Mark-Recapture Distance Sampling using Distance

Practical 8, Intermediate Distance Sampling workshop, CREEM, 2018

This version of the practical is for those who would like to conduct the analysis in Distance (Thomas *et al*. 2010). There is a separate version for conducting the analysis in R.

The first part of this practical involves analysis of a survey of a known number of golf tees. This is intended mainly to familiarise you with the double-platform data structure and analysis features in Distance. The second part of the practical involves analysis of the pack-ice seal survey data of Borchers *et al*. (2006) and Southwell *et al*. (2007).

I. Golf Tee Survey

The Golf Tee Data and Distance Project

The data come from a survey of clusters of golf tees in grass, conducted by 3rd and 4th year statistics students at the University of St Andrews. For the purposes of this exercise, we will assume that all the data were collected on one 210 metre long transect line out to a width of 4 metres (both sides), and that this comprises the study region. There were 250 clusters of tees in the study region and 760 individual tees in total.

The population was independently surveyed by two observer teams. The following data were recorded for each detected group: perpendicular distance, cluster size, observer (team 1 or 2), "sex" (males are yellow, females green and golf tees occur in single-sex clusters), and "exposure". Exposure was a subjective judgement of whether the cluster was substantially obscured by grass (exposure=0) or not (exposure=1). The lengths of grass varied along the transect line, and the grass was also slightly more yellow along one part of the line compared with the rest.

The data are stored in the distance project **GolfteesExercise.dst** or **GolfteesExercise.zip**. Open the project, and click on the **Data** tab to see how the data are stored. Notice that each detected cluster appears twice in the observation layer – once for observer 1 and once for observer 2. There is a field "object" which gives a unique number to each cluster and another field "detected" which indicates whether the cluster was seen (=1) or missed (=0) by each observer. There are also fields for the other covariates: perpendicular distance, cluster size, sex and exposure. In general, covariates could also be stored in other data layers (e.g., stratum- or transect-level covariates such as habitat).

Now click on the **Survey** tab and open the **Survey details** for "New survey". Click on **Properties** to see the survey's properties. Notice that the observer configuration is set to Double observer. We will analyse these data assuming that Observer 2 was generating trials for Observer 1 but not *vice versa* – i.e., trial configuration, where Observer 1 is the primary and Observer 2 is the tracker. (The data could also be analyzed in independent observer configuration – you are welcome to try this yourself.) Click on the **Data fields** tab, and notice that the Object, Observer and Detected roles are all filled by the appropriate fields – this is done by default when you are setting up a new project, if you tell Distance it is a double observer survey in the Setup Project Wizard.

Lastly, click on the **Analyses** tab. Two Analyses have been set up for you:

1. "FI – MR dist" is a model with only perpendicular distance as a covariate in the mark recapture detection function model.
2. "FI – MR dist + size + sex + exp" is a model with distance, cluster size, sex and exposure as covariates in the mark recapture detection function model.

Have a look at the **Properties** of the **Model Definition** for each of these models, and familiarize yourself with the contents of each tab. The difference between the two Model Definitions that have been set up for you is that they have different MR (mark recapture) models – see the MR Model button in the Detection Function tab. Notice that the names used in the model formulae are not the same as the field names in the Data tab – for example the "Perp distance" field is called "distance" in the formulae. (Recall from the lecture that some fields get renamed when the data are sent out to R for analysis. You can find out what renaming has been done in the Log tab once you've run an analysis – we'll do this in a second.)

Golf Tee Survey Analyses

1. Estimation of *p*(0): distance only

Run the "FI - MR dist" analysis. Once it is run, look in the **Analysis Details Log** tab, and find the part of the log where Distance tells you how it has renamed the fields. It's useful to know where this is in case you need reminding of what names to use when specifying formulae. Now look in the **Results** tab. Does the fit of the model look good?

Here's what the plots are:

a. The plots on the pages headed "Summary 1" and "Summary 2" show the histograms of distances for detections by either, or both, observer. The shaded regions show the number for observer 1 and observer 2, respectively.
b. The plot on the page 'Summary 3' shows the histograms of distances for duplicates (detected by both observers).
c. The plot on page headed 'Summary 4' shows the histograms of distances for observer 2. The shaded regions indicate the number of duplicates – for example, the shaded region is the number of clusters in each distance bin that were detected by Observer 1 given that they were also detected by Observer 2 (the "|" symbol in the plot legend means "given that"). [Note that if an 'io' configuration had been chosen, there would also be a plot showing the duplicates overlaid onto distances detected by Observer 1; in a 'trial' configuration we are only interested in Observer 2 setting up trials.]
d. Q-Q plot – this has exactly the same interpretation as a Q-Q plot in single platform analyses (please ask one of the instructors if you can't remember how to interpret them!)
e. The plot on the page headed "Detection probability 1" shows a histogram of Observer 1 detections with the estimated Observer 1 detection function overlaid on it. The dots show the estimated detection probability for all Observer 1 detections.
f. The plot on the page headed "Detection probability 2" shows the proportion of Obs 2's detections that were detected by Obs 1 (also see the "Summary" page). The fitted line is the estimated detection probability function for Obs 1 (given detection by Obs 2). Dots are estimated detection probabilities for each Obs 1 detection.

Is there evidence of unmodelled heterogeneity? What do these results tell you about *p*(0)?

2. Estimation of *p*(0): distance and other variables

Run "FI - MR dist+size+sex+exp". Can you explain the differences between the *p*(0) estimates and the abundance estimates between the two models? Which model would you use to estimate abundance? To decide, look at the goodness-of-fit test and AIC values from each model.

3. Specifying new models

The size covariate is the least significant of the covariates in the model "FI – MR dist + size + sex + exp" – its estimate is 0.078 with SE 0.183. So try creating a new model definition and analysis without this covariate. Does it have a lower AIC?

You can also try some models with interaction terms. For example, you would specify the interaction between sex and exposure as "sex:exposure". If you also want both sex and exposure in as main effects (which you usually do) then the notation "sex*exposure" is shorthand for "sex + exposure + sex:exposure". Don't try too many of these models – leave time for the next part!

4. Point independence

All the models we have tried so far assume full independence – i.e. that the detections are independent between platforms at all distances. A less restrictive assumption is point independence – that the detections are only independent on the line.

Let's start by seeing if a simple point independence model is better than a simple full independence one. Set up a new Analysis and Model Definition, and under Detection Function | Method, choose Trial,

Point independence. For the MR (mark-recapture) model, specify distance as the only covariate. You now also need to specify a DS (distance sampling) model. Start with a half-normal key function and constant scale parameter (i.e. no covariates).

Run this model, and compare it with the corresponding full independence model (i.e., the full independence model with the same MR model). Which has the lower AIC? Which has an estimate closer to the known true abundance?

Now try a point independence model that has a MR model the same as the MR model from your full independence analyses. Which has the lower AIC and bias? Finally, try some models where you introduce covariates into the scale parameter of the DS model (for example, try sex as a covariate).

What is your final best model? What is the estimate of $p(0)$ for this model? Was all this modelling necessary in this instance, given the value of $p(0)$? How else could you have obtained a robust estimate of abundance?

II. Crabeater Seal Survey

The Crabeater Seal Data and Distance Project

These data come from a helicopter survey of crabeater seals within pack-ice in the Antarctic conducted by the Australian Antarctic Division as part of their pack-ice seals program[1]. The helicopter can only operate within a relatively short distance from the icebreaker, which acts as its base, and the ice-breaker can only go where the pack-ice is thin enough, the aerial transects cannot be randomly located. This means that design-based abundance estimation was not a valid option – abundance was estimated using density surface modelling. In this exercise, we concentrate on detection function estimation.

The data are stored in two distance projects: **CrabbieMCDSExercise** and **CrabbieMRDSExercise**. The first contains a multiple-covariate distances analysis of the data (assuming $p(0)=1$); the second contains mark-recapture distance sampling analyses of the same data.

In addition to distance and cluster size, the following explanatory variables are available for modelling detection probability:

side:	the side of the helicopter from which the seals were seen
exp:	the experience (in survey hours) of the observer
fatigue:	the number of minutes the observer has been on duty on the current flight
gscat:	group size category (1/2/≥3)
vis:	visibility category (poor/good/excellent)
glare:	Yes or No, depending on whether there was glare or not
ssmi:	a measure of ice cover
altitude:	the height of the aircraft in metres
obsname:	individual observer identifier

The distance projects are set up as if the transects were random and survey area was 1,000,000 hectares. This is just a device to get Distance to produce an abundance estimate, which we can treat as a relative index of abundance.

The Crabeater Seal Analyses

The analyses described in Borchers *et al*. (2006) and Southwell *et al*. (2007) use the point independence method. Use the analyses in CrabbieMCDSExercise and CrabbieMRDSExercise distance projects to decide whether:
(a) an MCDS analysis would have been adequate (and if so, why), and
(b) a full-independence MRDS analysis would have been adequate (and if so, why).

[1] http://www.aad.gov.au/default.asp?casid=1164

The projects contain the following analyses (which have already been run, since running them takes a while):

1. MCDS like paper: MCDS model used for the MCDS component of the analysis in the papers.
2. PI dist only: Point independence model using only distance
3. FI like paper: Full independence model using the MRDS model used in the papers.
4. PI as in paper: Point independence model using the same MCDS model and MRDS model as used in the papers.

If you have time, try other models and see if you can do better than the model "PI as in paper".

References

Borchers, D.L., Laake, J.L., Southwell, C. and Paxton, C.G.M. (2006) Accommodating unmodeled heterogeneity in double-observer distance sampling surveys. *Biometrics* **62**: 372-378.

Southwell, C., Borchers, D.L., Paxton, C.G.M., Burt, M.L., de la Mare, W. (2007) Estimation of detection probability in aerial surveys of Antarctic pack-ice seals. *Journal of Agricultural, Biological & Environmental Statistics* **12**: 41-54.

Thomas, L., Buckland, S.T. Rexstad, E.A. Laake, J. L., Strindberg, S., Hedley, S.L., Bishop, J. R.B., Marques, T. A. and Burnham, K. P. (2010) Distance software: design and analysis of distance sampling surveys for estimating population size. Journal of Applied Ecology 47: 5-14. DOI: 10.1111/j.1365-2664.2009.01737.x