and additional named columns for covariates potentially useful in modelling detectability (for example `observer` or `seastate`). Some column names are reserved: `object` for an observation identifier (see “Extensions”), `size` for group or cluster size (see “Detection functions” and “Abundance and variance estimation”),

detected for whether an observation was detected (see “Extensions”) and the columns described in the next bullet.

- To estimate abundance beyond the area covered by expended survey effort, additional information is required. This consists of information about transect and occasion the observation was made (the **Sample.Label**), a column named **Effort** which gives the effort associated with that sample (for lines their length and for points the number of times that point was visited), the stratum the sample was located in (the **Region.Label**, which may be from pre- or post-survey stratification, see “Estimating abundance and variance”) and the stratum areas (**Area**). For transects which were surveyed but have no observations we enter **NA** in their **distance** column. 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 when using more complex models.

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 a set of cutpoints for the bins must be provided rather than the **distance** column, these are specified as two columns **distbegin** and **distend**. 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. For this reason, we recommended that these are converted to SI units before starting any analysis to ensure that resulting abundance and density estimates have sensible units.

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

```
library("Distance")
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:

```
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 during analysis below.

3. Detection functions

The detection function describes the relationship between observed distances and probability of detection. The detection function itself 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 , average in the sense of an average over distance from 0 to 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 (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 (known as the *truncation distance*). Including these larger distances in our analysis does not demonstrably improve the precision of abundance estimates (Buckland et al. 2001, sec. 4.3 and 5.3).

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 field procedure (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.
- *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 .

Table 1: Modelling options for key plus adjustment series models for the detection function. Adapted from Buckland et al. (2001), section 2.4.

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	$\sum_{o=2}^O a_o (y/w)^{2o}$

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).

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:

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

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 Figure 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 may be necessary to standardise the results to ensure that $g(0) = 1$, 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.

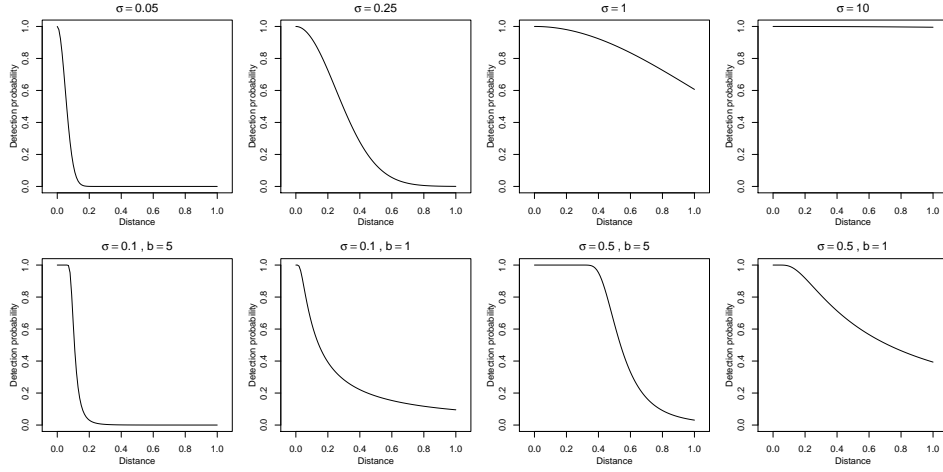


Figure 3: Half-normal (top row) and hazard-rate (bottom row) detection functions without adjustments, varying scale (σ) and (for hazard-rate) shape (b) parameters (values are given above the plots). On the top row from left to right, the study species becomes more detectable (higher probability of detection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

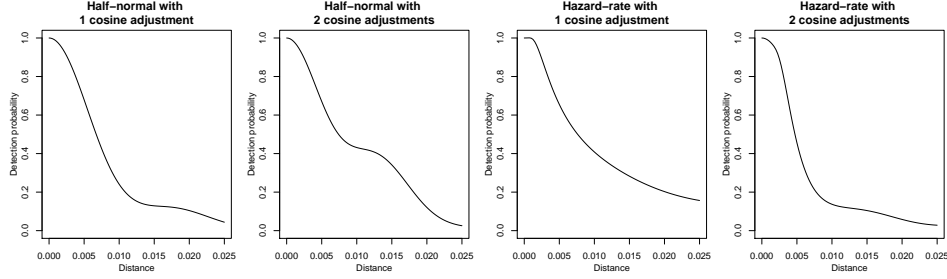


Figure 4: Possible (though not necessarily plausible) shapes for the detection function when adjustments are included for half-normal and hazard-rate models.

We do not always include adjustments (except in the case of the uniform key), in which case we refer to “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. These include things like the observer, size of group (if objects occur in groups), the vessel or platform used, the sea state, weather conditions and time of day. We assume that these variables affect detection 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 linear model (on the exponential scale) 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. Because we do not know the true distribution of the covariates, we must calculate the probability of detection conditional on the observed values of the covariates:

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

where \mathbf{z}_i is the vector of J covariates associated with observation i . For covariate models, we are calculating a value of “average” probability of detection (average in the sense of distance being integrated out) per observation. There will be 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, as we do not have any model for the distribution of the covariates. Without a distribution for the covariates, it’s not possible to know what the behaviour of the detection function will be, so we cannot setup meaningful constraints – so we cannot constrain the detection function to be monotonic. For this reason, we advise against 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

The workhorse of detection function fitting in **Distance** is the `ds` function. Here we show some of the possible formulations for the detection function we have seen above applied to the minke whale and amakihi data.

Minke whale

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

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

```
Starting AIC adjustment term selection.
```

```
Fitting half-normal key function
```

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

```
AIC= 46.872
```

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

```
AIC= 48.872
```

```
half-normal key function selected!
```

Note that `ds` will automatically select adjustment terms by AIC and show the selection steps. 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 to the next section).

We can specify the form of the detection function via the `key=` argument to `ds`. For example, a hazard rate model can be fitted as:

```
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!

`ds` fits a model with cosine adjustments (the default) but finds the AIC improvement to be insufficient to select the adjustment.

Amakihi

By default `ds` assumes the data given to it has been collected as line transects, but we can switch to point transects using `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 can be specified by (truncating sightings at 82.5m, given in Marques et al. 2007):

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

Cannot perform AIC adjustment term selection when covariates are used.

Fitting hazard-rate key function

AIC= 10778.448

No survey area information supplied, only estimating detection function.

Note that here `ds` warns us that we have not supplied enough information to estimate abundance, unlike with the minke whale data.

As with the minke whale model, we can plot the resulting model (Figure 5, middle panel). Rather than detection functions, Figure 5 shows the probability density functions for the amakihi models. Probability density function plots give a better sense of fit of the model than the detection functions, as for point data we must rescale the histogram when plotting the detection function to take into account the geometry of the point sampler. The amakihi model included covariates, so the plot shows the detection function averaged over levels/values of the covariate. Points on the plot indicate 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, we can see this is less clear when adding minutes after sunrise as a covariate into the model:

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

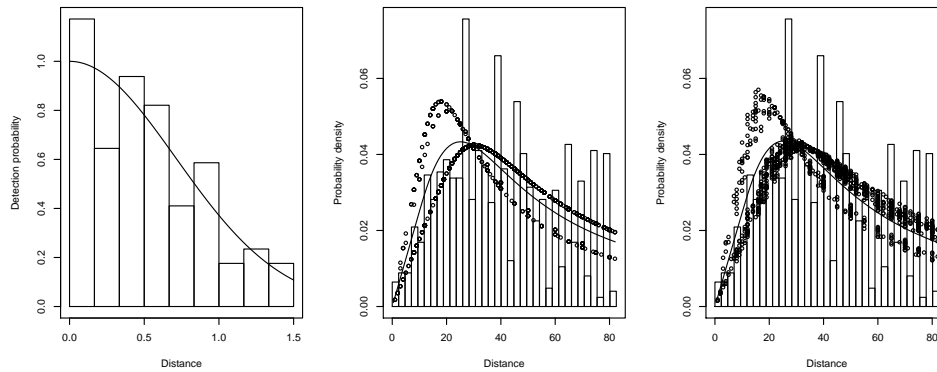


Figure 5: Left: fitted detection function overlayed on the histogram of observed distances for the minke whale data using half-normal model. Centre and right: plots of the probability density function for the amakihi models. Centre, hazard-rate with observer as a covariate; right, hazard-rate model with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

Cannot perform AIC adjustment term selection when covariates are used.

Fitting hazard-rate key function

AIC= 10777.376

No survey area information supplied, only estimating detection function.

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 for the amakihi, with observer as a covariate:

```
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):

	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 for the area covered by the survey (see the next section).

4.1. Goodness of fit

To judge goodness of fit for detection functions when exact distances are used, we want to compare the cumulative distribution function (CDF) and empirical distribution function (EDF) for the detection function via a quantile-quantile plot (Q-Q plot). The CDF evaluates the probability of observing an object at a distance less than or equal to some value. The EDF gives the proportion of observations for which the CDF is less than or equal to that of a given distance. 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 (where the “given values” are the observed distances). 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.

We can inspect Q-Q plots visually, though this is prone to subjective judgments. Instead we can quantify the Q-Q plot’s information using a Kolmogorov-Smirnov or Cramer-von Mises test (Burnham et al. 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 Cramer-von Mises test uses the sum of all the differences. As it takes into account more information and is therefore more powerful, the Cramer-von Mises is generally preferred. A significant result from either test indicates that the EDF and CDF do not come from the same distribution (and therefore the model is 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):

```

amakihi_hn <- ds(amakihi, truncation=82.5, transect="point", 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.

```

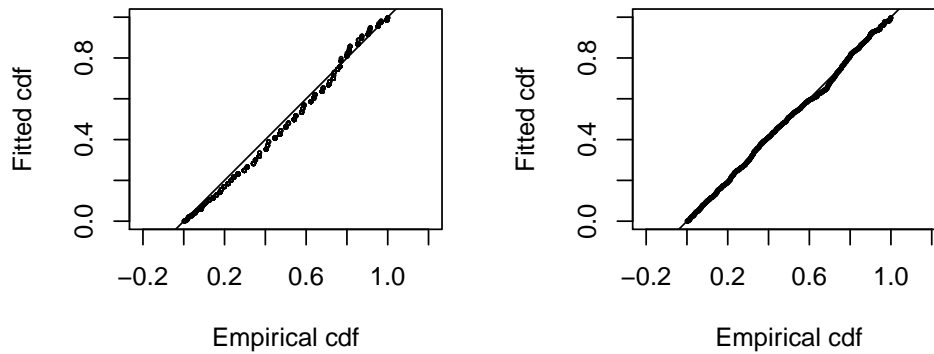


Figure 6: Comparison of quantile-quantile plots for a half-normal model (no adjustments, no covariates; left) and hazard-rate model with observer and minutes after sunrise (right) for the amakihi data.

```
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
```

```
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
```

So we can therefore conclude that the half-normal model does not pass our goodness of fit tests and should be discarded. 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.

4.2. Model selection

Once we have a set of models which fit well, we can use Akaike's Information Criterion (AIC) to select between models. **Distance** includes a function 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

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

Key function	Formula	C-vM p -value	\hat{P}_a	$se(\hat{P}_a)$	AIC	Δ AIC
Hazard-rate key function	obs + mas	0.389	0.319	0.020	10777.38	0.000
Hazard-rate key function	obs	0.271	0.314	0.020	10778.45	1.073
Half-normal key function	1	0.004	0.351	0.011	10833.84	56.465

when paired with **knitr**’s **kable** function to create summary tables for publication (Xie 2015). An example of this output 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)
```

5. Estimating abundance and variance

Though fitting the detection function is the primary modelling step in distance sampling, we are really interested in estimating detectability, and from that abundance. We also wish to calculate our uncertainty for each abundance estimate. 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 (representative) subset. To do this we first calculate the abundance in the area we have surveyed (the *covered area*) to obtain \hat{N}_C , we can then scale this up (based on the survey design) 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}_C), we use the estimates of detection probability (the $\{\hat{p}(\mathbf{z}_i); i = 1, \dots, n\}$, above) in a Horvitz-Thompson-like estimator:

$$\hat{N}_C = \sum_{i=1}^n \frac{s_i}{\hat{p}(\mathbf{z}_i)}, \quad (1)$$

where s_i are the sizes of the observed groups of objects, which is 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 group 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 groups 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:

$$\hat{N} = \frac{A}{a} \hat{N}_C,$$

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}_C and therefore calculate:

$$\hat{P}_a = n/\hat{N}_C.$$

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*. For example, strata may be defined by habitat types or animal gender (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 gender, 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 consists of two strata: `North` and `South` relating to a stratum further away and nearer the Antarctic ice edge, respectively. Figure 2 shows the two strata, along with observations and transect lines.

5.2. Variance

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

1. *Model parameter uncertainty*, from the estimation of the detection function parameters θ .
2. *Sampling uncertainty*, from the distribution of objects along the transect lines or between visiting occasions for points.

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. We assume that the observed group size (s_i) is recorded without error.

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}_C with respect to the parameters

of the detection function

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

where the partial derivatives of \hat{N}_C are evaluated at the MLE ($\hat{\boldsymbol{\theta}}$) 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}_C , we can trivially scale-up to variance of \hat{N} (by noting that $\hat{N} = \frac{A}{a} \hat{N}_C$ and therefore $\widehat{\text{Var}}_{\text{model}}(\hat{N}) = \left(\frac{A}{a}\right)^2 \widehat{\text{Var}}_{\text{model}}(\hat{N}_C)$).

Sampling uncertainty can be characterised by the *encounter rate*: 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}_C as this will take into account the effects of the covariates (Innes et al. 2002).

For line transects, by default, **Distance** uses a variation of the estimator “R2” from Fewster et al. (2009) replacing number of observations per sample with the estimated abundance per sample (Innes et al. 2002; Marques and Buckland 2003):

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

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 . Whereas for points we use estimator “P3” from Fewster et al. (2009) but again replacing n by \hat{N}_C in the encounter rate definition, we obtain the following estimator:

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

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}_C .

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.

We combine these two sources of variance by noting that squared coefficients of variation (approximately) add (Goodman 1960) (sometimes referred to as “the delta method”).

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:

```
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

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:

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

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

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, then a total for the whole study area.
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, then a total for the whole study area.
3. **Density:** lists the same statistics as **Abundance** but for a density estimate.

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 again 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

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

We noted at the start of the article that **Distance** is a simpler wrapper around the package **mrds**. Additional features are available in **mrds** including the ability to model data where the assumption that detection is certain at zero distance from the line or point is violated, using mark-recapture type methods for double observer surveys (see Burt et al. 2014 for an introduction).

The abundance estimates calculated here are based on the assumption that within a given stratum density is uniform. We may extend this approach to many strata, making the area of each very small to account for small-scale variation in space. A more rigorous approach is to build a spatial model incorporating spatially-referenced environmental data (for example derived from GIS products). **Distance** interfaces with one such package to perform this type of analysis: **dsm** (Miller et al. 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 et al. 2013).

Further distance sampling complexities such as 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 2015a). In addition, **mads** can also incorporate sightings for which species identification could not be achieved. This is done by estimating the abundance of these ‘unidentified’ sightings and pro-rating them as described in (Gerrodette and Forcada 2005).

In addition to the analyses, an equally if not more important stage of a distance sampling survey is the survey design. We have therefore developed a package which allows users to test out different designs in their specific study region and tailor population attributes to reflect the species they are working with. In this way **DSsim** (Marshall 2015b) allows users to more easily identify challenges unique to their study and select a survey design which is likely to yield the most accurate and precise results.

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 2015b) 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 and extract 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>.

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 et al. 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 hope that this paper provides useful examples for those wishing to pursue this. 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. 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** 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.

Bibliography

- Allaire, J J, Joe Cheng, Yihui Xie, Jonathan McPherson, Winston Chang, Jeff Allen, Hadley Wickham, Aron Atkins, and Rob Hyndman. 2015. *rmarkdown: Dynamic Documents for R*.
- Becker, Earl F, and Aaron M Christ. 2015. “A Unimodal Model for Double Observer Distance Sampling Surveys.” *PLoS ONE* 10 (8): e0136403–18.
- Becker, Earl F, and P X Quang. 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): 207–23.
- Borchers, David L, and Kenneth P Burnham. 2004. “General formulation for distance sampling.” In *Advanced Distance Sampling*, 6–30. Oxford University Press, Oxford, UK.
- Branch, T A, and D S Butterworth. 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, Stephen T. 1992. “Fitting Density Functions with Polynomials.” *Applied Statistics* 41 (1): 63.
- Buckland, Stephen T, David R Anderson, Kenneth P Burnham, David L Borchers, and Len Thomas. 2001. *Introduction to Distance Sampling*. Estimating Abundance of Biological Populations. Oxford University Press, Oxford, UK.
- Buckland, Stephen T, David R Anderson, Kenneth P Burnham, Jeffrey L Laake, David L Borchers, and Len Thomas. 2004. *Advanced Distance Sampling*. Estimating Abundance of Biological Populations. Oxford University Press, Oxford, UK.
- Buckland, Stephen T, E A Rexstad, Tiago A Marques, and Cornelia S Oedekoven. 2015. *Distance Sampling: Methods and Applications*. Methods in Statistical Ecology. Springer International Publishing.
- Burnham, Kenneth P, Stephen T Buckland, Jeffrey L Laake, David L Borchers, Jon R B Bishop, and Len Thomas. 2004. “Further topics in distance sampling.” In *Advanced Distance Sampling*, 385–89. Oxford University Press, Oxford, UK.
- Burt, Mary Louise, David L Borchers, Kurt J Jenkins, and Tiago A Marques. 2014. “Using mark-recapture distance sampling methods on line transect surveys.” *Methods in Ecology and Evolution* 5 (11): 1180–91.
- Eidous, Omar M. 2005. “On Improving Kernel Estimators Using Line Transect Sampling.” *Communications in Statistics - Theory and Methods* 34 (4): 931–41.
- Fewster, Rachel M, Stephen T Buckland, Kenneth P Burnham, David L Borchers, Peter E Jupp, Jeffrey L Laake, and Len Thomas. 2009. “Estimating the Encounter Rate Variance in Distance Sampling.” *Biometrics* 65 (1): 225–36.
- Gerrodette, T, and J Forcada. 2005. “Non-recovery of two spotted and spinner dolphin populations in the eastern tropical Pacific Ocean.” *Marine Ecology Progress Series* 291 (April): 1–21.
- Ghalanos, Alexios, and Stefan Theussl. 2014. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*.
- Giammarino, Mauro, and Piero Quatto. 2014. “On estimating Hooded crow density from line transect data through exponential mixture models.” *Environmental and Ecological Statistics* 21 (4): 689–96.

- Goodman, Leo A. 1960. "On the Exact Variance of Products." *Journal of the American Statistical Association* 55 (292): 708.
- Hedley, Sharon L, and Stephen T Buckland. 2004. "Spatial models for line transect sampling." *Journal of Agricultural, Biological, and Environmental Statistics* 9 (2): 181–99.
- Innes, Stuart, M P Heide-Jørgensen, Jeffrey L Laake, Kristin L Laidre, Holly J Cleator, Pierre Richard, and Robert EA Stewart. 2002. "Surveys of belugas and narwhals in the Canadian High Arctic in 1996." *NAMMCO Scientific Publications* 4 (0): 169–90.
- Laake, Jeffrey L, David L Borchers, Len Thomas, David L Miller, and Jon Bishop. 2015. *Mrds: Mark-Recapture Distance Sampling*. <http://CRAN.R-project.org/package=mrds>.
- Lahoz-Monfort, José J, Gurutzeta Guillera-Aroita, and Brendan A Wintle. 2013. "Imperfect detection impacts the performance of species distribution models." *Global Ecology and Biogeography* 23 (4): 504–15.
- Marques, Fernanda F C, and Stephen T Buckland. 2003. "Incorporating covariates into standard line transect analyses." *Biometrics* 59 (4): 924–35.
- Marques, Tiago A, Len Thomas, Steven G Fancy, and Stephen T Buckland. 2007. "Improving estimates of bird density using multiple-covariate distance sampling." *The Auk* 124 (4): 1229.
- Marshall, Laura. 2015a. *Mads: Multi-Analysis Distance Sampling*. <http://CRAN.R-project.org/package=mads>.
- . 2015b. *DSsim: Distance Sampling Simulations*. <http://CRAN.R-project.org/package=DSsim>.
- Miller, David L. 2015a. *Distance: Distance Sampling Detection Function and Abundance Estimation*. <http://CRAN.R-project.org/package=Distance>.
- . 2015b. *Readdst: Convert Distance for Windows Projects to R Analyses*. <https://github.com/distancedevelopment/readdst>.
- Miller, David L, and Len Thomas. 2015. "Mixture models for distance sampling detection functions." *PLoS ONE*.
- Miller, David L, Mary Louise Burt, Eric A Rexstad, and Len Thomas. 2013. "Spatial models for distance sampling data: recent developments and future directions." *Methods in Ecology and Evolution* 4 (11): 1001–10.
- Miller, David L, Eric Rexstad, Louise Burt, Mark V Bravington, and Sharon Hedley. 2015. *Dsm: Density Surface Modelling of Distance Sampling Data*. <http://CRAN.R-project.org/package=dsm>.
- R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Thomas, Len, Stephen T Buckland, Eric A Rexstad, Jeffrey L Laake, Samantha Strindberg, Sharon L Hedley, Jon R B Bishop, Tiago A Marques, and Kenneth P Burnham. 2010. "Distance software: design and analysis of distance sampling surveys for estimating population size." *Journal of Applied Ecology* 47 (1): 5–14.
- Thompson, S K. 2002. *Sampling*. 2nd ed. Wiley.
- Wood, Simon N. 2006. *Generalized Additive Models. An Introduction with R*. CRC Press.
- Xie, Yihui. 2015. *knitr: A General-Purpose Package for Dynamic Report Generation in R*.

Affiliation:

David L Miller

University of St Andrews

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

Laura Marshall

University of St Andrews

Centre for Research into Ecological and Environmental Modelling, The Observatory, St Andrews, Fife KY16 9LZ, Scotland

E-mail: lh@st-andrews.ac.uk

Jeffrey L Laake

NOAA

National Marine Mammal Laboratory Alaska Fisheries Science Center 7600 Sand Point Way N.E., Seattle, WA 98115, USA

E-mail: Jeff.Laake@noaa.gov