14. Data analysis in DISTANCE

1. Analysing your data

Open DISTANCE by double clicking on your **.dst** file. In the Project browser window, select the **Analyses** tab, under which you set up and store the results of your analyses. On the toolbar you have buttons for:

Set – for organising your analyses;
Analysis – for creating, deleting, selecting details, running and moving your analyses.

These functions can also be accessed from **Analyses** in the main **DISTANCE** toolbar.

Under the Project browser toolbar are two panels. The left panel stores names, etc of different analyses. The right panel stores summary diagnostics and results of the analyses.

To set up your first analysis, select **Analysis > Analysis Details** from the **Project Browser** toolbar or click the third icon from the left next to **Analysis:** on the **Analyses** tab toolbar. This brings up the Analysis Details screen. There are three tabs on the right side – you are in the **Inputs** tab.

IMPORTANT POINT ABOUT DATA FILTERS AND MODEL DEFINITIONS

An important concept in DISTANCE analysis is the independence of two separate parts: data selection through a **Data filter** tab; and defining your analysis through a **Model definition** tab (see below). For a particular set of data selected under the **Data filter** tab, you can run multiple different models set up under the M**odel definition** tab. Equally you can run any particular model (defined under the **Model definition** tab) on multiple sets of data (defined under the **Data filter** tab).

In this **Analysis** window, you can see sections for the **Data filter** and the **Model definition**.

1.1 Data selection (under Data filter*)*

First, you need to select which data to analyse by creating a **Data filter**, which is the panel in the middle of the screen. Click on **Properties …** on the right side of this panel to bring up the **Data Filter Properties** window. There are four tabs.

Under the **Data selection** tab you tell DISTANCE which data to analyse. For this first analysis, use all the data so no selection is necessary and you can ignore this and all the other tabs for now but give your Data filter a **Name**, e.g. “All data”, and click **OK**.

1.2 Detection function model (under Model definition*)*

Now define the model for fitting the detection function to estimate average detection probability.

Click on **Properties …** on the right side of the **Model definition** panel at the bottom of the screen to bring up the **Model Definition Properties** window. At the top, under **Analysis Engine**, select the kind of analysis you want to do. The choices are:

CDS – Conventional distance sampling	Basic line transect analysis
MCDS – Multiple covariate distance sampling	Allows inclusion of covariates when fitting the detection function
MRDS – Mark-recapture distance sampling	Uses double team data to estimate g(0) and account for responsive movement
DSM – Density surface modelling	Model-based abundance estimation

Select **CDS – Conventional distance sampling** for this analysis.

This window has six tabs.

First select the **Estimate** tab. Under **Stratum definition**, select **Use layer type**; the only option for Layer type is *Stratum*. Under **Quantities to estimate and level of resolution**, check the *Stratum* box at the top to check all the other quantities. If you leave these all checked, DISTANCE will estimate Density, Encounter rate, Detection function and Cluster size separately for each stratum.

Encounter rate must be estimated separately for each *Stratum*. However, because detection probability is usually unrelated to Stratum, the Detection function is often estimated for all data pooled (i.e. at the *Global* level). Check the box under column *Global* for Detection function.

Leave Cluster (group) size at the *Stratum* level, but it is also possible to estimate it at the *Global* level.

Leave Density checked at the *Global* as well as *Stratum* level. DISTANCE will then calculate an overall estimate of Density as well as for each Stratum. The *Global density estimate* should be the **Mean** *of stratum estimates*, *weighted by* **Stratum area**.

If you had a single survey area (no strata) or wanted an unstratified estimate, you leave **Stratum definition** as **No stratification** and all estimates are made at the *Global* level.

Select the **Detection function** tab, which sets up how to fit the detection function. It has four tabs.

Use the **Models** tab to select the form of the detection function from a range of **Key functions** (basic models) and **Series expansions** (the form of the adjustment function). Keep the default **Half-normal** key function with **Cosine** series expansion for this analysis.

Use the **Adjustment terms** tab to select how to add the series expansion terms. For this analysis, leave this on **Automated selection** to let DISTANCE decide how many, if any, adjustment terms are needed for the best fitting model, using AIC as the model selection criterion.

Ignore the **Constraints** and **Diagnostics** tabs.

Select the **Cluster size** tab in the row above.

In line transect sampling of animals that occur in groups, there is a tendency to miss smaller groups at greater perpendicular distances, which can lead to bias in the estimation of average group size. This bias can be accounted for by regressing group size (or log group size) against perpendicular distance (or against estimated detection probability).

DISTANCE gives you several options; choose the third one: **Use size bias regression method if regression significant …** . Leave the alpha-level at 0.15. Under **Size-bias regression method**, choose the third one: **Regress ln(cluster size) against distance x**.

Ignore the other tabs but give your Model definition a **Name**, e.g. “Half-normal”, and click **OK**.

The last step is to give your Analysis a **Name** at the top of the Analysis Details screen. Use something short-hand but informative such as: “All data HN” (to represent the analysis of all data using the Half-normal detection function).

Run your analysis using the button in the top right corner of the screen.

If there are any warnings, the **Log** tab will be coloured orange. Warnings generally do not invalidate your analysis and can usually be ignored.

For example, you might get a warning saying:

** Warning: Parameters are being constrained to obtain monotonicity. **

This relates to ensuring that the detection function has the right shape (monotonically decreasing – i.e. detection probability cannot increase as perpendicular distance increases). This is OK and the warning can be ignored.

Depending on the size of your dataset (number of observations), you might also get a warning saying something like:

** Warning: only 500 out of a total of 549 observations will be shown on the qq plot. **

This is the DISTANCE default; to see all your data on the QQ plot, select the **Input** tab, open the *Model definition* **Properties** window, select the **Detection** tab and then the **Diagnostics** tab. Type in, say, 1000 for *Maximum num points in qq plots*, and click OK.

The **Confirm Change** window appears, as it does whenever you modify a *Model definition* that is used by an analysis that has already been run. This is to make sure you haven't made the mistake of modifying an existing *Model definition* instead of creating a new one (easily done).

In this case, we do want to change the *Model definition* so select **Yes**.

Return to the **Project Browser** by selecting **View > Project Browser**, or closing the Analysis Details screen (don't close DISTANCE).

You'll see that the dot at the left hand side for this analysis is grey. This means you need to run the analysis again (because you changed the *Model definition*). This will also be the case for any other analysis that uses this particular *Model definition*.

Select **Analyses > Run analysis** (or the yellow button with a running man) to run the model again.

If there are no warnings, the **Results** tab will be green. If your analysis fails, the **Log** tab will be red. The dot at the left hand side of the analysis on the **Project Browser** will be the same colour.

1.3 Results

The **Results** tab contains several pages (scrolled from the top of the screen) including:

- The options you selected for analysis
- Details of the model fitting and the best model selected, including summary input data
  - At the top of pages 2 and 3, Effort is the total transect length, # samples is the number of transects, Width is the maximum perpendicular distance in your data (or the truncation width – see later), # observations is the number of sightings.
- Results for the best-fitting model, including average detection probability (p) and effective strip half width (ESW)
- Diagnostics to tell you how well the detection function model has fitted the data
  - QQ plot
  - Kolmogorov-Smirnov (K-S) and Cramer-von-Mises goodness of fit (GoF) test results
  - Plots of the fitted detection function together with histograms of the observed data and Chi-squared GoF test results for three different histogram "bin" widths
- For each Stratum, cluster (group) size regression results and a (crude) plot of the regression line
- For each Stratum, estimates of density of groups (DS), group size (E(S)), density of animals (D) and abundance (N), together with their SE, CV and 95% confidence intervals
- Summary pages, including the pooled abundance estimate on the final page

When you've looked through these, return to the **Project Browser** to see the summary for your analysis.

First, change the name of the **Set** to something like "All data".

In the left panel, the colour of the dot indicates whether the analysis ran without warnings (green), ran with warnings (orange), failed (red) or has not yet been run (grey). Hover the mouse pointer over the "1" in the 4th column to see the name of the Data Filter used in this analysis. Do the same with the next column to see the name of the Model Definition used in this analysis.

In the right panel there are columns for:

- # parameters – the number of parameters in the fitted detection function
- Delta AIC – the difference in AIC between this model and the best-fitting model
- AIC – Akaike's Information Criterion – a measure of relative model fit
- ESW/EDR – Estimated effective strip half width
- D, D LCL, D UCL, D CV – Density of animals with lower/upper confidence limits and coefficient of variation

You can customize this panel by adding/removing columns using **Analysis > Arrange columns** (or the far right button on the toolbar). HOWEVER, a better way to do this is to set these columns as a default so you don't have to modify them every time you do a new analysis. To do this, select **Analyses > Preferences …** under the DISTANCE toolbar to bring up the **Distance Preferences** window. Under *Analysis browser window*, click the *Analysis Browser* button. Now add additional columns (or remove columns you don't want) using the arrows to move them from right to left (or left to right).

Useful columns to add include:

- **GOF K-S p** and **GOF CvM (cos) p** (useful diagnostics)
- **P**, **P CV** (detection probability)
- **DS**, **DS CV** (density of groups)
- **CS**, **CS CV** (group size)
- **N**, **N CV**, **N LCL**, **N UCL** (abundance)

The **D CV** is the same as **N CV** so you may wish to remove **D CV** (and also **D LCL**, **D UCL**).

2. Extending the scope of your analysis

There are a lot of possible ways to extend your analysis to improve your results. In the first instance we will look at:

- Truncating your perpendicular distance data to improve model fit
- Trying other models for the detection function

2.1 Truncating perpendicular distances

Data for sightings a long way from the transect line are not very informative but they may adversely affect the fit of the detection function and increase variance. It is worthwhile, therefore, examining the distribution of perpendicular distances to see if truncating the data could improve model fit.

Bring up the details for your first analysis and look at the **Results** page *Detection Fct/Global/Plot: Detection Probability 3*. If the distribution has a long tail you could discard the observations with the largest values. You could choose to discard observations greater than a certain perpendicular distance or you could discard a given percentage of the largest perpendicular distances. Try this in your next analysis to see what happens.

Select **Analyses > New Analysis** (or the first icon next to **Analysis:**) to create a new analysis; its name will be the same as the previously highlighted one with "1" at the end. Bring up the Analysis details screen. Select **New …** next to the **Data filter** panel to bring up the **Data Filter Properties** window. Select the **Truncation** tab.

We want to right truncate the data. Decide if you want to discard perpendicular distances larger than a certain value (e.g. 5m) or a certain percentage of perpendicular distance (5% is one recommendation). Give your new Data filter a name, e.g. "All data trunc 5m" or "All data trunc 5%".

Keep your **Model definition** the same, give the new analysis a name, e.g. "All data trunc 5% HN", **Run** the analysis and check the results. Diagnostics to look at include: the QQ plot, the K-S goodness of fit test results, plots of the detection function overlaid on the data.

Question 1: Are your results better with the truncated data?

NOTE that you can only compare AIC values for models fitted to the same data set. Truncating changes the data so you cannot use the AIC as a diagnostic in this case. Instead, you can look at the Goodness of Fit diagnostics - QQ plots and the GOF K-S p and GOF CvM (cos) p; the higher the probability value, the better the fit.

Generally, it's a good idea to store analyses based on different datasets in different **Set**s so that model diagnostics, such as AIC, are comparable for all models within a **Set**. However, when you're deciding whether or not to truncate, you can leave them in the same set so you can compare the GoF diagnostics.

2.2 Other models for the detection function

So far we have used the Half-normal model for the detection function but there are others we could try to see if we get a better fit to the data.

Highlight your favoured Analysis in the left panel of the **Project Browser**, select **New Analysis** and bring up the Analysis Details screen. This time, select a **New Model definition**. Under the **Detection function** tab select the **Models** tab. Select **Hazard-rate** under **Key function**. Change the **Name** to "Hazard-rate" and click **OK**. Change the name of the analysis to something like "All data trunc 5% HR", **Run** the analysis and look at the results. See how the shape of the detection function is slightly different.

Look at the diagnostics again; this time we can also compare the AIC values with those from the analysis with the same truncation (or no truncation) but using the Half-normal detection function.

NOTE that models with AIC values that differ (Delta AIC) by less than 2 units effectively have the same support from the data, i.e. they fit equally well.

Question 2: Does the Hazard-rate model fit better than the Half-normal model?

We ran separate analyses for the two model forms for the detection function. An alternative is to do this in the same analysis by including both Half-normal and Hazard-rate in the same **Model definition**. Do this simply by creating a new Model definition and adding another model under the **Detection function** > **Model** tabs by clicking on the + button on the right. Don't forget to give your Model definition a new name, e.g. "Half-normal + Hazard-rate". And don't forget to give your Analysis a new name.

DISTANCE will then try both model forms and select the one giving the best fit to the data, based on AIC. Try this and you should get the same results as one of the previous analyses.