Convenient analysis of numerous distance sampling data sets in R

Eric Rexstad, David L. Miller, Lindesay Scott-Hayward

Centre for Research into Ecological and Environmental Modelling, University of St. Andrews

University of St Andrews FOUNDED 1413

Challenges of analysis of archival data

- As questions about animal populations become more complex, so do the data and analytical requirements,
- Few of us analyses single species, collected at a single point in time,
- It is increasingly common to use archival data (defined as data collected for one purpose but used for another purpose) to assess spatial and/or temporal change in animal density.

We describe a generic set of tools to permit thoughtful analysis of archival data. This involves exploratory data analysis, collaborative decisions about analysis steps, reproducible audit trail of analysis steps. As described in an ISEC2010 plenary talk by Jeff Laake, reproducible research is a desirable goal of any analysis project. However with multiple interested parties, reproducibility climbs the priority list.

Purpose of analysis

Produce seasonal density surfaces of seabirds combining aerial and shipboard surveys for 41 species groupings.
Large volume of data collected (>2000000 records), requirement to produce density surfaces for species $\times$ season combinations. Detection functions to adjust for incomplete detectability created for each survey platform (planes and boats).
Visual inspection of the distribution of detection distances was necessary before analysis because:

- detections for some season/platform combination were sparse,
- geographic coverage for some season/platform combinations was poor, or
- distribution of detection distances was not conducive to detection function modelling.

Familiarity with data

- If analyst = data collector, this step happens by default.
- Otherwise, exploratory data analysis is more elaborate.
- If there are other collaborators, exploratory analysis may be protracted.
  - Analysis decisions may involve not only data provider and analysts but other parties who commissioned the analysis.
  - This is made simpler through the use of *analysis notebook* that can be shared among investigators.
  - Audit trail is produced to permit tracing of evidence used in analysis decisions.

Inventory phase of EDA

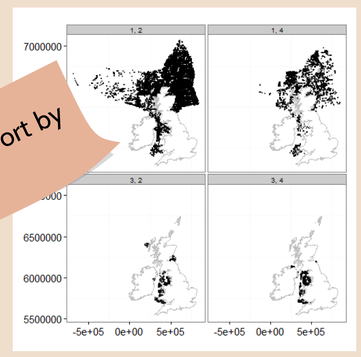
Archival data are high dimensional. There are requisites necessary for robust estimates of animal density derived from distance sampling methods. These requisites include adequacy both of survey ***effort*** and ***sightings*** mutually in the dimensions of

- species
- time (year or season)
- geographical area
- In a perfect world, there would be a balanced design in which there were sufficient data for all combinations of species $\times$ season $\times$ region, but data are seldom perfect
- Some combinations are likely to be missing or under-represented,
- With insufficient effort or detections, estimation of animal density will be fruitless at best and misleading at worst.

Example EDA results

We developed R Markdown code to perform the exploratory analyses and prepare a notebook page for each of the 41 species groupings of interest.

1 NORTHERN FULMAR Data (F.)



Spatial effort by season

Figure 1: Detections by season (2=summer, 4=winter) and by platform (1=boat, 3=aerial)

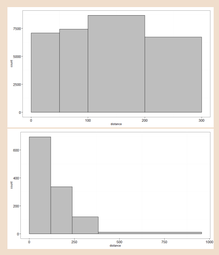
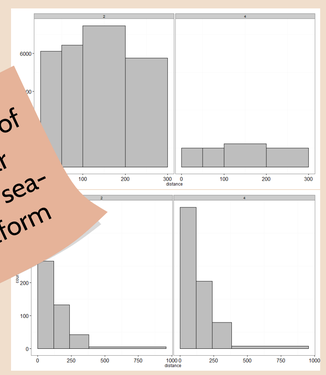


Figure 3: Detections by platform (top=boat, bottom=aerial)

Table 1: Table of number of detections by season and by platform

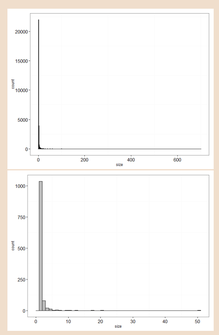
	Season 2	Season 4	Total
Boat	25752	4208	29960
Plane	444	719	1163
Total	26196	4927	31123

Sightings by season $\times$ platform



Distribution of perpendicular distances season $\times$ platform

Figure 2: Detections by season (2=summer, 4=winter) and by platform (top=boat, bottom=aerial)



Distribution of observed cluster sizes

Figure 4: Group size by platform (top=boat, bottom=aerial)

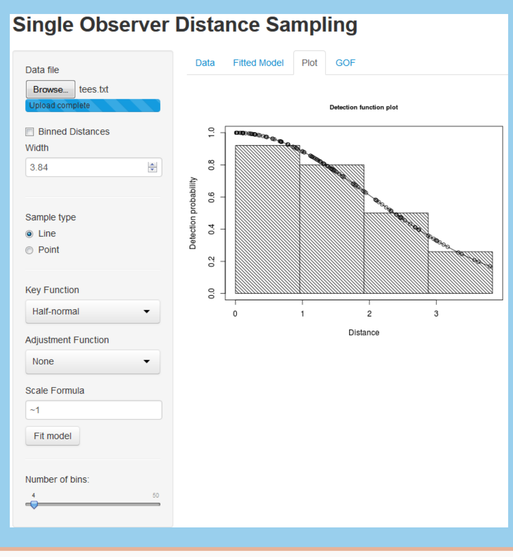
"Industrial-scale" analysis

- Facilitated by use of detection function fitting routines in Distance R package (Miller 2014)
- We visually assessed perpendicular distance distributions for all species $\times$ season $\times$ platform combinations
  - The appearance of the perpendicular distance distributions was placed into one of three categories:
    - shape did not support fitting of a detection function
      - remove that species $\times$ season $\times$ platform combination
      - treat as strip transect
    - shape supported fitting after truncation
    - shape could have detection function fitted to untruncated data
- After categorisation, it was straightforward to perform "bespoke" modelling for the 164 species $\times$ season $\times$ platform combinations.

Web-based analysis

The next step in collaborative analysis between data provider, analyst and funding agent is placing the entire enterprise on the web.

Some of this can be done using the Shiny interface for R. Jeff Laake, National Marine Fisheries Service in the U.S. is developing such an interface for components of Distance.



Fitting detection functions using Distance

Distance is an uncluttered interface to the MRDS R package for fitting detection functions. Data that contain rudimentary fields (region label, region area, transect length, perpendicular distance) are amenable to analysis without extensive data preparation.
The multi-species seabird database was organised such that fitted detection probabilities can be extracted, following the EDA with code such as

```
library(Distance)
pr.detect <- numeric(0, length=3)
for (species.name %in% c("FULMAR","GANNET","SCOTER")) {
pr.detect[i] <- ds(data[species==species.name,],
truncation="0%", key="hr",
adjust=0)$dht$individuals$average.p
}
```

If there is a balanced design (adequate effort and sightings for all combinations of strata), the "split-apply-combine" strategy can be employed to conduct wholesale analyses.

```
models <- dlply(data,.(Region.Label),ds)
sapply(models, plot) # view fitted models
count.adjustment <- sapply(models, function(x)
x$dht$individuals$average.p) # p_i for this stratum
```

To learn more

Distance package (CRAN)

Distance Shiny app (jlaake.shinyapps.io/Distance)

distancesampling.org (New Distance website)

International Statistical Ecology Conference 1 - 4 JULY 2014 - MONTPELLIER - FRANCE

University of St. Andrews is a charity registered in Scotland: SCO13532

LaTeX TikZposter