# Convenient analysis of numerous distance sampling data sets in R

**E. Rexstad** and **D. Miller** and **L. Scott-Hayward**

Centre for Research into Ecological and Environmental Modelling
University of St. Andrews
St. Andrews Scotland KY16 9LZ
eric.rexstad@st-andrews.ac.uk, dave@ninepointeightone.net, lass@st-andrews.ac.uk

**Keywords:** distance sampling, big data

**Abstract:** Institutions such as governmental agencies and non-governmental organisations attempt to make greater use of historical data they have accumulated. Statisticians are called upon to analyse data sets of increasing size with greater rapidity. The gulf between data gatherers and data analysts is very broad in these circumstances and challenges the opportunity for interactive data analysis. We are developing reproducible research (Gandrud 2013) tools to couple the convenience of R for data manipulation with the Distance package (Miller 2013).
Our tools perform exploratory analysis of data sets and produce a laboratory notebook style digest of preliminary results. The analyst and client then examine the preliminaries to determine the course of action (that may include merging of datasets, degree of truncation or elimination of problematic datasets) for final analyses.
The Distance R package provides a convenient way of fitting detection functions to sighting data. Models involving double observer protocols or covariates influencing detectability are easily incorporated. The detection function fitting routines provide metrics for model selection. These adaptable tools conveniently fit inside many types of analytical systems, we present a system tailored for reproducible methodology when analysing a large number of data sets.
Tools used for the final analysis include a series of standardised graphics and goodness of fit diagnostics. This provides uniformity across data set and allows analyst and client to have a consistent means of examining findings. This addresses the common concern of distance sampling novices "how do I present the results."
We demonstrate use of these tools with a 30-year dataset for a multi-species survey consisting of $\geq$30 species of sea birds. We also discuss the potential to extend the set of tools in a web-based environment to heighten the collaborative aspects.

## References

Gandrud, C. (2013) Reproducible Research with R and RStudio. CRC Press.
Miller, D.L. (2013) Package Distance reference manual.
`http://cran.r-project.org/web/packages/Distance/Distance.pdf`