Distance sampling: Estimating densities of wildlife populations by modelling detectability

David Maphisa & Florian Weller
Distance sampling is applicable to many species groups (birds, plants, mammals, whales etc.) Surveys can be done on foot, plane, boat etc.

1. Point counts

2. Transects counts
The key is to account for observation process

Many factors can determine your ability to observe the species accurately eg. vegetation (grassland birds)

Time of the day (or temp)

Cloud cover etc.........................

How many animals can you see on the 2nd pic below?
We present two approaches to analysing your distance data

1. Package *unmarked* (function disamp) in R (Chandler et al. 2015)

I demonstrate this by using bird and vegetation data from my Phd thesis

My study site (FS-KZN boundary)

2. Florian will demonstrate Program DISTANCE (standalone) and talk more about assumptions and biases
The way package *unmarked* is implemented in R

**Observation process/detection process**
Allows inclusion of factors that may affect your ability to see or not see animals etc

**Hierarchical distance sampling**

**Biological process/density (ha)**
include factors that affect density
Data requirements

1. Bird distance sampling data (in my case fixed distance bands)

2. Habitat data - in my case detctn & density are affected by
   (i) fire – whether each transect was burned or not burned
   (ii) Grazing – in 3 categories (none, light or heavily grazed)
   (iii) My main focus was on grass height (avh) and grass cover (cover)

all these datasets (birds and vegtn must be loaded in R ) – merged into one object

Once this is done actual analysis is carried out (first detection and then density)
Out of eight species – work on one spcs at a time

data.PA <- bldata[bldata$Species=="African Pipit",]
data.PA$Season <- factor(data.PA$Season)

summary(data.PA)

ltUMF <- with(data.PA, {
  unmarkedFrameDS(y = cbind(X50m, X100m. X100m.1),
  siteCovs = data.frame(Year, Season, Area),
  dist.breaks = c(0, 50, 100, 500)
  tlength = rep(500,dim(data.PA)[1]), survey = "line", unitsIn = "m")
ltUMF
summary(ltUMF)
hist(ltUMF)
Model fitting – testing which model best fit your data
Several functions are available

(LcCm1_default <- distsamp(~ 1 ~ 1, ltUMF)) #
default same as below

(LcCm1_halfNorm <- distsamp(~ 1 ~ 1, ltUMF,
keyfun = "halfnorm", output = "density", unitsOut = "ha")) #

(LcCm1_hzrd <- distsamp(~ 1 ~ 1, ltUMF, keyfun = "hazard", output = "density", unitsOut = "ha")) #
Upon running the above codes /models - AIC values are produced to decide on best fit functn
For demonstration purposes this is what I did for African Pipit

\[
\text{PAmdetc}n \leftarrow \text{distsamp}\left(\sim (\text{avh} + \text{cover}) \sim 1, \text{ltUMF}\right)
\]

(PAm8 \leftarrow \text{distsamp}\left(\sim (\text{avh} + \text{cover}) \sim \text{Grazing}, \text{ltUMF}\right)
(PAm9 \leftarrow \text{distsamp}\left(\sim (\text{avh} + \text{cover}) \sim \text{Burning}\right)
(PAm10 \leftarrow \text{distsamp}\left(\sim (\text{avh} + \text{cover}) \sim \text{avh}, \text{ltUMF}\right)
(PAm11 \leftarrow \text{distsamp}\left(\sim (\text{avh} + \text{cover}) \sim \text{cover}, \text{ltUMF}\right)
• Multiple covariate distance analysis with program DISTANCE
  • example 1: data in intervals
  • example 2: exact data

• More about the basic assumptions of distance sampling
  • how to deal with possible violations

• General considerations about the modeling process
Distance analysis with program DISTANCE

DISTANCE is a standalone Windows program for designing and analysing DS surveys. It is developed and maintained by a group based at the Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St Andrews.

- the same group maintains a number of DS R packages (primarily package Distance)
  - (note, this is a different implementation than unmarked)
- compatible with R on various levels; several of its components are in fact R modules
  - package readdst can port data and models directly from DISTANCE to R
Distance analysis with program DISTANCE

DISTANCE is a standalone Windows program for designing and analysing DS surveys. It is developed and maintained by a group based at the Centre for Research into Ecological and Environmental Modelling (CREEM) at the University of St Andrews.

DISTANCE vs R

Pro:
- easier entry and learning curve (learning R and DS at the same time can be daunting)
- better documentation
- convenient output, analysis and error checking options
- survey design functionality (using spatial data)

Con:
- R allows greater customization and better data handling
- development happens primarily with R packages and is then ported, thus DISTANCE can be expected to lag a little in development
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

- import data
  - store data in external file and import (using import wizard), rather than entering it directly in DISTANCE
- set up data filter and model definition
- run analyses as combinations of filters and definitions
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Basic model (halfnormal function, no covariates)
• detections are in 4 distance bands
Example analysis:

Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Hazard rate function, covariate: observer (*average of covariate levels*)
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Hazard rate function, covariate: observer (*covariate level 1*)
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Hazard rate function, covariate: observer (*covariate level 2*)
Example analysis:
Chaffinch data from multi-species bird surveys in kiwifruit orchards, New Zealand

Hazard rate function, covariate: observer (covariate level 3)
Covariates may be factorial (e.g., observer) or continuous (e.g., wind speed)
Covariates may be factorial (e.g. observer) or continuous (e.g. wind speed)

Keep in mind: all covariates that enter into this model fitting process are covariates of detectability, not abundance. Here we are only concerned with effects on the detection process, to derive a largely unbiased survey result.

Many covariates can of course affect both detectability and abundance (e.g. site identity, vegetation type, time of year, ...). However, modelling their effects on density/abundance itself is a separate (later) job, and not part of DS proper.

(unmarked combines these steps sequentially, but they are still different processes) 

perpendicular distance (m)
Another example:
Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

While the chaffinch data was collected in specific distance bands, the skylark detection distances (collected with a range finder) were analysed as exact measurements.

Pro:
• biases in detection process are easier to identify in high resolution data
• better model fit can be achieved

Con:
• requires a large number of detections to be feasible
• prone to overfitting
• you will likely end up partitioning the data into intervals anyway, to deal with biases (see below)

→ some partitioning into intervals / detection bands is generally a good idea; if not in the field then during the analysis stage
Allowing the detection function to fit more closely to the data

<table>
<thead>
<tr>
<th>Detection function</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>$1/w$</td>
</tr>
<tr>
<td>Half-normal</td>
<td>$\exp\left(-y^2/2\sigma^2\right)$</td>
</tr>
<tr>
<td>Hazard-rate</td>
<td>$1 - \exp\left(-\left(y/\sigma\right)^b\right)$</td>
</tr>
<tr>
<td>Negative exponential</td>
<td>$\exp\left(-ay\right)$</td>
</tr>
</tbody>
</table>

Extended form:  
detection function = key function + series expansion  
g(y) = key (y) [1 + series (y)]

Series expansions add **flexibility** to the shape of the key function

Cosine  
$\sum_{j=1}^{m} \alpha_j \cos\left(j\pi y / w\right)$

Simple polynomial  
$\sum_{j=1}^{m} \alpha_j (y / w)^{2j}$

Hermite polynomial  
$\sum_{j=2}^{m} \alpha_j H_{2j}(y / \sigma)$

where $y$ is distance and $w$ is truncation distance
Allowing the detection function to fit more closely to the data
Song thrush data

Basic model (halfnormal function, no covariates) + 2 series expansion terms
Another example:
Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

Basic model (halfnormal function, no covariates)
Another example:
Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

Basic model (halfnormal function, no covariates) + 5 series expansion terms
-> overfitting
Another example:
Skylark data from multi-species bird surveys on sheep and beef farms, New Zealand

Hazard rate function, no covariates
• partitioned into 8 m intervals
• Basic assumptions of distance sampling
• General considerations about model choice
Assumption #1:

All animals on the line are detected (i.e., detection probability at distance 0 is 1)

• depends on study subject; often an issue in shipboard surveys because ship may block field of view, and/or animals may dive
• in terrestrial surveys, can generally be assumed
• animals may move away from the line before observer passes that point; however, this does not result in a non-detection but in movement before detection (see below)

dealing with violations:

• this is tricky, because \( g(0) = 1 \) is a fundamental assumption of the fitting process
• can use a 2\textsuperscript{nd} observer to “guard the centerline” (may result in overestimation)
• there are procedural fixes but they aren’t pretty
• try not to violate this one
Assumption #2:

Animals are randomly and independently distributed

- three sources of bias:
  a) populations are clustered (flocks etc.) but individual detections are treated as independent
     - artificially shrinks the confidence interval of the detection estimate
  b) transects are not placed independently of gradients of density (e.g. roads)
     - can lead to strongly biased detection at specific distances
  c) transects are too close together

dealing with violations:

a) record not individuals but clusters + cluster size, then incorporate cluster size into the detection function
b) place transects either randomly, or across gradients of density
c) make sure that maximum detection ranges do not overlap between transect
Recording clusters instead of individuals

<table>
<thead>
<tr>
<th>Effort</th>
<th># samples</th>
<th>Width</th>
<th># observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>34955</td>
<td>71</td>
<td>80.0000</td>
<td>167</td>
</tr>
</tbody>
</table>

Model
Half-normal key, \( k(y) = \exp\left(-y^2/\left(2^2 A(1)^2\right)\right) \)
\( A(1) \) bounds = (0.000000, 0.10000007)

<table>
<thead>
<tr>
<th>Iter</th>
<th>L(1) (likelihood)</th>
<th>Parameter Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-649.744</td>
<td>30.6225</td>
</tr>
<tr>
<td>2</td>
<td>-649.744</td>
<td>30.6225</td>
</tr>
<tr>
<td>3</td>
<td>-649.635</td>
<td>30.9778</td>
</tr>
<tr>
<td>4</td>
<td>-649.496</td>
<td>31.9478</td>
</tr>
<tr>
<td>5</td>
<td>-649.494</td>
<td>32.0689</td>
</tr>
<tr>
<td>6</td>
<td>-649.494</td>
<td>32.1064</td>
</tr>
<tr>
<td>7</td>
<td>-649.494</td>
<td>32.1067</td>
</tr>
</tbody>
</table>

Results:
Convergence was achieved with 7 function evaluations.
Final L(1) (likelihood) value = -649.493081
Akaike information criterion = 1300.9877
Bayesian information criterion = 1304.0438
AICc = 1301.0184

<table>
<thead>
<tr>
<th>Estimate</th>
<th>%CV</th>
<th>df</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average cluster size</td>
<td>1.1911</td>
<td>4.62</td>
<td>156.00 1.0872</td>
</tr>
<tr>
<td>Half-normal/Cossins</td>
<td>1.1366</td>
<td>2.56</td>
<td>156.00 1.0808</td>
</tr>
</tbody>
</table>

Analysis Engine: NCDS - Multiple occasions distance sampling
Cluster size estimation method:
- Use size bias regression method
- Use mean of observed clusters
- Use size bias regression method if regression significant at an alpha-level of 0.15. Use mean of observers if not significant

Size bias regression method:
- Regress in(cluster size) against estimated g(x)
- Regress cluster size against estimated g(x)
- Regress in(cluster size) against distance a
- Regress cluster size against distance a

SECC: Statistics in Ecology, Environment and Conservation
Assumption #3:

Animals do not move before detection

- bias resulting from movement is **negligible if movement is random**
- movement in response to the observer can result in negative (**avoiding**) observer or positive (**attracted**) to observer bias in detectability
- particularly avoidance behaviour is quite common

dealing with violations:

- average out & bridge dips/spikes in the distribution by
  - partitioning data into **intervals**
  - using **models with a shoulder** (e.g. hazard-rate)
Indications of **observer avoidance** in the data
Assumption #4:

**Measurements are exact** (angles and distances)

- the effect of *random* errors is negligible, but *systematic* errors can introduce bias; this often happens with *rounding*
- rounding angles or distances to preferred values can result in *heaping*
- *angle rounding to zero* is particularly common

**dealing with violations:**

- *avoid dead reckoning* in the field by using *tools*: range finders, angle boards, etc.
- smoothen & average instances of rounding in the data by *partitioning into intervals*
Indications of angle **rounding to zero** in the data
Hazard rate (shoulder), partitioning into intervals
When to use distance sampling?

- **Census**
  (complete count)

- **Double count**
  (rapid large-scale counts, calibrated by a few small-scale censuses)

- **Distance sampling**
  (modelled detectability)
  + large improvement in accuracy by modelling detectability
  + much lower effort than census
  – requires reasonably high number of records (>60)

- **Index count**
  (count not corrected for detectability, may be sufficient for tracking relative changes over time)

- **Presence/absence**
  (e.g. Occupancy)

available number of samples / accuracy of abundance estimate
When to use **multiple covariate** distance sampling? (as opposed to basic DS)

-> this is a question of **level of detection function** vs **level of abundance/density estimate**

Assume e.g. that one species was surveyed at several sites, and you want an estimate for each site. There are three basic options:

a) Fit a **separate** detection function to each site dataset
   → most accurate outcome, best adaptation to different site parameters; **BUT** requires sufficient sample sizes at each site

b) Fit a **global** detection function using all pooled detection, then use this function on **site datasets** to get site-level estimates
   → may save your bacon if sample sizes at individual sites don’t support fitting a function; **BUT** assumes that detection does **not vary** (substantially) between sites

c) Use **multiple covariate distance sampling**: fit a global detection function, specifying site as a **covariate** (i.e., allow the global function to vary for data subsets defined by covariate levels)
   → this is halfway between the other two options in requirements and result accuracy

The same choice applies to any covariate (canopy cover, season, sex, etc.)
→ this is essentially a **model selection question**; if the choice isn’t clear, AIC can be used to help choose between approaches (as all three use the same data set)
Useful resources

• The Distance project website http://distancesampling.org/
  • software downloads (R packages & program Distance)
  • Distance mailing list / Google group
  • huge distance sampling article bibliography

• Books (by Steve Buckland et al.):
  • Introduction to Distance Sampling: Estimating Abundance of Biological Populations. (2001) Oxford University Press. (in Main Library)

• Program Distance documentation (very useful)