Distance sampling workshops 2019: Data format

Introduction

One of the goals of the 'Advanced distance sampling' workshop is to enable people who have already collected distance sampling data to do some preliminary analysis of their data using the 'Distance' package (Miller *et al*. 2019b) in R (R Core Team 2019). This page explains how to get your survey data into a format that 'Distance' can easily read for conventional distance sampling and also density surface modelling.

Please note that there will be little time during the 'Introductory distance sampling' workshop for participants to work with their own data, but the details on the basic data format are provided for information.

Basic data format for distance sampling

The data below illustrates a simple example of how survey data should be arranged. The first line of the data contains the column names.

Region.Label, Area, Sample.Label, Effort, distance
Stratum 1,100,Line 1,10,14
Stratum 1,100,Line 1,10,8
Stratum 1,100,Line 1,10,22
Stratum 1,100,Line 2,10.3,7
Stratum 1,100,Line 2,10.3,37
Stratum 1,100,Line 2,10.3,13
Stratum 2,123,Line 1,5.7, NA
Stratum 2,123,Line 2,8.4,27
Stratum 2,123,Line 2,8.4,76
Stratum 2,123,Line 2,8.4,44
Stratum 2,123,Line 2,8.4,7

In the data above, the columns are separated by commas:

- column 1 is the region (or stratum) name (Region.label)
- column 2 is the region area (Area)
- column 3 is the transect name (Sample.Label)
- column 4 is the transect length (Effort)
- column 5 is the perpendicular distance (distance).

Notice that all transects from the same stratum are grouped together on adjacent lines, and all observations from the same transect are grouped together (this is important: sort your data by transects within region). Notice also that the record "Line 1" in "Stratum 2" has no distance in the final column – this is a transect along which no objects were seen – and so 'NA' indicates there were no detections on that transect.

Which columns should be included in your data file? As a minimum, the file should contain a column for line transect, or point transect, name (i.e. Sample.Label) and a column for observed distance (i.e. distance). For line transect surveys you will also need a column for transect length (Effort). If your survey involved stratification then you will need to include columns for stratum name and stratum area. If your objects are

in groups or clusters, rather than individuals, then you should include a column for cluster size (called 'size').

The columns should be separated by a delimiter (ASCII character), which can be either a tab, semicolon, comma or space; R can handle most formats (including excel). For ease of reading it helps if the columns are in some logical order, as above. Each row should finish in a Carriage-return + Line-feed combination. Excel file can easily be exported to delimited text files (although some formats of dates can cause problems).

The description above is for the most basic data structure to be imported. Additional columns of data that you may wish to use in your analysis (such as year of survey in multi-year surveys) can easily be included. To keep maximum flexibility, it is best to bring your data along as a text file, as outlined above, but also in its original spreadsheet or database format in case you decide to take on a more complex analysis in a later part of the workshop.

With software such as a spreadsheet, it is easy to arrange, sort and filter columns and export them into a text file that Distance can read. Alternatively, if you are bringing along your own computer then you can use your favourite software to do the required re-formatting.

If you are still confused about what to do, don't worry – we will be on hand to help when you get to the workshop. Just bring some of your data along in an electronic format

Data format for density surface modelling (Advanced topic)

One of the topics to be covered in the 'Advanced distance sampling' workshop is density surface modelling (Miller *et al*. 2013) and this requires data in a particular format. This description is taken from the 'dsm-data' help file in 'dsm' package – this R package will be used extensively during the 'Advanced' workshop (Miller *et al*. 2019a).

Two sets of data must be provided to 'dsm'. They are referred to as 'observation data' and 'segment data'.

The segment data table has the sample identifiers which define the segments of search effort, the corresponding effort (segment length) expended and the environmental covariates that will be used to model abundance/density.

Observation data provides a link table between the observations used in the detection function and the samples (segments), so that we can aggregate the observations to the segments (i.e. observation data is a "look-up table" between the observations and the segments).

Observation data - the observation data must have (at least) the following columns:

- object – unique object identifier
- Sample.Label – the identifier for the segment that the observation occurred in
- Distance – perpendicular distance to observation
- Size – the size of the observed group or cluster (this will be 1 if objects occurred individually)

The observation data can be used to fit a detection function and so additional columns for detection function covariates are allowed in this table.

Segment data - the segment data table must have (at least) the following columns:

- Sample.Label – unique identifier for the segment
- Effort – the length of the segments
- ??? – environmental covariates, for example, location (projected latitude and longitude), and other relevant covariates (sea surface temperature, foliage type, altitude, bathymetry etc).

The columns should be separated by a delimiter (ASCII character) as described in the basic data format section above. QGIS can export databases to comma delimited text files.

References

Miller DL, Burt ML, Rexstad ER and Thomas L (2013) Spatial models for distance sampling data: recent developments and future directions. Methods in Ecology and Evolution 4: 1001-1010

Miller ML, Rexstad E, Burt L, Bravington MV and Hedley S (2019a) dsm: Density Surface Modelling of Distance Sampling Data. R package version 2.2.17. https://CRAN.R-project.org/package=dsm

Miller DL, Rexstad E, Thomas L, Marshall L, Laake JL (2019b) Distance Sampling in R. Journal of Statistical Software, 89(1), 1-28. doi: 10.18637/jss.v089.i01 https://doi.org/10.18637/jss.v089.i01

R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.