# Project

Chris Martin

05/01/2022

### Abstract

### Introduction

The purpose of this report is to detail to extensions made to the dsims R package and conduct analysis to compare the abundance estimates generated through a design based approach, distance sampling, and a model based approach, density surface modelling. This will allow researchers the opportunity to use the best possible model to fit the circumstances of their own study.

main objective is that the use of density surface modelling can be extended beyond the original survey area while distance sampling is more restricted in this approach. The simulations of different designs can be used to evaluate the how well these extended areas can be modeled without the requirement to sample there, potentially allowing for the survey design to be optimised to allow the maximum area to be estimated within a given accuracy.

## Background research

Informed from Buckland et al 2015 One of the key aims in areas of applied ecological research is to determine the abundance of a particular population of interest, such as in a periodic way to monitor its development over time and determine changes, or to evaluate the potential effect of a new factor, such as a human disturbance. The size of the population can determine the importance of any new factors, with a smaller populations more under threat from a given factor compared to an abundant one. One option for determining a populations size is to count every single individual, known as a census, similar to the UK completing a Census of its population every 10 years. However, in the natural world, this is only realistically possible in the simplest instances and therefore a different approach must be used. Researches often use some form of sampling method to conduct a sample of the target population and draw conclusions for the overall population based on this sample.

Distance sampling is an approach to estimating population abundance and density by surveying a small portion

need to define truncation distance, detection function, abundance and density from literature.

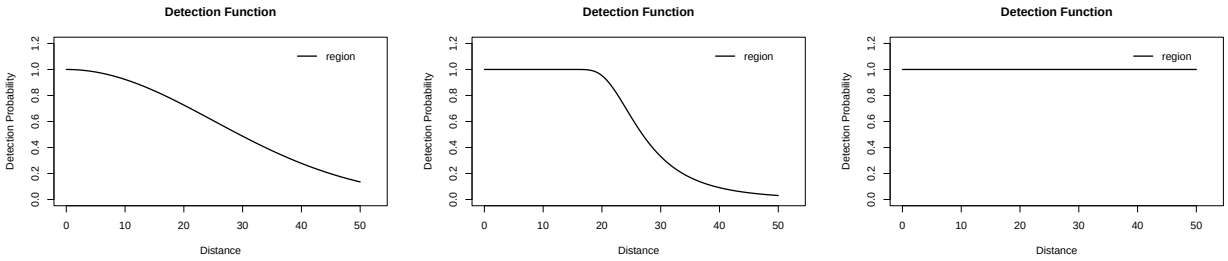
explanation of how the ds and dsm models generate their abundance estimates in R and under what circumstances one may be better than the other to inform potential comparisons

### Detection function estimation

### Distance Sampling Simulations

All the material in this section is based on Buckland et al (2015) Prior to the simulation for a particular design being run, a number of objects must be first be defined. The first object is the study region, this can

either be the default generated by R or user defined from a shapefile. Following this, a spatial distribution or density surface must be defined, from which animal locations can be generated based on the population description. The desired population size can be user defined and set for a series of simulations or be generated based on the spatial distribution supplied by the user. The desired truncation distance must then be defined and based on this an appropriate design can be generated. The main considerations when constructing the design are the type, either line or point transects and the desired number or length of transects. If line transects are used, the design angle may be altered from its default of 0. Based on the design, a set of survey transects can be generated, during which the detection process is simulated. The user can then define a detection function, based on either a half normal ('hn'), hazard rate ('hr') or uniform distribution ('uf') with a defined scale parameter and the desired truncation distance $w$, examples of which can be seen below:




Therefore for an animal at distance $x$ from the closest transect the probability of the animal being detected is given by the detection function evaluated at $x$, provided $x$ is less than or equal to $w$. The distance data generated during the survey is then analysed to estimate the abundance $N$ of the study area, with options available for several models to be analysed for each set of distance data, with a model selection criteria used to select the best, using AIC as the default. These operations are then repeated the specified number of times, say $R$, for each density and design, to obtain a set of simulations of animal distribution and survey design, alongside a corresponding set of estimates *Nhat* of N. typical values for $R$ are between 100 and 1000. In the case where the design is intrinsically selected by the user, as opposed to randomised, the exact same design will be used for all $R$ simulations

## Density Surface Modelling

This sections contains material base on D.L.Miller et al 2013 In order to construct a density surface model, initial the approach must be decided upon. The choice is between using a two stage approach, whereby the detection function is fitted first then subsequently fitting a spatial model, while the one stage approach leads to estimating the detection and spatial parameters simultaneously. Miller et al states that 'Generally, very little information is lost by taking the two stage approach' as transect width is comparably smaller than that of the study region, therefore, provide the population does not differ spatially within the transect, no information is lost by the two stage approach. This may lead is issues occurring where the density of the species has significant variability at the transect level. However, one drawback of the two stage model is that, to accurately evaluate the model uncertainty, the uncertainty in both the detection function and the spatial models should be suitably combined. For the remainder of this report only the two stage approach will be discussed. Initially, the detection function must be fitted, with the specification being the same as mentioned in the distance sampling section above. Following this, the density surface model can be fitted. To enable this to occur, the data must be separated into segments. This is easily done for point transects with each point being a segment however it more complicated for line transects. With line transects, they must be split up into J segments of length $l_j$. It is normally from the segments to be approximately square, with dimensions of 2 $w$ x 2 $w$ where $w$ is the truncation distance of the design. From here, the segment areas enter the model as part of an offset, to allow for non-constant segment areas. This leads the line transect segments to have an area of $2wl_j$ and the point transect segments with an area of $\pi w^2$. In the model, the counts or abundances are using a generalised additive model (GAM) using the sum of the smoothed covariates.

**Response models**

The model used when the count per segment is used as the response is:

$$\mathbb{E}(n_j) = \hat{p_j} A_j exp(\beta_0 + \sum_k f_k(z_{jk})$$

Where $f_k$ are the smoothed functions of the covariates and $\beta_0$ is the intercept term. By multiplying the segment area $A_j$ by the estimated probability of detection $p_j$ this gives the effective area of the segment, acting as an offset to account for different segment areas. Where distance is the only convariate in the detection function, $p_j$ is constant across all segments and therefore $\hat{p_j} = \hat{p} \forall j$. The distribution of $n_j$ can then be modeled using an overdispersed Poisson, Negative binomial or Tweedie distribution.

An alternative to using this is to use abundance estimates for each segment generated by distance sampling as the response.. To do this, the response $n_j$ is replaced by an estinator of the abundance in each section, $\hat{N_j}$ where this is defined as:

$$\hat{N_j} = \sum_{r=1}^{R_j} \frac{s_{jr}}{\hat{p_j}}$$

Where $R_j$ is the number of observations in the jth segment and $s_{jr}$ is the size of the rth group observed, with this being 1 if only individuals are observed. As identified by Buckland et al 2015, this is an Horvitz–Thompson-like estimator of the segment abundance, allowing for covariates to be included through $\hat{p_j}$. The fitted model then becomes:

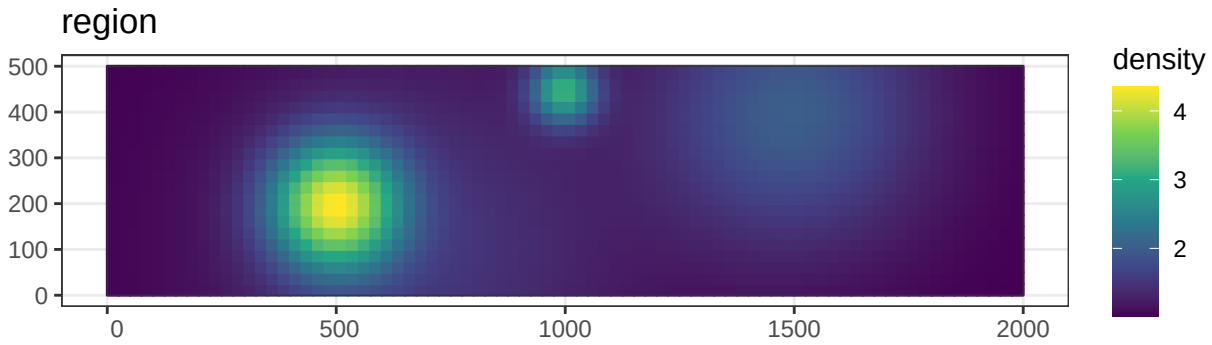
$$\mathbb{E}(\hat{N_j}) = A_j exp[\,\beta_0 + \sum_k f_k(z_{jk})]$$

Where the model follows the same three distributions as before. The main difference between these models is that the offset is now the physical area of each segment, as opposed to the effective area in the first model for $n_j$.

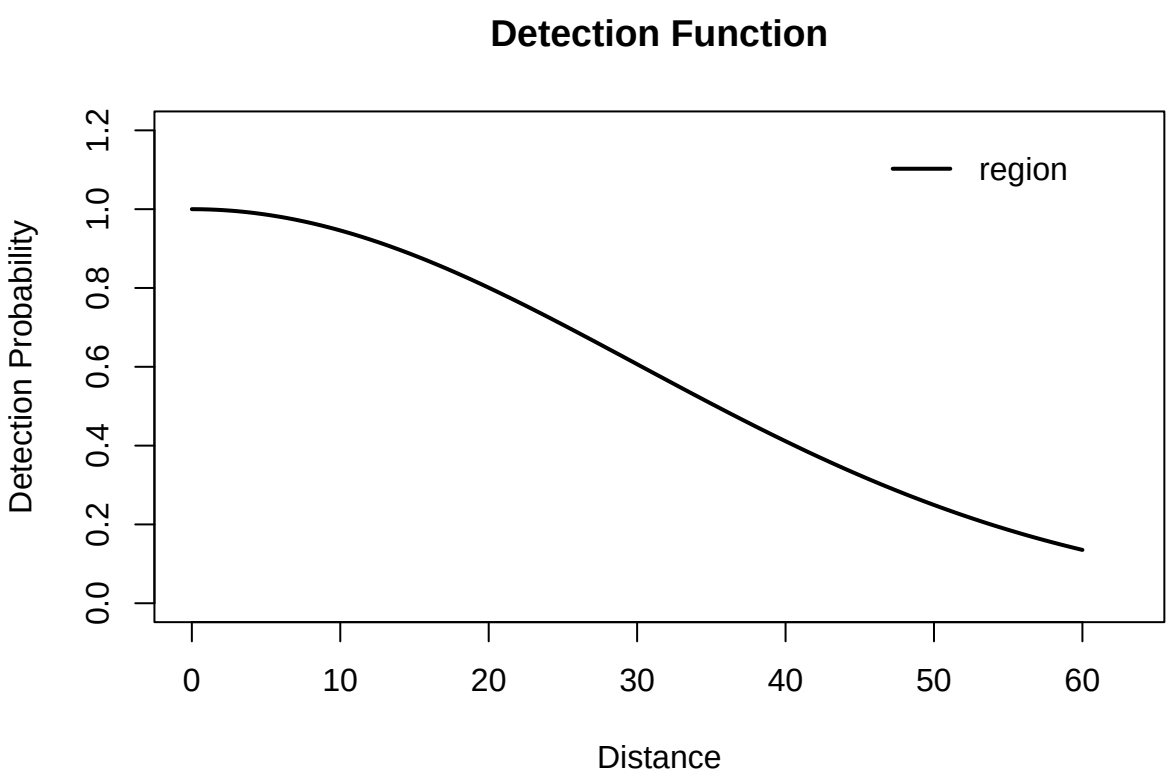
To allow for a DSM to predict abundance, a series of prediction cells must be defined. These are not necessarily restricted to just the original study region, allowing for regions outside the study area to be predicted over. Each of the prediction cells must include the same covariates as specified in the dsm, including the area of each cell. Predictions can then be made for the abundance in each cell and by summing these over the whole region, an overall abundance estimate can be obtained. The size of the prediction cells may be specified by the user, however cells 'smaller than the resolution of the spatially referenced data' do not have an influence on the abundance estimates produced by dsm.

## Modelling

The simulation was initially tested using the default region generated by dsims with a truncation distance of 60. Both point and line transect designs were run with an aim of 25 and 12 samplers for the respective designs. A basic test density was then constructed for the region with high and low spots as seen below with relatively gently gradients.




A population description was then constructed based on this density surface with a true population of 1000. A detection function was then defined as a half normal with scale parameter of 30 for both designs, producing the following detection function:




A prediction grid is then constructed across the study region and is identical for each simulation iteration, with the resolution of the grid set at the truncation distance of the design,XXX as Miller et al notes that using cells smaller than the spatial data resolution will not have an effect on the abundance estimates provided by dsmXXX.

The simulation loop then begins. For each iteration, a new survey constructed. From this, the observation data and segmented data is extracted.

In the case of line transect designs, the transects are split to allow them to be modelled as points. Each transect is split into segments of approximate length $2w$ with $w$ being the truncation distance, as suggested by Miller et al 2013 and each segment assigned its own unique sample label. For Point transect designs, each point is treated as its own segment. The polygons of each segment is then created using the st_buffer command (sf package, ref required), using $w$ as the distance. This leads to squares of approximately $2w * 2w$ for line transects and circles of radius $w$ for points. These allow the area of each segment to be calculated, a requirement for the dsm model. This is calculated using st_area, on the intersection between the polygons and the study region, to ensure only areas within the study region are counted towards segment area. Failure to do this results in the areas of each segment being larger than they are in the survey, and as a result the dsm abundance estimate is smaller than true, since the prediction grid is only over the survey area. Once the segment areas have been calculated, they can be linked to the observation data by allocating each observation to the nearest segment and giving this the respective segments Sample label in the observation data, overwriting the original allocation.

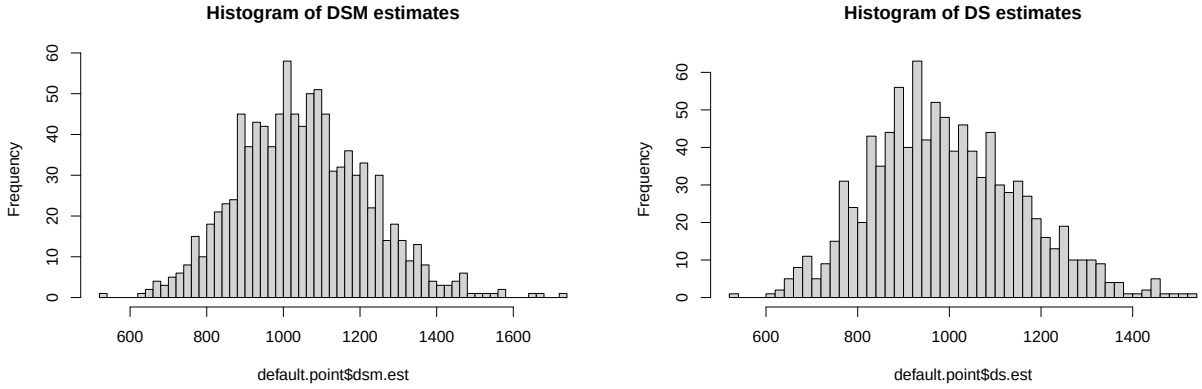
Based on this data, both a distance sampling model and density surface model are constructed, with the dsm modeling the counts against the smooth of spatial locations using a tweedie error distribution. In the smoothed term, the degrees of freedom is restricted to the total number of transects. The abundance estimates are extracted from both models and stored alongside the prediction variance and deviance explained from the dsm model.

## Results

Having completed 1000 bootstrap simulations for both the distance sampling and density surface models with each design, we can now examine and compare these to give us an insight into the circumstances under which a particular model is better of worse than the other.

### Default Region Point design

For the initial default region with the point transect design, the histograms of both the distance sampling and dsm abundance estimates are displayed below:




These plots show somewhat similar data since both estimates are generated by the same data set. If we now compare the means and 95% confidence intervals:

```
## [1] 1052.032
```

```
## [1] 992.0967
```

```
## [1]  722.3986 1381.6645
```

```
## [1]  674.1186 1310.0748
```

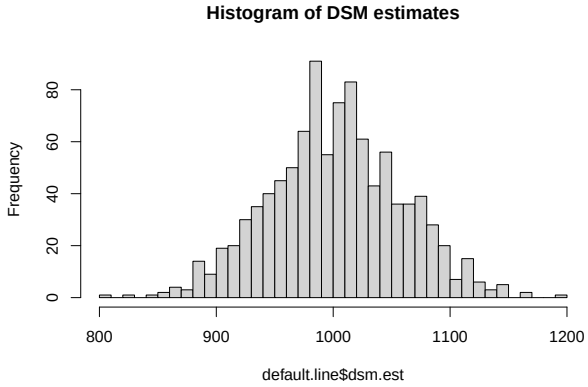
It can be seen that the mean of the dsm estimates appears to very close to the true population while the distance sampling method is closer. Examining the confidence intervals for both methods, the interval for the dsm model is in this case slightly wider than that of the ds model, indicating it is slightly less accurate in the case of this region and density surface however this effect is very small.

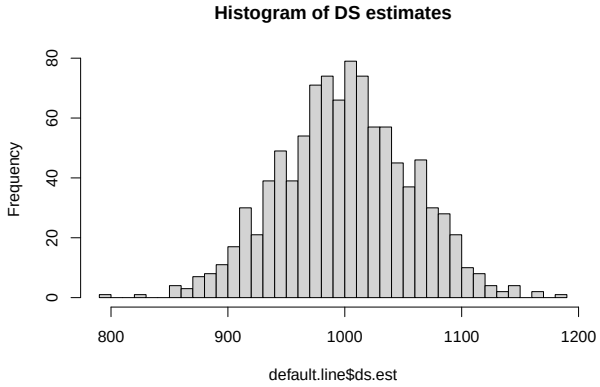
Next, we can evaluate how many times the true abundance, 1000, was within the 95% for every model computed.

```
## [1] 1000
```

### Default region parallel Line design

Now examining the results of the line transect design, we see the histograms of the two estimates below:




```
## [1] 1002.038
```

```
## [1] 1000.358
```

```
## [1]  890.8191 1113.2573
```

```
## [1]  890.8782 1109.8380
```

## conclusion