Introduction to Distance Sampling

Exercise 1: Line transect estimation by hand

1) Plot a histogram of the following duck nest data, and fit a detection function by eye.
From your histogram, estimate the proportion of nests within 2.4m of the line that are seen, P_a.
Hence estimate nest density D (number of nests per square meter or per square kilometer – be careful of units!).

n=534 nests. L=2575 km.

Perpendicular distance band (meters)	0.0-0.3	0.3-0.6	0.6-0.9	0.9-1.2	1.2-1.5	1.5-1.8	1.8-2.1	2.1-2.4
Frequency	74	73	79	66	78	58	52	54

Having produced your fit to the histogram, to assist in producing your estimate of nest density, fill in these blanks.

Area of rectangle =

Area under your fitted detection function =

$$P_a = \frac{area_{curve}}{area_{rectangle}} =$$

$$\hat{N}_a = \frac{n}{P_a} =$$

$$\hat{D} = \frac{\hat{N}_a}{a} = \frac{\hat{N}_a}{2wL} =$$

Complete only part 1) of this exercise until instructed to go further.

---

2) Now use your histogram to estimate the effective half-width of search μ. Again estimate nest density D. How does it compare to your estimate from part (a)?

3) Rescale the y-axis to make your curve into the probability density function $f(x)$. Read off $f(0)$, and again estimate nest density D. How does it compare with your previous estimates?

Introduction to Distance Sampling

Exercise 2: Line transect estimation using Distance: Ducknests

GETTING STARTED ON THE COMPUTER

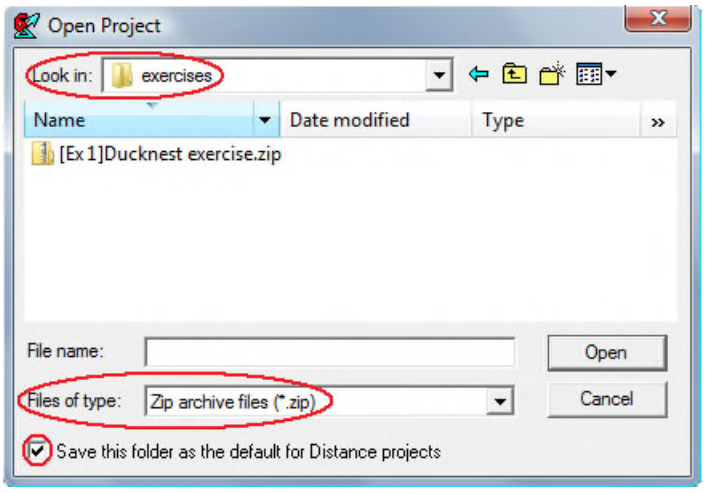
Click on **"Start"**. A list will be displayed. Click on **"Programs"**, then **"Distance"**. Now click on **"Distance 7.0"**. (Or double-click the **Distance 7.0** icon on the desktop.) This opens Distance.

1. Refer to the data from the graph paper exercise (Exercise 1). These data have been set up as a Distance project, which have been archived in a compressed (.zip) file on your thumb-drives. You should copy the subdirectory containing the Distance projects for the workshops to the My Distance Projects folder under My Documents, or to a location of your choice (to make it easier to find for subsequent exercises).

- Select **File** followed by **Open project**. Under "**Files of type**" choose **Zip archive files (*.zip).**

- Next to "**Look in:**", browse for the thumb drive directory (or wherever you have saved the exercises). **Note:** Distance includes some sample projects. The Sample Projects folder is the default folder Distance opens when instructed to open a new project, and contains a different duck nest data set, so make sure that you are looking in the right place!

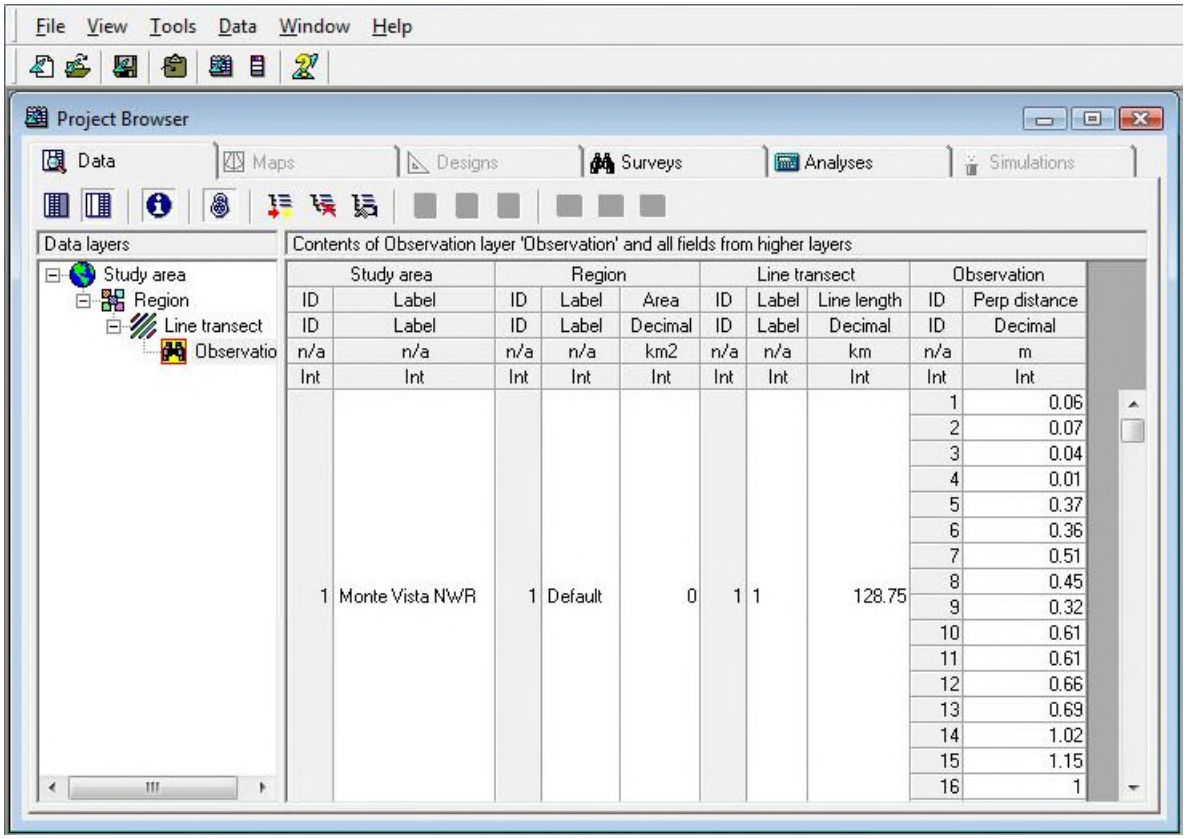
- You can change the default folder to one of your choice by checking the box next to "Save this folder as the default for Distance projects". If you do this, the next time you open a project, Distance will look in the folder you specified – containing all the exercises relevant to this workshop. The Sample Projects folder will still exist in Distance, and you may want to look at those projects at a later date.



- Double click on **Ducknest exercise.zip**. Click **OK** to unpack the project into the current directory and open it. Next time you open the project, you can open the file *Ducknest exercise.dst* directly.

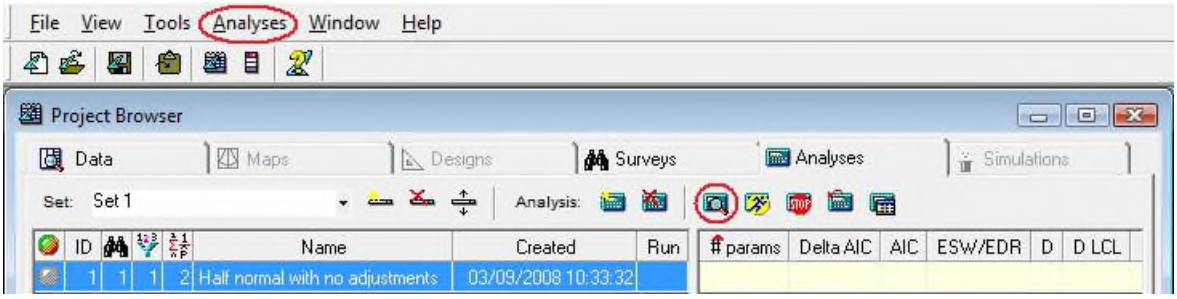
Examining the data

- Click on the **Data** tab of the **Project Browser** to show the **Data Explorer**. Look at the data structure and in particular how the distance data have been entered. (You will need to click on **Observation** in the left hand pane of the Data Explorer to see this.)

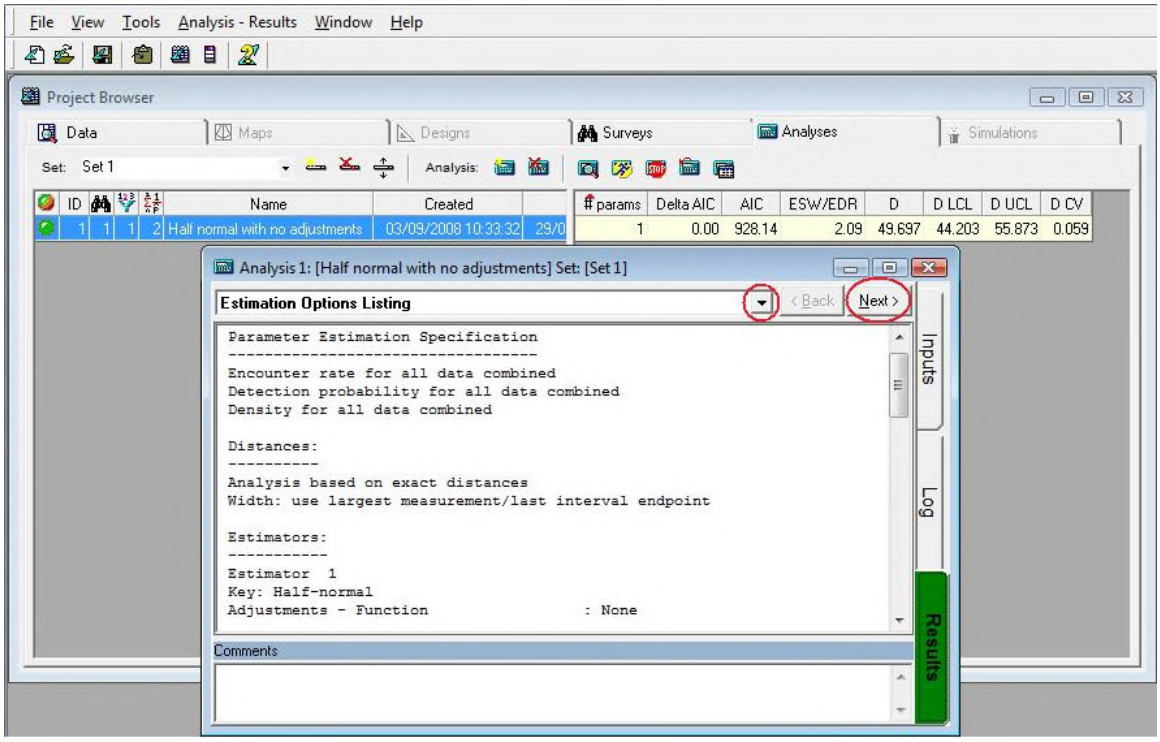


Studying the first analysis

- Now click on the **Analysis** tab of the **Project Browser** to show the **Analysis Browser**. You should see one analysis listed, called "Half-normal no adjustments." Double-click on the grey status button for this analysis to open the **Analysis Inputs** tab for this analysis (you can do the same thing by clicking the 3rd button after "Analysis:" on the Analysis Browser menu bar, or by choosing **Analyses** then **Analysis Details...** from the menu bar at the top).

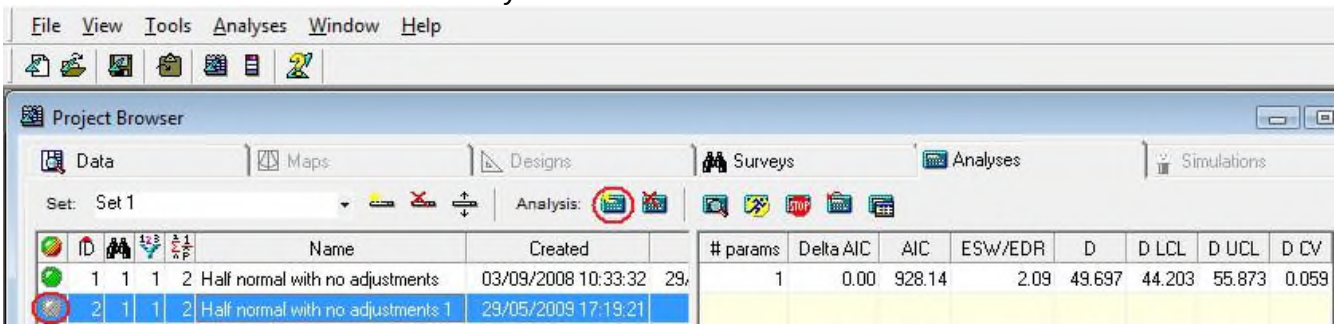


- A grey status icon indicates that this analysis has yet to be run. Click on the **Run** button in order to run the analysis. The **Results** tab should turn green.
- Click on the **Results** tab to see the results, and use the **Next >** button to move through the pages of results, looking at each page and trying to relate the analysis given here to the one you did by hand. (**Note:** These are the analysis details (Inputs/Log/Results) for one analysis – you can resize this window so that you can view details from multiple analyses when you have more than one analysis to compare).

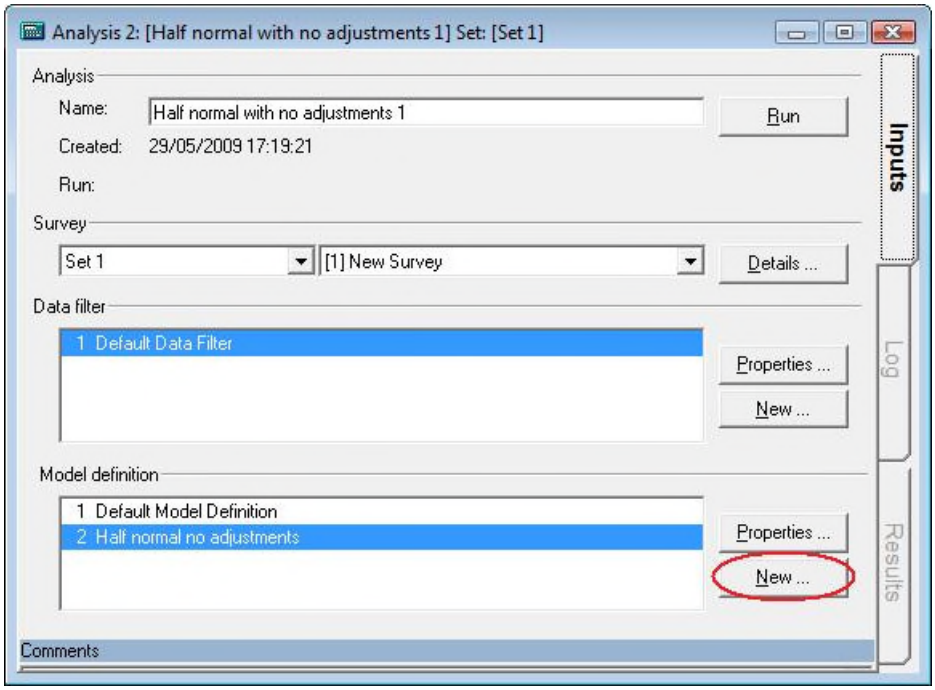


Creating a new analysis

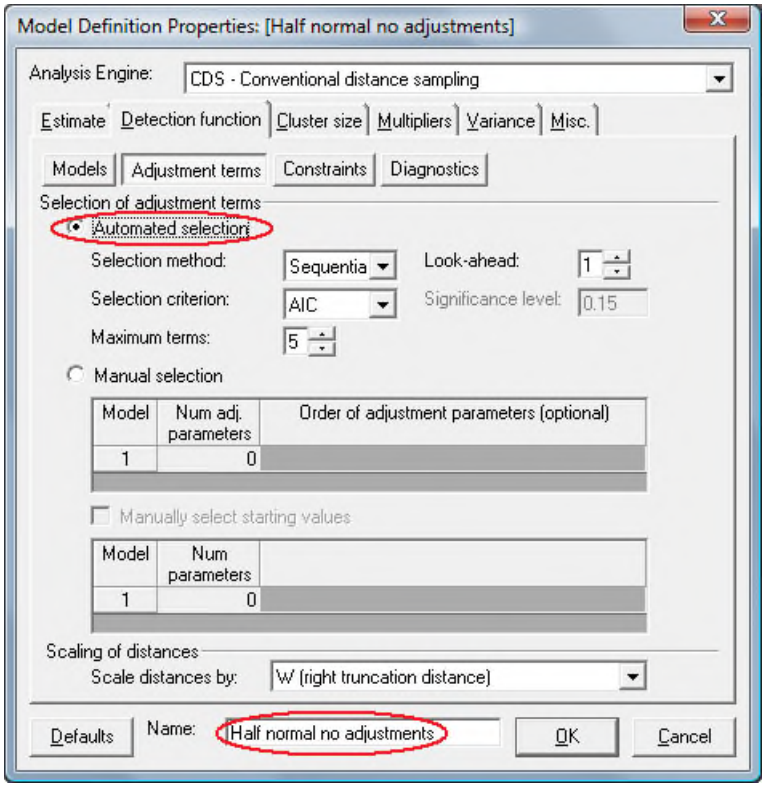
- Return to the Analysis Browser, and click on the first button after “Analysis:” on the Analysis Browser menu bar (“New Analysis”). Double-click on the status button to go to the Analysis Details window for this new analysis.



- Because the analysis is not run, you are taken to the **Inputs** tab. You will not need to edit the Survey or Data Filter for this example, but click on **New** in the **Model Definition** section.

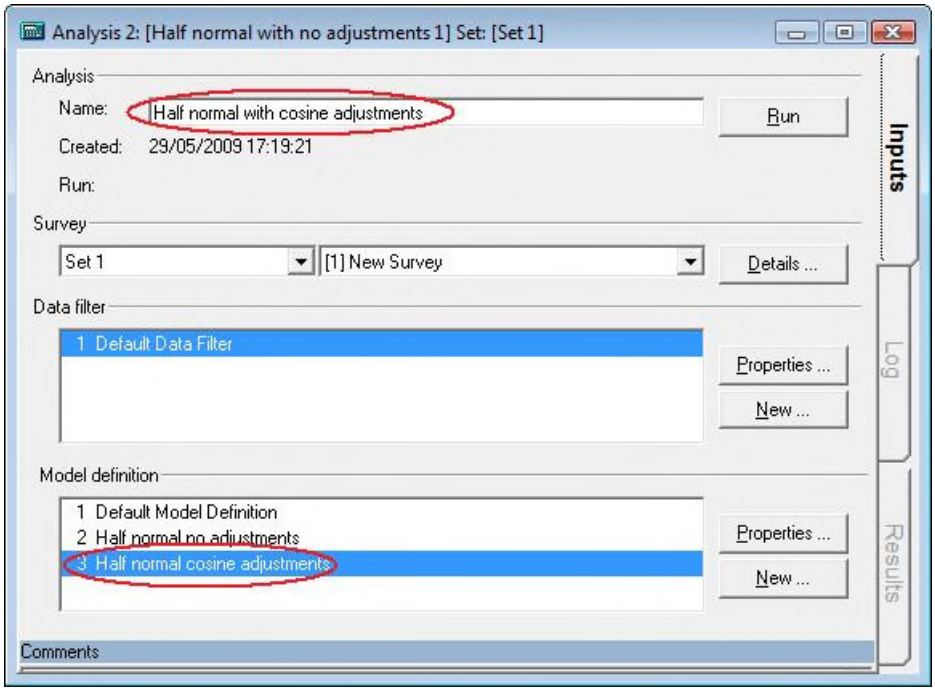


- Specify a half-normal key function with a cosine series adjustment, allowing selection of adjustment terms. When you have defined your new model, give it a suitable name (one that reflects the options you have set) and select **OK**.



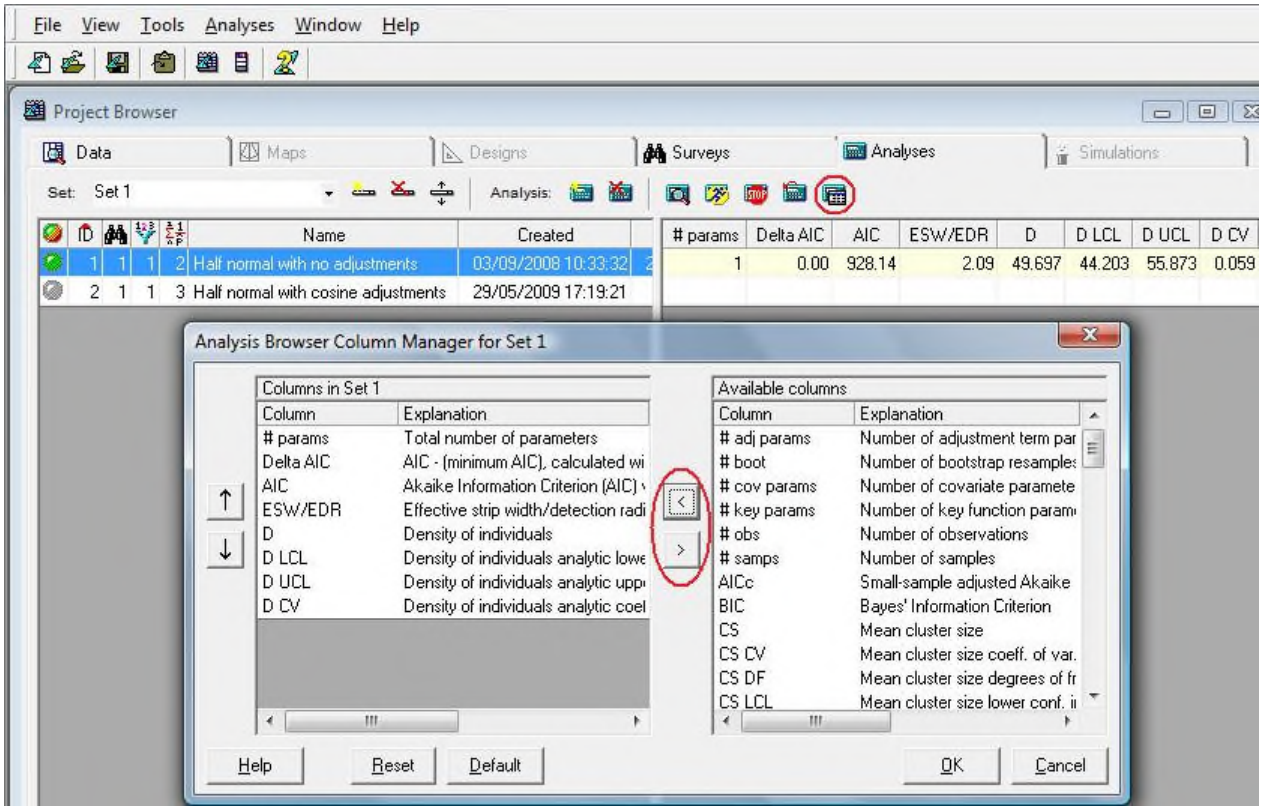
- Now give your analysis a suitable name, and click the run button. When the analysis finishes, it will automatically take you to the log tab if there were problems, or the results tab if the

analysis ran without errors or warnings. From the results tab, you can investigate the result of your analysis.

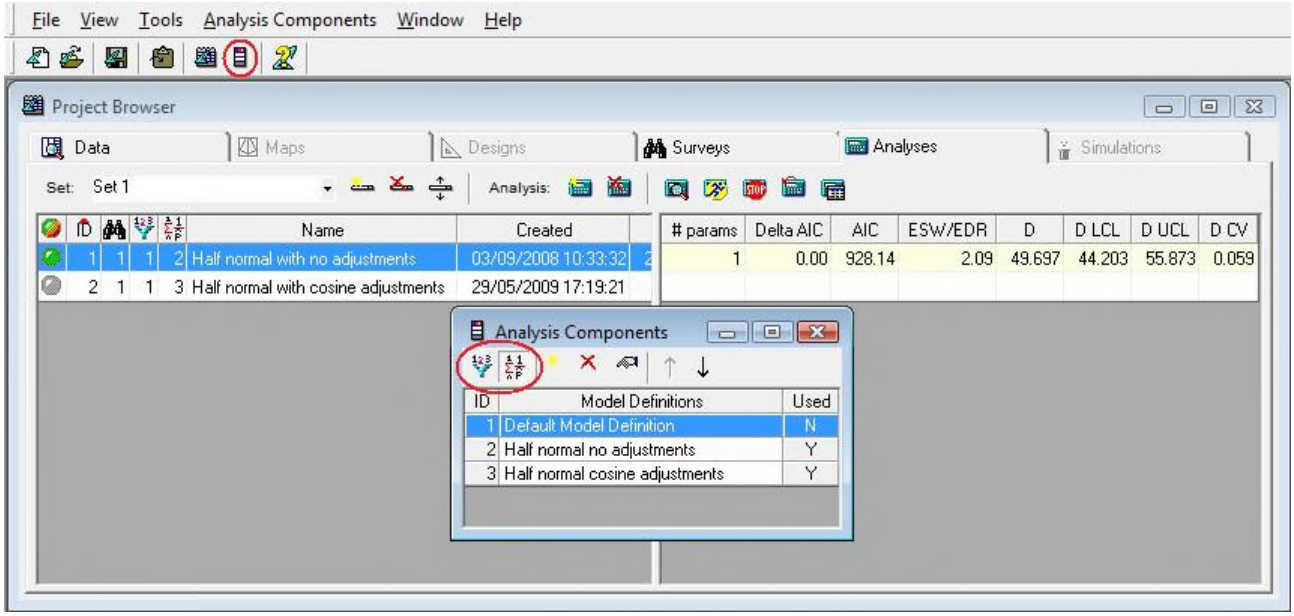


- Create one more detection function model, this time specifying the hazard rate as key function, and Hermite polynomial as the adjustment. Compare the performance of the 3 models you fitted to this dataset. **Note:** when you create a new analysis (or model definition or data filter), Distance copies the settings from whichever analysis (or model definition or data filter) was highlighted at the time (the name is also copied). The default settings are not restored automatically.

- It is easiest to compare results from different analyses using the Analysis Browser. You can change the default columns in the browser using the **Column Manager** (furthest button on the right of the Analysis Browser menu bar).



- As you create more Data Filters and Model Definitions, you may find that you want to change their order, rename or delete them. A convenient way to do this is using the **Analysis Components** window – click the 6th button from the right on the main menu bar ("View Analysis Components"). In the Analysis Components window, clicking the first button lists the Data Filters and clicking the second button lists the Model Definitions.

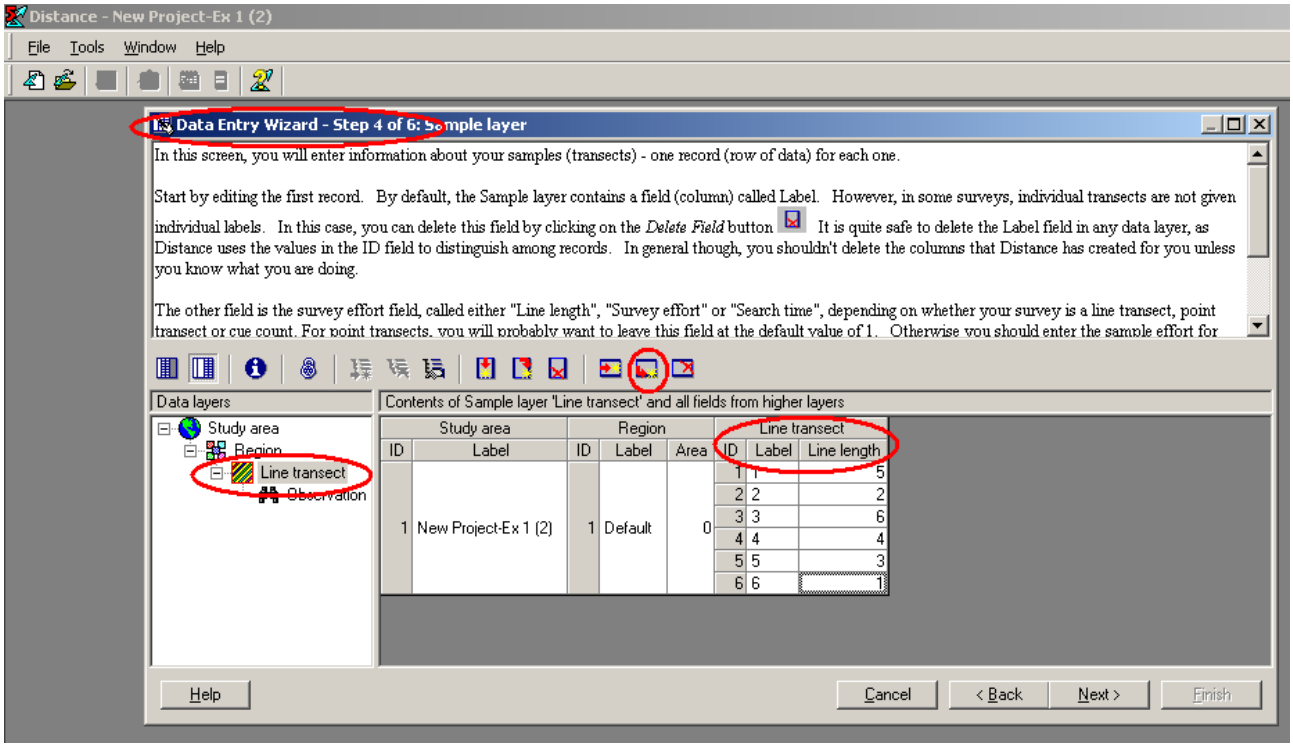


Introduction to Distance Sampling

Exercise 3: Line transect estimation using Distance

1(a). The line transect data immediately below were generated from a half-normal model.

- Open a new project (click on **File** then on **New project ...**), name it, and click on **Create**. Step through the New Project Setup Wizard (you should not need to change any of the defaults, except the units for density estimates to km^2 not the default hectares, but study each page) and click on **Finish**. This takes you to the Data Entry Wizard. Click **Next** until you get to the "line transect" page Step 4 of 6: Sample layer. Enter say the first 6 line labels (e.g. "line 1", "line 2", …) and lengths (5, 2, …). You need to click on the "append new record after current" button on the menu bar or type CTRL + Enter together before entering the information for each line.



- When you have finished, click on **Next** and enter the distances corresponding to each observation in a similar fashion (using CTRL + Enter between each observation). Once you have entered the distance data, go to the analysis browser, and carry out an analysis of these data using the half-normal detection function key.

Perpendicular distances in metres generated from a half-normal line transect model.

Line 1; length 5km
7.9 10.2 12.4 3.8 4.8 8.5 13.4 5.8 7.5 11.5
0.9 9.2 12.5 6.1
Line 2; length 2km
9.1 6.4 21.2
Line 3; length 6km
3.8 12.6 4.7 17.9 14.5 5.1 4.2 3.6
Line 4; length 4km
11.2 12.2 1.8 35.8 2.6 6.2 9.7 4.0 9.7
Line 5; length 3km

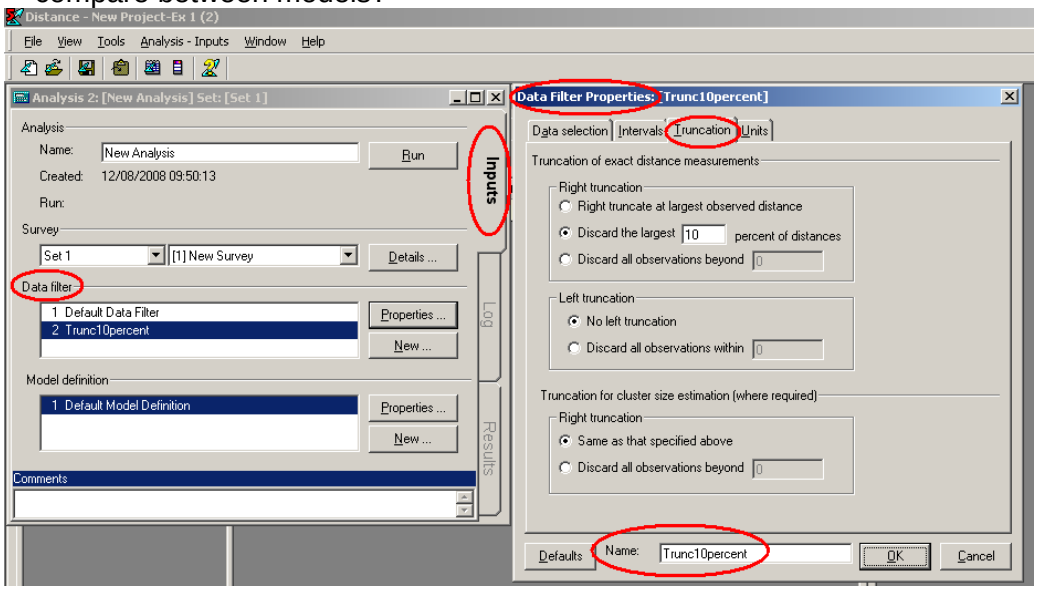
```
 6.9   5.1   3.3
Line 6; length 1km
 6.0  18.4   3.8   2.9
Line 7; length 4km
 3.3   2.9   3.7  13.2   1.0   2.3  13.4  16.2   3.8  19.3
11.1
Line 8; length 4km
 0.8   1.5   0.7  10.2  10.0   0.6   7.6   4.4
Line 9; length 5km
 1.0   1.0   1.2   4.6   9.2  15.8   1.9   3.3   3.7   5.8
 5.9   4.8  12.4   7.6  10.6  17.8   5.8
Line 10; length 7km
 0.0   0.6   2.0   6.9   7.2   7.7  10.2   1.3   1.7   8.4
13.4  19.4  12.8  13.2   6.3  10.0  12.4  19.5   1.7   3.1
 3.3  19.4  16.6
Line 11; length 3km
no detections
Line 12; length 4km
1.0   6.6  12.4   4.9  15.4
```

(b) The full data set is in project **Exercise3.zip** Choose Open project and select zip file type.

- Experiment with keys other than the half-normal (uniform, hazard-rate and negative exponential), to assess whether these data can be satisfactorily analysed using the wrong model.

- For each key, determine a suitable truncation point, and decide on whether, and which, adjustments are needed. Truncation points come under the data filter – click **New…** in the **Data Filter** section and create and name your own data filter, including truncation. In the example data filter below, the largest 10% of distances were truncated – you may want to truncate at a specific distance, depending on the data.

- Given that the true density was 79.8 animals / km^2 for these data, how do bias and precision compare between models?



Additional question

2. Below are perpendicular distance data (m) from line transect surveys of capercaillie (a large grouse) in Scotland. Total line length was 240km. The data are also in a text file **capercaillie.txt** in the Distance project directory. In the text file, column 1 is the transect number, column 2 is the transect length and column 3 is perpendicular distance. Columns are separated by tab characters. Create a new Distance project and either enter the data by hand or use the **Data Import Wizard** (Tools > Import Data Wizard) to import the data from the text file. Then decide on a suitable model for the detection function and estimate bird density.

CAPERCAILLIE, MONAUGHTY FOREST

n=112

28.0	17.0	15.0	14.0	18.0	0.0	38.0	6.0	50.0	65.0
75.0	1.0	70.0	28.0	40.0	40.0	40.0	15.0	40.0	30.0
5.0	55.0	60.0	40.0	24.0	30.0	0.0	50.0	55.0	10.0
40.0	10.0	30.0	34.0	24.0	30.0	15.0	20.0	14.0	48.0
0.0	30.0	2.0	52.0	11.0	48.0	28.0	38.0	25.0	35.0
45.0	0.0	16.0	12.0	2.0	14.0	12.0	24.0	70.0	50.0
49.0	40.0	80.0	18.0	27.0	30.0	30.0	60.0	58.0	14.0
0.0	56.0	40.0	19.0	21.0	0.0	38.0	20.0	28.0	30.0
20.0	16.0	0.0	69.0	40.0	46.0	50.0	40.0	70.0	67.0
28.0	12.0	12.0	22.0	40.0	48.0	48.0	15.0	12.0	0.0
15.0	20.0	17.0	30.0	30.0	32.0	48.0	20.0	10.0	20.0
42.0	30.0								

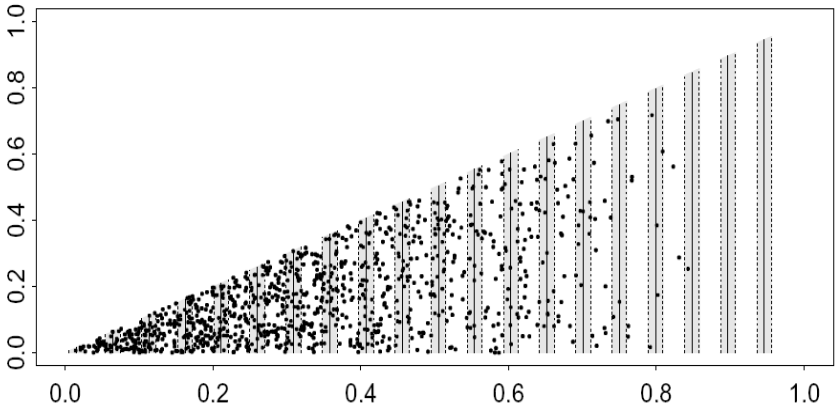
Introduction to Distance Sampling

Exercise 4: Variance estimation for systematic designs, the bootstrap method

In the lecture describing measures of precision we explained that systematic survey designs usually have the best variance properties, but obtaining good estimates of the variance is a difficult problem for statisticians. In this exercise we give an example of a situation where the systematic design gives a density estimate with much better precision than a random design. This means that the usual variance estimators used in Distance, which are based on a random design, give variance estimates that are far too high. The true variance is low, but the estimated variance is high. We will see how to implement a post-stratification scheme that enables us to get a better estimate of the variance. We also look at another case to see that the unstratified variance estimates provided by Distance are usually fine for a systematic design: things only go wrong when there are very strong trends in animal density, especially when the strong trends are associated with changes in line length (e.g. the highest densities always occur on the shortest lines, or vice versa).

We begin with the population and survey shown below. All the populations in this exercise are simulated on a computer: they are not real data. Note the characteristics: extreme trends with very high density on short lines and very low density on long lines. Additionally, the systematic design (search strips are shaded) covers a fairly large portion of the survey area. These are the danger signals that the usual Distance variance estimators might not work well, and a post-stratification scheme should be considered.

The survey region is a triangle, with dimensions 1km by 1km. The systematically placed search



strips are shaded above.

Basic variance estimation, with bootstrapping

1. Open the Distance project *Systematic_variance_2.zip*.
2. On the Analysis tab, click New Analysis. Rename it *Without post-stratification*.
3. Under Model definition, click Properties. Rename the new model: *No_adjustments_plus_bootstrap*.
4. Click the tab for Detection function, and click Adjustment terms. Select Manual selection so that no adjustment terms are fitted. Select the Constraints button, and click No constraints. These options will reduce the work that has to be done during bootstrapping.
5. Click the tab for Variance, and check the box for Bootstrap variance estimate: Select non-parametric bootstrap. The box Resample samples should be checked (this means resample transect lines). Leave the other settings at default, noting that there will be 999 bootstrap resamples conducted.
6. Click OK and then run the model. You can see the progress of the bootstrap in the bar at the top. Wait a few moments until the bootstrapping is completed.

7. Your analytic output should look like this:

```
                    Estimate      %CV     df     95% Confidence Interval
                    ----------------------------------------------------
Half-normal/Cosine
               D       2044.6       27.70    20.74  1161.0       3600.6
               N       1022.0       27.70    20.74  581.00       1800.0
```

8. Because we have simulated these data, we know what the true values are. The true number of animals in the survey region is N=1000, and the true density is D=2000 km^{-2} (1000 animals in an area of size A=0.5 km^2). The point estimates are good, but what do you think about the precision in the output above?
9. Find the bootstrapped confidence intervals for D and N, and check whether they are similar to the confidence intervals above.
10. What percentage of the total density variance is attributed to encounter rate estimation and what percentage to the detection function estimation?

Variance estimation for systematic designs using post-stratification

Recall we have a particular situation in which we have systematically placed transects, unequal in length. Furthermore there exists an east-west gradient in animal density juxtaposed such that the shortest lines are those that pass through the portion of the study area with the highest density. We examine a means by which we can use post-stratification to produce a better estimate of the variance in estimated density.

Post-stratification to improve variance estimation

The estimation of encounter rate variance in Exercise 4 used estimators that assumed the transect lines were randomly placed throughout the triangular region. In our case, the transects were not random, but systematic. In some circumstances, this can reduce the encounter rate variance a great deal. The data we are working with is an example of this. There are very high densities on the very shortest lines. In samples of lines collected using a completely random design, the sample by chance might not contain any of these very short lines, or it might contain several. The variance is therefore very high, because the density estimates will be greatly affected by how many lines fall into the short-line / high-density region: we will get very low density estimates if there are no short lines, but very high density estimates if there are several short lines. By contrast, in a systematic sample, we cover the region methodically and we will always get nearly the same number of lines falling in the high density region. The systematic density variance is therefore much lower than the random placement density variance.

Although there is no way of getting a variance estimate that is exactly unbiased for a systematic sample[1], we can greatly improve on the random-based estimate by using a post-stratification scheme. This works by grouping together pairs of adjacent lines from the systematic sample. Each pair of adjacent lines is grouped into a stratum. The strata will improve variance estimation, because the systematic sample behaves more like a stratified sample than a random sample.

Follow the steps below.

1. Open the Distance project we used in the previous section (**Systematic_variance_2.dst**; is has the ".dst" extension because you uncompressed it minutes ago).
2. Click the Analyses tab, and click the "New analysis" button to create a new analysis. Double click the grey ball and the Analysis Details Window should come up. Name the new

[1] because it is effectively a sample of size 1 – only the first line position was randomly chosen, and the rest followed on deterministically from there.

analysis something like *With post stratification*.

3. Under Model Definition, click New. Change the name at the bottom of the dialogue box to *Poststratified_no_adjustments_no_bootstrap*. (We don't want to conduct a bootstrap for our poststratified data, because it would involve some extra confusion and is not necessary.) In the Variance tab, click Advanced..., and select the option “Post-stratify, grouping adjacent pairs of samplers”. Un-tick the option “Select non-parametric bootstrap”.
4. Click OK and then Run to run the analysis. How does the variance and confidence limits compare with those you obtained in the previous section? What are the implications? Note what percentage of the overall variance now comes from encounter rate and from estimating the detection function, and compare this with the earlier percentages.
5. Now try the overlapping post-stratification option. A simulation study in Fewster et al. (2009) concluded that its performance was very similar to, but marginally better than the regular post-stratification. When the sample size of lines is small, it gives more post-strata and so is to be preferred for that reason. Create a new analysis, called say *With overlapping post stratification*, and then a new Model Definition for that analysis, in which you choose the Advanced variance option “Post-stratify, with overlapping strata made up of adjacent samplers”. How does the variance compare with those you previous obtained? How do the degrees of freedom in the Estimation Summary – Encounter Rate page of output compare with that from the previous question?
6. (Optional) If you wish, you can try manual post-stratification. This is good practice if you need to do post-stratification for point transect studies. In this case you will have to add a new field to the sample layer, and then set up a new model definition in which you tell Distance to use post-stratification. Here goes:

a) Click the Data tab. Click the padlock button on the toolbar to unlock the data sheet for modification.

b) On the left-hand outline, click Line transect. The data sheet expands to 20 rows, each row corresponding to one line transect. This is the best format for the data sheet to be in when entering a new stratum number for each transect.

c) Click on the cell corresponding to Line transect Label 1. Several buttons on the tool-bar should become live. Click on the button corresponding to *Append field after current*. (The button has an arrow pointing sideways then downwards.)

d) You are prompted for Field name: enter VarGroup to indicate that you are grouping lines together for the purpose of variance estimation. Click Field type: Integer, and click OK.

e) You can now enter the line groupings for post-stratified variance estimation. Enter label 1 for lines 1 and 2; label 2 for lines 3 and 4; label 3 for lines 5 and 6; and so on, to finish with label 10 for lines 19 and 20. You have now defined 10 strata, each containing two adjacent transect lines from the systematic sample of lines.

f) After entering the column of VarGroup labels, click the padlock button again to lock the data sheet.

g) Now we will analyse the post-stratified data. Click the Analyses tab. Create a new analysis with a suitable name - .e.g, *Manual post stratification*

h) Create a new Model definition, with a suitable name. In the Estimate tab, click the button for Poststratify. Select Layer type: sample, and Field name: VarGroup. This means that we want to poststratify at the sample (transect) level, using our newly defined groupings VarGroup to delimit the strata.

i) For the levels of resolution, select the following:

- Density: Global *and* Stratum
- Encounter Rate: Stratum only
- Detection function: Global only

- Cluster size (not required): Global only

These settings ensure that it is only encounter rate variance that is affected by the post-stratification scheme; the detection function is still pooled over all observations as before.

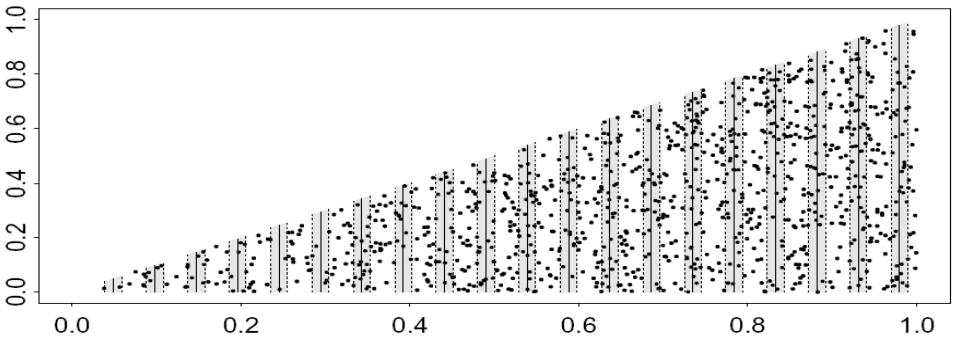
j) In the next field, enter Global density estimate is *Mean* of stratum estimates, and in the next field select Weighted by *Total effort in stratum*. Do **not** tick the box saying Strata are Replicates.

k) Click OK and run the new model. The point estimates should be the same as the previous non-overlapping post stratification run.

Note: The precision of D and N are greatly improved in the post-stratified analyses. Note that we are not getting something for nothing: the second analysis is giving us an answer much closer to the true answer, while the first analysis was simply giving us the wrong answer. We have not *changed* the true variance by our post-stratification scheme: we are just getting a better *estimate* of the true variance. Because the data above were generated by simulation, we can use repeated simulated surveys to check that the second answer is indeed close to the true density variance over the repeats.

Systematic designs where post-stratification is not needed

The following simulated population does not exhibit strong trends across the survey region. Otherwise, the strip dimensions and systematic design are the same as for the previous example.



Open the project **Systematic_variance_1.zip**. Add the new data column VarGroup before conducting any analyses this time. With the augmented data, repeat the analyses you performed on the Systematic_variance_2.zip project. Find the relevant outputs. Has the post-stratification scheme been necessary in this case?

Introduction to Distance Sampling

Exercise 5: Point transect exercises

1. Simulated point transect data from 30 points are given in project **PTExercise1.zip**. These data were generated from a half-normal detection function, and true density was 79.6 animals / ha. Experiment with keys other than the half-normal (uniform, hazard-rate and negative exponential), to assess whether these data can be satisfactorily analysed using the wrong model. For each key, determine a suitable truncation point, and decide on whether, and which, adjustments are needed. (Truncation points come under the data filter.) How do bias and precision compare between models?

2. The projects **Wren1.zip**, **Wren2.zip**, **Wren3.zip** and **Wren4.zip** contain winter wren data, collected at Montrave, Scotland in 2004. Each project corresponds to a different method of data collection. Thirty-two points were defined through 33.2 ha of parkland (Fig. 1), and detection distances were measured in metres with the aid of a laser rangefinder. Three types of point transect data were collected: 1. standard five-minute counts; 2. the 'snapshot' method; and 3. a cue count method. In addition, line transect data were collected (method 4), and territory mapping was conducted, which gave an estimate of 43 wren territories (1.30 territories ha^{-1}).

    a) Select a single model for exploratory data analysis. Experiment with different truncation distances *w*, and select a suitable value for each method. Do you see potential problems with any of the data sets?
    b) Try other models and other model options. Use plots, AIC values and goodness-of-fit test statistics to determine an adequate model.
    c) Record your estimates of density for each method. Record also the corresponding confidence intervals. Compare your answers with those of others in the workshop.

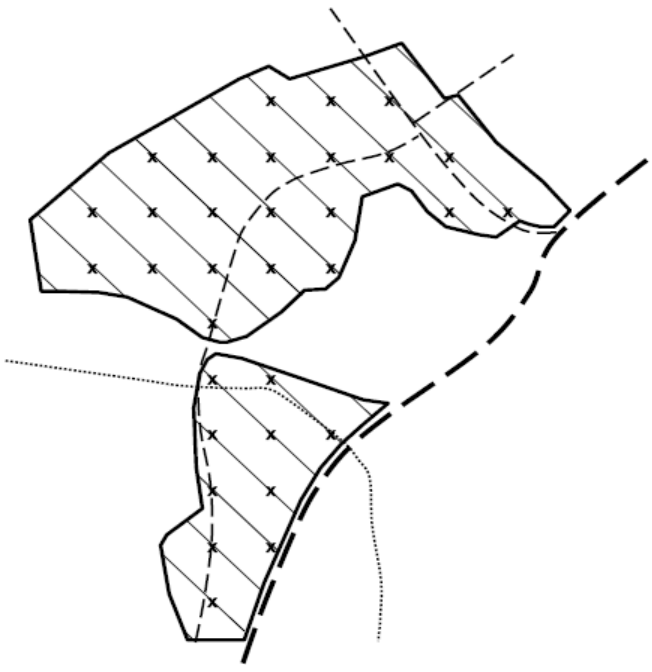


Figure 1: The study site at Montrave in Fife, Scotland. The dotted line is a small stream, the thin dashed lines are tracks, and the thick dashed line a main road. The 32 points are shown by crosses, and are laid out on a systematic grid with 100m separation. The diagonal lines are the transects used for method 4.

3. The Sample Projects directory contains two point transect projects, **Savannah Sparrow 1980.zip** and **Savannah Sparrow 1981.zip**. These were part of a large data set collected in Arapaho National Wildlife Refuge, Colorado. For both data sets, consider an appropriate truncation distance, decide on a suitable model for the detection function, and estimate density, both for each stratum individually and for the whole study area. You should include in your analysis an assessment of whether the detection function can be estimated from data pooled across strata, or whether separate estimates are needed per stratum. (This will be covered in the lecture discussing stratification if you don't already know how to do it.)

Introduction to Distance Sampling

Exercise 6: Automated Survey Design Exercises

1. Point transect survey of North-eastern Mexico

Reviewing the data

Extract and open the project **MexicoUnPrj** from the archive **MexicoUnPrj.zip**. This project contains data from 4 states in North-eastern Mexico. Let's begin by reviewing the data. On the top menu-bar select **File**, **Project Properties**. The **General** tab gives you information about the location of the project file and its associated data folder (MexicoUnPrj.dat). The **Geographic** tab gives you information about the default geo-coordinate system of the geographic data, and the default map projection. The geo-coordinate system is used to locate the geographic data (which is stored in decimal degrees of latitude and longitude) on the earth's surface. The projection is used to convert these data from the curved surface of the earth into a flat plane that can be used for displaying maps and designing surveys. The resulting projection has linear units, such as metres or kilometres. If you are planning a survey that will take place over a small geographic area, and you are inputting your data by hand, then you don't need to worry about geo-coordinate systems or projections and can set both these options to [None]. In this example, however, the survey area is quite large and the projection chosen will make some difference to the results. Click **Cancel** to close the Project Properties window without saving any changes you have made.

Click on the **Data** tab of the **Project Browser** to view the **Data Explorer.**. In the left-hand pane, under Data Layers, you can see that there are four layers in the project: “Mex”, “MexStrat”, “Grid1” and “Grid2”. You can tell the layer types by looking at the icons beside the names: Mex is a Global layer, MexStrat is a Stratum layer and Grid1 and Grid2 are Coverage layers. When you open a new project, the Global layer is selected by default, so the layer Mex is now selected. Click on the **Data Layer Properties** button on this tab (7th button from left) to find out more about this layer. The **Layer Properties** window opens, and under the **Geographic data** tab, you can see that the geographic data is stored in a shapefile called Mex.shp in the data folder, and that the shapes in this layer are Polygons (i.e. solid shapes). Click **Cancel** to return to the Data Explorer.

In the right-hand pane of the Data Explorer, you can see its’ fields: ID, Label, Area and Shape. There is one record, with ID = 1. The Shape field holds the geographic information for that record. Because this layer holds polygons, the shape record has the word "Polygon" in it. Double click on this word to open the **Shape Properties** window. This is where you edit the geographic information inside Distance (an alternative is to edit the shapefile Mex.shp from outside of distance, using a GIS package such as ArcGIS). Click **Cancel** to return to the Data Explorer.

The coverage layers Grid1 and Grid2 contain a grid of points that will be used for determining probability of coverage for our survey designs. If you click on “Grid 1” in the left-hand pane of the Data Explorer, its records open in the right-hand pane, and you can see that it has 177 records. A better way to look at the grid points is to view them in a map. Click on the **Maps** tab in the **Project Browser**. Click on the **New Map** button (3rd button along) to create a new map. Double-click on the words "New Map" to edit the name of the map, and call it "Grid 1". To view the map, click the **View Map** button (5th button along), or double click on the map's ID. A **Map Window** opens.

The map starts life blank. You add layers to the map by clicking the **Add Layer to Map** button (7th button along). Click this button and select "Mex" from the list of layers. Then click the button **Add Layer** button again and select "Grid 1" from the list. Now you can see the grid points.

You could also add the points from Grid 2 to the same map. If you do this, you will see that the grid points for Grid 2 are much closer together than those for Grid 1. (Grid 2 was generated with a spacing of 20 km, while for Grid 1 the spacing was 50km.) The points for Grid 2 obscure those from Grid 1 – you can change the order of the map layers by clicking on a the legend "Grid 2" in the left-hand pane of the map and holding the mouse button down while you drag it down to below "Grid 1". Click the [X] button in the top right corner of the Map Window to close the map. (Say "Yes" if it asks you to save changes.)

In the **Data Explorer**, click on the MexStrat layer to see those data. You can see that there are 4 strata. If you want to see where they are, you could create a new map in the **Maps** tab and add the MexStrat layer to the map. If there are layers on the new map that you don't want, you can remove them with the **Delete Selected Layer** button in the **Map** window.

When you've finished exploring the data, move on to create a new design.

Creating a new design

Click on the **Designs** tab of the **Project Browser**. To create a new design, click the **New Design** button (1st one after the word "Design:"). A new record appears in the left-hand pane, called "New Design". Double click on the name, and edit it to call the new design "150 random points". If you need more room, click and drag to the right the vertical splitter that divides the Designs window into two. Click the **Show Details** button (3rd one after the word "Design") to open the **Design Details** window. Look under "Type of design" to see the sampler and design class; the default sampler is "Point" and the default design class is "Simple Random Sampling". Click the **Properties** button to set the properties for this design. The **Design Properties** window opens. The options you see on the design properties tabs depend on the type of design. In this example, choose the following options:

- Under Stratum layer, choose the stratum layer "MexStrat".
- Under Design coordinate system, make sure the box "Same coordinate system as stratum" is unchecked. The projection should say "Plate Carree" and the units “Metre”.
- In the **Effort Allocation** tab, under Edge Sampling select the “Plus” option. Uncheck the box "Same effort for all strata". A list of the four strata in the MexStrat layer appears. Under "Allocation by stratum", click the "Percentage from" radio button, and enter "150" as the number of points. In this example, we will put most of our effort into the two Baja strata (perhaps because this is where we think most of the animals of interest live). Under "Effort %" enter 10 for Sinaloa, 10 for Sonora, 40 for Baja Sur (south) and 40 for Baja Norte (north).
- In the **Sampler** tab, select Kilometre for the point sampler radius units. Let's imagine we're surveying for a very vocal species and that our truncation distance will be 5km, so we enter 5 under radius (for this example we'll assume same sampler properties for all strata).
- Lastly, in the **Coverage Probability** tab, click on "Estimate by simulation" and enter 100 as the number of simulations. This is far too few for an accurate simulation, but will do for the purposes of demonstration. Under grid layer, choose "Grid 2", which is the one with the grid points closer together.

Now click OK to close the **Design Properties** window. The properties window closes and we are back with the design details.

Automated generation of new surveys

Click the **Run** button on the **Design Details** window. A window pops up offering you two choices: (1) Calculate coverage probability statistics, and (2) Generate a new Survey. Choose the second option, and give the new Survey a useful name like "150 points survey" and the new layer a name like "150 points". Then click OK. A **Survey Details** window opens, and the status bar at the top of the Distance window says "Running Survey". At this point you have to be patient while the survey runs. Distance is creating a set of randomly located survey points, based on the design. When it is finished, the **Survey Details Results** tab opens, and you can review some statistics about the new survey. Click the "Next >" button to see a map of the points – you should be able to see that there are more in the

Western strata (Baja) than the eastern. Click “Next>” again to see a list of the points, with latitude and longitude for each. (You could, for example, use this to make a survey plan to take into the field. To copy this text to another file, press the “Copy current window” button, 4th from the left on the top toolbar. Then open, say, a Word document and click Paste to copy it there. You can also copy the map of points by displaying the map and pressing the copy button, or choosing the menu item Survey – Results | Copy Map to Clipboard)

Click on [X] to close the **Survey Details** window, and click on the **Surveys** tab of the project browser. You can see that your new survey has been added there. If you select it and click the "Show Details" button (3rd from left after the word "Survey") you get back to the **Survey Details** window **Results** tab. Click on the **Inputs** tab and then **Properties ...** button. Under **Data Layers**, you can see that the new Sample data layer "150 points" has been entered as the lowest sample layer. Close the **Survey Properties** and **Survey Details** windows, and click on the **Data** tab of the **Project Browser**. You can see that the new sample data layer "150 points" has been added below the "MexStrat" data layer.

Design statistics

Go back to the **Design Details** window for your design, and click **Run** again. This time, choose the top option (Calculate probability of coverage statistics) and click OK. You have to wait while Distance generates multiple simulated surveys and uses these to work out the probability that each grid point will be covered by the survey. When it has finished, you can see the results in the **Results** tab, and a map of coverage probability by pressing the "Next >" button. In theory, this design should produce an even probability of coverage within stratum. However, you can see that there is considerable variation. Why is this? What would happen if you repeated the run with more simulation runs (say 500, or 1000)? (Before you spend a lot of time running simulations with this project, read the next section.)

Working with projected raw data

There is a technical problem with the way the geographic data are stored in MexicoUnPrj. Each time you view a map or run a survey design, the data have to be projected from the latitude and longitude format in which they are stored using the projection you have specified (Plate Caree in this case). This takes some computer time, so if you’re doing lots of survey design work there’s a trick to make things more efficient. The trick involves projecting the raw data files.

We used ArcGIS to project the shapefiles in MexicoUnPrj using the Plate Caree projection, and stored this new data in the project MexicoPrj. So rather than being stored in latitude and longitude, the data in MexicoPrj is stored in meters. Run a second instance of Distance, and then extract and open the project **MexicoPrj**. Look under the **Project Properties**, and you will see that the GeoCoordinate system and Projection are both set to [None], and that the units are meters. So, we’ve projected the raw data, and so long as we project all the data layers the same way we don’t need to tell Distance anything about the coordinate systems used.

As a check that the data really are projected, go the Data Explorer and double-click on the global layer’s Polygon record. The first value is something like x= -12594701 y=3230255 – this gives the number of meters of that point on the polygon from some reference point on the earth. If you do the same thing in the MexicoUnPrj project, you’ll see that the first value is something like x=-113 y=29, which is the latitude and longitude of that point.

If you’re going to do lots of experimenting with the Mexico data this evening, or at home, it’s better to use the MexicoPrj project, as you’ll find the probability of coverage simulations run quite a bit faster. Meanwhile, move on to the next exercise.

2. Entering geographic data into Distance, and generating Coverage grids

The purpose of this exercise is to show you how to enter geographic data by hand into Distance, and how to generate Coverage grid layers.

Create a new project and enter data

On the top menu-bar select **File**, **New Project** (or click the toolbar button). In the **Create New Project** dialog box give it the File Name "Trapezium" and then click on **Create**. The new project setup wizard starts up. Under "I want to", select the option to "design a new survey", and click **Next**. Then click **Finish**.

The **Project Browser** will open up, showing the **Data** tab. Click on the menu **File | Project properties**, and look under the **Geographic** tab to confirm that there is no geographic coordinate system for this project (i.e. non-earth referenced), and that the default units are metres. Click **OK** to close the **Project Properties** window.

In the **Data** tab of the **Project Browser**, you can see that Distance has created a global data layer called Study Area, with default fields ID, Label and Shape. Double click on the word "Polygon" to open the **Shape Properties** window to edit the new survey region. Click on the **Insert Point** button 4 times and fill in the following (X,Y) coordinates: (0,0), (0,100), (120,20) and (120,0). Click **OK** to return to the Data Explorer.

Generate a coverage grid layer

To generate a coverage grid layer click on the **Create New Data Layer** button (5th from left) in the **Data** tab of the **Project Browser**. Enter "TrapGrid" as your Layer Name and set the Parent Layer to "Study Area" and the Layer Type to "Coverage". You should now be able to click on the **Properties…** button. In the **Grid Properties** that pops up set the "Distance between grid points" to 2.5 and the "Units of distance" to "Metre". Once you press **OK** you should proceed to add the grid points to the layer. This may take a few moments.

Create a new map on the **Map** tab of the **Project Browser** and add your new global and coverage layers to take a look at them.

Creating a new design

Click on the **Designs** tab of the **Project Browser** and then the **New Design** button. Rename your design "equal angle zigzag" and then click the **Show Details** button to open the **Design Details** window. Select the "Line" sampler and set the design class to "Equal Angle Zigzag". Click the **Properties** button to set the following properties for this design:

- As the Trapezium survey region is non-earth referenced you don't need to make any changes on the **General Properties** tab.
- In the **Effort Allocation** tab, under "Effort determined by" select the Sampler angle option. In the "Allocation by stratum" section set the Line length units to be Metres. Make sure the "Update effort in real time" check box is ticked. As there is only one survey stratum it does not matter whether the "Same effort for all strata" check box is ticked or not. The "Absolute values" radio button is the only one available when effort is determined by sampler angle. Enter 75 in the "Angle" (measured in degrees) column of the table. The "Length" column should now read 463.644. The accuracy of this approximation of zigzag length depends on the shape of the survey region, but should be relatively accurate for convex survey regions.
- In the **Sampler** tab, set the width to 1 meter.
- Lastly, in the **Coverage Probability** tab, click on "Estimate by simulation" and enter 100 as the number of simulations. Under grid layer, choose previously created "TrapGrid".

Click OK to close the **Design Properties** window and return to the design details.

Design statistics

Run your design to work out the coverage probabilities - this design will take a while to run! In the second page of the **Design Details Results** tab that opens when its finished, take a look at the coverage probabilities map. Note how uneven these probabilities are and how they increase as the trapezium height decreases for the equal angle zigzag design.

Additional investigations

If you are particularly interested in zigzag surveying, you might want to come back to this exercise after completing exercise 3, and compare the coverage probabilities of the three different types of zigzag designs. You can do this when you get home. For now, skip ahead to exercise 3.

For work on your own:
Create two new designs - one for the equal spaced zigzag and one for the adjusted angle zigzag. To facilitate comparisons, you want to set properties for both that are somewhat equivalent to those for the equal angle design. You can see the mean trackline length for the equal angle design in its Results tab (about 460 metres). You can then set the effort allocation for the two new designs to be the same as this value. Make sure that the Coverage probability tab shows "Estimate by simulation" and that you have an appropriate Grid Field Name.

Try creating a few surveys for each design, so you can see how they differ. Then run the coverage probability simulations. How do the coverage probabilities for the 3 designs differ? You may need more simulations to see a strong difference between the equal spacing and adjusted angle design.

3. Systematic parallel line aerial survey of marine mammals in St Andrews bay

Reviewing the data

Open the project archived in **StAndrews.zip**. This project contains the survey region for an aerial survey of porpoise, common dolphins and seals in and around St Andrews bay. (For locals: the nearer St Andrews bay region has been extended in an easterly direction out past bell rock, as there are some pockets of deeper water out there that are of interest with regard to the distribution of cetaceans. The survey region has a chunk missing due to a no-fly zone around Buddo Ness, just below Carnoustie). To take a look at the survey region create a new map in the **Maps** tab and add the layer **StAndrews** to the map.

The small survey plane permits a total flight time of approximately 250 km (excluding the flight time to and from the landing strip in Fife Ness, just down the coast). A systematic line sampling design is going to be used. The survey plane permits easy movement between survey lines, but it would still be efficient to spend as much of the 250 km flight time on effort surveying rather than on movement between the sampler lines. The aim of this exercise is to decide on a systematic line spacing that gives about 200 km on-effort trackline with the total trackline length constrained to 250 km. To do this create a number of systematic line sampling designs with different line spacings, generate the design statistics for these designs and then the statistics for the total trackline length to the on-effort trackline length for different designs.

Before proceeding to the design stage you need to generate a coverage grid layer, as this will be needed to generate design statistics.

Generate a coverage grid layer

To generate a coverage grid layer click on the **Data** tab of the **Project Browser** and then the **Create New Data Layer** button (5th from left). Enter “Grid5” as your Layer Name and set the Parent Layer to “StAndrews” and the Layer Type to “Coverage”. You should now be able to click on the **Properties…** button. In the **Grid Properties** that pops up set the “Distance between grid points” to 5 and the “Units of distance” to “Kilometre”. (This is too far apart for estimating probabiltiy of coverage, but we know coverage is even for this design, so choosing a wide spacing makes the simulations run faster.) Once you press **OK** you should proceed to add the grid points to the layer. This may take a few moments.

Now create and generate a couple of designs with a spacing of your choice (some suggested spacings include 4.5, 5, 5.5 & 6 km)

Creating a new design

Click on the **Designs** tab of the **Project Browser** and then the **New Design** button. Rename your "New Design" something like "systematic line test" and then click the **Show Details** button to open the **Design Details** window. Select the “Line” sampler and set the design class to "Systematic Random Sampling". Click the **Properties** button to set the following properties for this design:

- On the **General Properties** tab under Stratum layer, the StAndrews stratum layer should be selected. Under Design coordinate system, the design coordinate system should be “Non-earth referenced”. (The data have already been projected from the OSGB 1936 geo-coordinate system using the transverse mercator projection – the same trick we used for the MexicoPrj project.)
- In the **Effort Allocation** tab, under Edge Sampling select the “Minus” option. In the “Allocation by stratum” section set the Line length units to be Kilometres. Make sure the “Update effort in real time” check box is ticked. As there is only one survey stratum it does not matter whether the "Same effort for all strata" check box is ticked or not. Click the "Systematic line spacing" radio button and enter the line spacing in the “Spacing” column of the table. When you enter a 5 km spacing for instance the “Length” column should then read 226.203 and the “Samplers” column 8. The accuracy of this approximation of on-effort line length and total number of line samplers depends on the shape of the survey region, but should at least give you some indication of what to expect.
- In the **Sampler** tab, select Kilometre for the line sampler width units. Set the truncation width to 2 km.
- Lastly, in the **Coverage Probability** tab, click on "Estimate by simulation" and enter 100 as the number of simulations. This is too few to give accurate coverage probabilities, but sufficient for the on-effort and total trackline length statistics. Under grid layer, choose previously created "Grid 5". Make sure the Grid field name is the same as your design name.

Click OK to close the **Design Properties** window and return to the design details.

Design statistics

For each design run Distance generates multiple simulated surveys and uses these to work out the statistics for on-effort and total trackline length. Run your designs and in the **Design Details Results** tab that opens review the statistics to decide on suitable systematic line spacing.

Automated generation of new surveys

To see an example survey, go back to the **Design Details** window for your selected design click **Run** again this time choosing the “Generate a new Survey” option. The second page of the survey results displays a map of the survey region with the systematic lines superimposed. You can add this map to the **Map browser** and manipulate it there by clicking on the 6th button on the Survey map results page.

Introduction to Distance Sampling

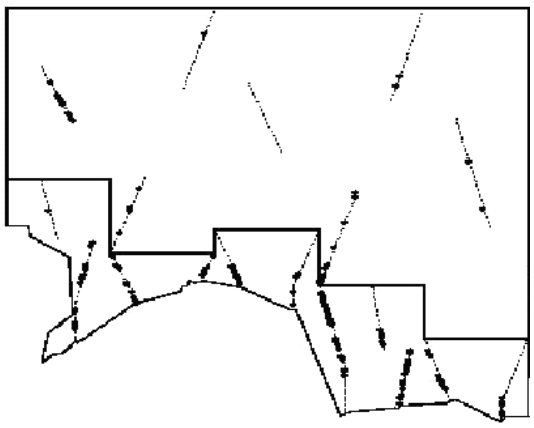
Exercise 7a: Analysis of Stratified Data

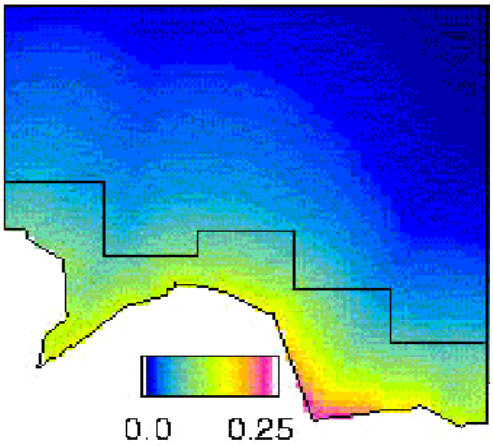
The Data

The Distance project **Stratify exercise** contains data from a stratified survey of Antarctic minke whales. The data are "exact" insofar as they are calculated directly from the estimates of radial distance and angle recorded by the observers. While angle boards and reticule binoculars were used for estimation of angles and distances when possible, the transitory nature of cues (usually blows) and the pitch and roll of the vessel, among other things, leads to in errors in estimating angles and distances. Angular errors are typically of the order of a degree or two; the coefficient of variation of distance estimation errors is typically of the order of 10%.

The two strata were surveyed by different vessels at the same time. Because the whales tend to be found in high densities against the ice edge, where they feed, densities in southern strata are typically higher than those in northern strata. In fact this is the primary reason for using a stratified survey design. It is also the reason for covering the southern strata more intensely; in this survey the transect length per unit area in the southern stratum, is more than 2.5 times that in the northern stratum.

Here are pictures of the sort of design used and a typical density gradient. The irregular bottom border is the ice-edge; the "steps" define the boundary between southern and northern strata; dotted lines are transects; solid dots are detections.





Analysis Exercises

Begin by opening the project from its archive **Stratify exercise.zip**. The project contains one analysis specification, called "Full geog stratification". This is a fully stratified analysis of the data. Seven equal perpendicular distance intervals, truncation at 1.5 nautical miles (nm), and a hazard rate detection function form with no adjustment terms are used to estimate the detection function. As the focus of these exercises is stratification, do not investigate other perpendicular distance intervals and detection function forms; the given models are adequate. Use the **Analysis browser** to familiarise yourself with the details of this analysis specification.

1. Having done that, run the analysis "Full geog stratification". Look at the results, and note the AIC statistics from each detection function fit.

2. To stratify *f*(0) or not to stratify?: Create a new analysis identical to "Full geog stratification" by clicking the **New Analysis** icon in the **Analysis** tab of the **Project browser** after selecting the existing analysis. The new analysis will be a copy of the existing one.

   Create a new model definition for this new analysis by going to the **Inputs** tab and highlighting the "haz rate+no adj full strat" model, then clicking the **New** tab. This will copy the existing model definition – modify the new model definition so that f(0) is to be estimated from the pooled strata (click the **Detection function** x **Global** cell of the table on the **Estimate** tab of the **Model Definition Properties** window you get after clicking **New**). Give this new model definition a suitable name and then click **OK**.

   Run the new analysis and look at the output. By comparing the AIC from this analysis with the sum of the AICs from the analysis "Full geog stratification", and considering the fits of each detection function, decide whether or not to pool strata for estimation of *f*(0).

   If you have time, here's a more difficult question.

3. Create an analysis without any stratification and estimate density using it. Why is the density estimate so much higher than those from 1. and 2. above?

Introduction to Distance Sampling

Exercise 7b: Analysis of Clustered Data

The Data

Cluster exercise.zip contains "exact" perpendicular distance and cluster size data from a survey of Antarctic minke whales (the same data as are in the project file stratify exercise.zip).

Open the **Cluster exercise.zip** project in Distance. Use the data explorer to familiarise yourself with the data (click the **Data** tab in the **Project Browser**, followed by the **Region**, then **Line Transect**, then **Observation** symbols in the left window). Ignore the "Cluster strat" data column for the moment, it is dealt with below.

Analysis Exercises

This exercise will allow you to explore some of the different methods of dealing with clustered data, as discussed in the lecture. The following methods will be used:

- Regression
- Truncation
- Post-stratification

The project contains one analysis specification, called "E(s) by ln(s)_g(x)". Use the **Analysis browser** to familiarise yourself with the details of this analysis specification. This analysis uses a regression of the log of school size (s) against the estimated detection function to estimate mean school size (look under the **Cluster size** tab in **Model Definition Properties**). Seven equal perpendicular distance intervals, truncation at 1.5 nautical miles (nm), and a hazard rate detection function form with no adjustment terms are used to estimate the detection function. As the focus of these exercises is mean cluster size estimation, do not investigate other perpendicular distance intervals and detection function forms; the given models are adequate.

<u>Using regression</u>

1) Run the analysis "E(s) by ln(s)_g(x)". Look at the results and the cluster size estimation pages in particular.

    a) Is the regression method estimate of E(s) bigger than the observed mean cluster size?

    b) What percentage of the variance of the density estimate is due to cluster size estimation?

<u>Using truncation</u>

2) Using the fitted detection function, decide on an appropriate point at which to truncate the data in order to use the mean observed cluster size as an estimate of E(s). Create a new analysis, identical to "E(s) by ln(s)_g(x)" except that it should use the truncation method to estimate E(s). To do this, click the "**New Analysis"** icon in the **Analysis browser** after selecting the existing analysis, then add a new **Data filter** in which the right truncation for cluster size estimation on the **Truncation** tab

has been set appropriately. Create a new Model Definition where the mean of the observed clusters, rather than the regression method, is used (specified in **Model Definition Properties/Cluster size)**. Having run the analysis, look at the cluster size estimation pages.

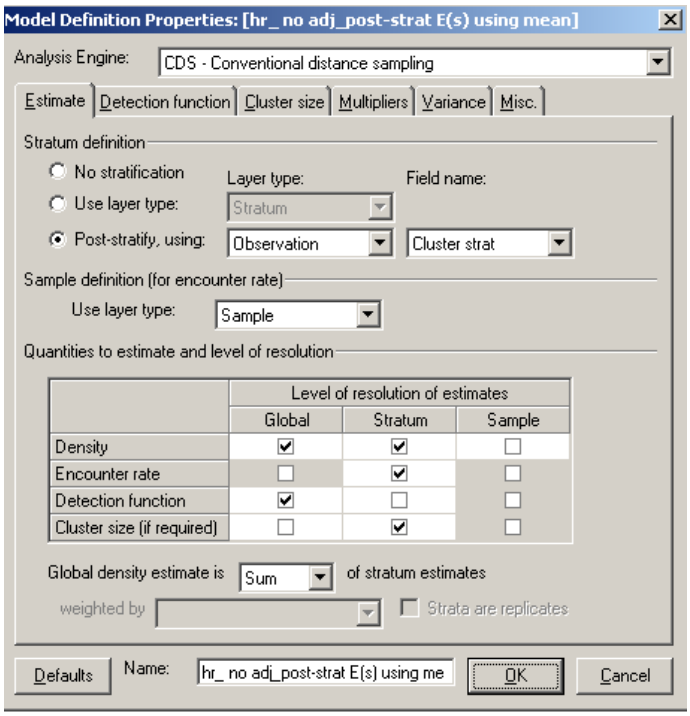
a) Why is the "Mean cluster size" on the **Cluster size/Global/Estimates** page different from the mean cluster size in analysis 1 above?

b) Why is the standard error of "Mean cluster size" in this analysis larger than that of the "Expected cluster size" in analysis 1 above? (Hint: look at the sample sizes.)

Using post-stratification by cluster size

3) Now we come to the "Cluster strat" column in the Observation layer of the data. It was added after the data were entered and is just an indicator column for stratification on the basis of cluster size. All observations with cluster size 1 have been defined to be in cluster stratum 1 and hence have 1 in the "Cluster strat" column. Similarly for cluster size 2. Due to small sample sizes it was not possible to create separate strata for cluster sizes of 3 and above. Therefore, all observations with size 3 or greater have been put in cluster stratum 3 and hence have 3 in the "Cluster strat" column.

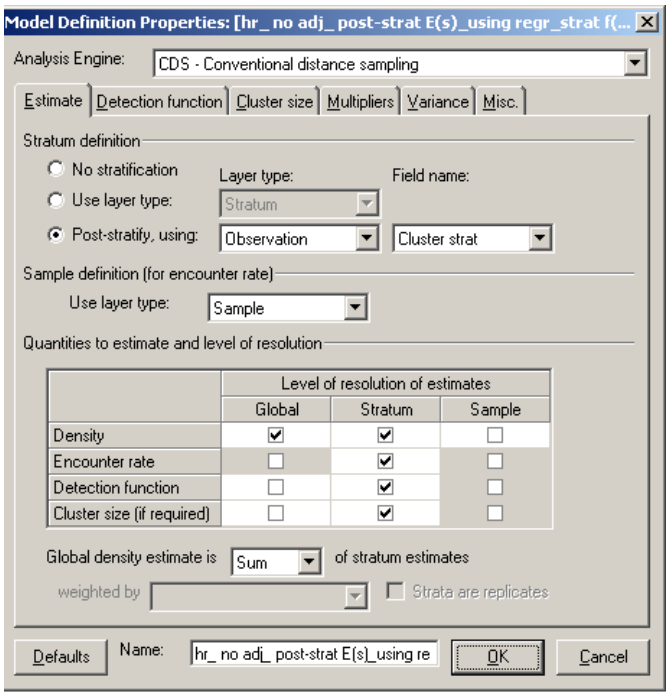
a) Use the "Cluster strat" column as a basis for performing an analysis post-stratified by cluster size. Do this by creating a new analysis with a new Model Definition that uses post-stratification at the Observation level. Fit a detection function pooled across strata, but estimate mean cluster size separately for each stratum (see the picture below for help). There should be no size bias within the strata, so theoretically it should be sufficient to use the mean of the observed cluster sizes for each stratum. Once the analysis is run, note the mean cluster size for the third stratum.



b) However, when forced to use strata that contain a range of cluster sizes due to small sample sizes (such as stratum 3 in this case), you may suspect that size

bias is still present. It is possible to use the regression method to check this. Create another post stratified analysis which uses the regression method to estimate E(s) in each stratum (again, estimate a pooled detection function and separate cluster size estimates). Compare the regression estimate of E(s) with the mean cluster size (the mean should be identical to the estimate you found in 3(a)). Does it suggest that size bias is present in this third stratum?

c) Another consideration when using regression with post stratification is the following: is the detection function you are using for the regression the correct one (recall that the explanatory variable in the cluster-size regression is g(x))? In other words, in 3(b) the pooled detection function was used for the regression in the third stratum. However, if you suspect you have size bias in the first place, then you would expect the detection function for larger and smaller cluster sizes to be different - you would expect the detection function for larger cluster sizes to have a wider shoulder (i.e. larger effective strip width and a smaller f(0)). Therefore, perform an analysis where you estimate a detection function for each stratum. Look at the results – are the detection functions different between strata? Do they seem plausible? Are you satisfied with the sample sizes used to estimate the detection functions?



Overall question: consider all the analyses conducted – which would you use for this dataset?

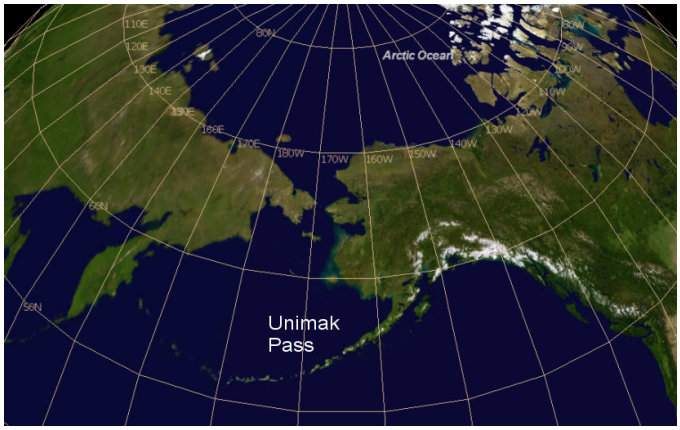
Introduction to Distance Sampling

Exercise 8: Covariates in the detection function

This exercise consists of four datasets of increasing difficulty. Everybody should work through the first dataset, and the other datasets can be examined later if you wish, or you may work on them when you complete the first analysis. The first analysis will show you the rudiments of conducting an analysis, while the remaining analyses take you deeper into the heart of understanding multiple covariates.

1 A whale of a dataset

Rather than relaxing here in the serenity and tranquility of the Scottish coast, image instead that you are a research biologist collecting distance sampling data during December on gray whales as they migrated through the Aleutian chain near Unimak Pass en route to their wintering grounds off Baja California (some luckier, more senior researcher, got the job of data collection on their wintering grounds). These data will now be the focus of your attention for this exercise examining the potential utility of covariates in explaining variation in animal detectability.



Detections were of individuals (not groups), and you chose to record not only distance, but also time of observation (at this latitude at this time of year, the crew was restricted to making observations between 1000 and 1500 during the day). However, because of the low sun angles during much of this time, there was some reason to believe that time of day might play a role in whale detectability. [In what manner might you wish to incorporate this covariate?]

Under extreme weather conditions, observer motion sickness can influence the performance of the observers. An additional covariate, "motion sickness tablet effective dosage at time of observation (MSTDO)" was recorded each time a whale was detected.

The data are available for your inspection in the Distance project **adv_practical_1.dst**. Notice the extreme precision with which the perpendicular distances were measured (how do you suppose this could happen on a rolling ship in the Bering Sea?).

Describe your candidate model set (what models did you construct) and your rationale for the final estimates you provide. You may also comment upon the use of time of observation as a measure of glare from oblique sun angles.

If you have been successful in performing the analysis of this dataset (which can now be revealed to have been simulated), you can continue to sharpen your skills in using covariates in your analysis of distance sampling data by exploring two other data sets, that are considerably more elaborate.

2 Golf tees Data

2.1 The 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. It was conducted as a double platform survey; double-platform methods are described later in the workshop and so

we will only use detections recorded for one observer (or platform) for the purposes of this exercise.

Assume that all the data were collected on one 210 metre long transect line, and that this comprises the study area. There were 250 clusters of tees in the study area and 760 individual tees in total.

The population was independently surveyed by two observer teams, of which we will use the data recorded by observer 1. The following data were recorded for each detected group: perpendicular distance, cluster size, observer (team 1 or 2), "sex" (males are yellow=1, females green=0 and golf tees occur in single-sex clusters), and "exposure". Exposure was a subjective judgment 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**. Open the project. Notice that there is already a data filter and several model definitions set up. To avoid overwriting these as they will be used in the double platform exercise, first create a new 'set' on the **Analysis** tab and then create a new analysis for that set. Open the model details for the new analysis. First, we need to select only the sightings detected by one observer – we will use observer 1 sightings. Create a new data filter and on the **Data Selection** tab in the data filter, click '+'. Choose Layer type 'Observation' from the dropdown menu that appears when you click on the cell. Under Selection criteria type 'observer=1 AND detected=1'. This selects only the detections made by observer 1. The data is already truncated at 4 metres and we will use the same truncation distance.

2.2 CDS analysis of the golf tee data

Now create a new model definition. Start by performing a conventional distance sampling (CDS) analysis using a half-normal key function. To do this, edit the new model definition. Under the Analysis Engine, choose CDS and use the default setting for the detection function. Give it a sensible name and run it.

Look at the results (in the **Analysis** details, **Results** tab). Don't worry about the warning – this is because there is only one transect and so the encounter rate variance is estimated assuming that the observations are from a Poisson distribution so that $\hat{V}(n) = n$ rather than from inter-transect variation. Make a note of the estimated abundance and associated coefficient of variation (CV). Also have a look at the percentage of variance that was due to the detection function.

2.3 MCDS analysis of the golf tee data

Create a new analysis and a new model definition. This time choose the MCDS analysis engine.
Check that under the **Detection function** tab, the selected key function is half normal and under the Adjustment terms button we have manual selection of zero adjustment terms. MCDS analyses are much harder for the analysis engine to fit than single covariate ones (and a different algorithm is used). In general, it is better to avoid automated selection of adjustment terms and use manual selection instead. Start with zero adjustments terms, and gradually build up 1, 2 etc. checking AIC or one of the other criteria to see if this gives a better fit. It is also a good idea to tick the option in the **Misc.** tab to 'Report results for each iteration of detection function fitting routine' (it is ticked by default for the MCDS engine) – this will help you to diagnose any problems that may occur during fitting.

There were 3 additional covariates recorded along with perpendicular distance; cluster size, sex and exposure. Obviously, sex and exposure are factor variables. Sometimes cluster size can be treated as both a factor variable or as a continuous variable: if there are only a few cluster sizes then it can be treated as a factor; however, if cluster size

ranged over a large number of values it would have to be treated as a continuous variable. In this data, cluster sizes ranged from 1 to 8 and it is debatable as to whether you would want to treat it as a factor variable as there are very few large clusters detected. When including cluster size don't forget to tick the cluster size box on the **Covariates** tab – this tells Distance that this covariate is the cluster size covariate. When cluster size is included as a covariate, Distance uses a 'Horvitz-Thompson-like' estimator of abundance (this will have been covered in lectures). In this case, Distance changes a number of options in the **Estimate** and **Cluster size** tabs. In **Estimate**, it changes the 'Sample definition' option and doesn't allow stratification and in **Cluster size** it removes all the options.

Select each of these terms in turn and also in combination on the **Covariates** tab. After running a model, look at the results. The presentation of results is like that in CDS analyses, with a **Log** tab where any warnings or error messages are written, and the **Results** tab which contains details of the analysis. Make a note of the AIC value and look at the detection function plots – notice the difference in the detection function plots when the covariate is specified as a factor variable or a continuous variable.

Once you have decided on the best model, make a note of the estimated abundance, associated CV and percentage of variance accounted for by the detection function. How has this changed?

3 Dolphin Sightings Data

This exercise is optional – so feel free to switch to your own data if you have some. In this example there are several potential covariates and no 'right' answers!

3.1 Reviewing the data

In this example we have a sample of eastern tropical Pacific (ETP) offshore spotted dolphin sightings data, collected by observers placed on board tuna vessels (the data were kindly made available to us by the Inter-American Tropical Tuna Commission – IATTC). In the ETP, schools of yellow fin tuna commonly associate with schools of certain species of dolphins, and so vessels fishing for tuna often search for dolphins in the hopes of also locating tuna. For each school detected by the tuna vessels, the observer records the species, sighting angle and distance (later converted to perpendicular distance and truncated at 5 nautical miles), school size, and a number of covariates associated with each detected school. Many of these covariates potentially affect the detection function, as they reflect how the search was being carried out.

A variety of search methods are used to find the dolphins, but currently the most commonly used are 20x binoculars from the crow's nest, 20x binoculars from another location on the vessel, from a helicopter, or through "bird radar" (high power radars which are able to detected seabirds flying above the dolphin schools). In the example dataset these are coded as 0, 2, 3, and 5, respectively. Some of these methods may have a wider range of search than the others, and so it is possible that the effective strip width varies according to the method being used.

For each sighting the initial cue type is recorded. This may be birds flying above the school, splashes on the water, floating objects such as logs, or some other unspecified cue. In the example they have been coded as 1, 2, 4 and 3, respectively.

Another covariate that potentially affects the detection function is sea state, as measured by Beaufort. In rougher conditions (i.e. higher Beaufort levels), visibility and/or detectability may be reduced. For this example Beaufort levels are grouped into two categories, the first including Beaufort values ranging from 0 to 2 (coded as 1) and the second containing values from 3 to 5 (coded as 2).

The sample data encompasses sightings made over a three month period: June, July and August (months 6, 7 and 8, respectively).

Begin by extracting and opening the project from the archive **Dolphin.zip**. Once it is open, you will see the **Project Browser**, from which you can have a look at the data (**Data** tab).

3.2 Analysis of Dolphin Sightings data

Start by running a set of conventional distance analyses. Are there any problems in the data and if so how might you mitigate them? (Hint – check out the q-q plot, and also try dividing the data into a large number of intervals in the Model Definition | Detection Function | Diagnostics.)

As there are a number of potential covariates to be used in this example, try fitting models with different covariates and combinations of the covariates. All of the covariates in this example are factor covariates except cluster size.

Keep in mind that this is a large dataset (> 1000 observations), and hence estimation may take a while, particularly if you are allowing up to 5 adjustment terms to be fitted. It will be generally more efficient to start fitting models without any adjustment terms, and then adding one at a time if appropriate. Consider also whether to standardize by w or by σ (or try both!).

You will likely end up with quite a few models, so think about how you are going to name and organize them in the Project Browser (for analyses) and Analysis Components window (for model definitions).

Discuss your choice of final model (or models) with your neighbours - did you make the same choices?

4 Passerine data from Marques et al. (2007)

The data from the Auk paper by Marques et al. (2007) is also available on your data stick. It is zipped as the project **fTAMAUK07.zip**. See if you can produce results comparable to those presented in the manuscript (also on your data stick).

Introduction to Distance Sampling

Exercise 9a: Analysis with the use of multipliers

The Problem

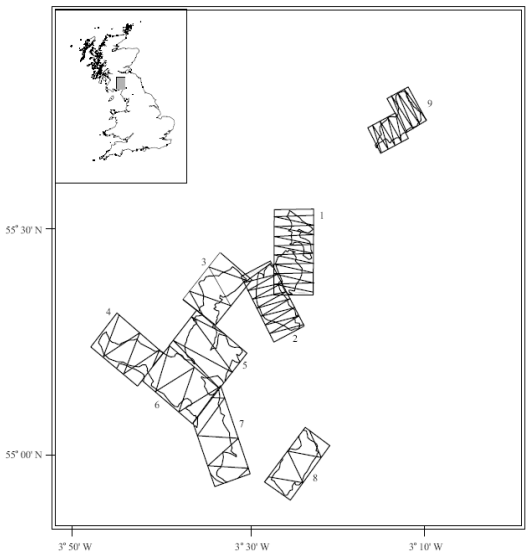
The question is how to estimate of the density of sika deer in a number of woodlands in the Scottish Borders. These animals are quite shy and often will be alert to the presence of an observer before the observer detects them, making surveys of the deer challenging. As a consequence, indirect estimation methods have been applied to this problem. In this manner, an estimate of density is produced for some sign generated by deer (faecal pellets) and this estimate is transformed to density of deer by

$$\hat{D}_{deer} = \frac{\dfrac{\hat{D}_{\text{pellet groups}}}{\text{mean time to decay}}}{\text{dung production rate (per animal)}} = \frac{\text{dung deposited daily}}{\text{dung production rate}}$$

We will produce a pellet group density estimate, then adjust it accordingly to account for the deposition and decay processes operating during the time the data are being acquired. We will also take uncertainty in the production and decay rates into account in our final estimate of deer density.

The Data

Data from 9 woodlands were collected according to the survey design shown below (note differing amounts of effort in different woodlands based on information derived from pilot surveys).



In addition to these data, we also require estimates of the defecation rate. From a consultation with the literature, we learn that sika deer deposit 25 pellet groups daily. The literature source did not provide a measure of variability of this estimate. During the course of our surveys we also followed the fate of some marked pellet groups to estimate the disappearance (decay) rates of a group. A thorough discussion of

methods useful for estimating decay rates and associated measures of precision can be found in Laing et al. (2003) [found on your thumb drive].

There are many factors that might influence both deposition and decay rates, and for purposes of this exercise we will make the simplifying assumption that decay rate is homogeneous across these woodlands; with their mean time to decay of 163 days and a standard error of 13 days. However if you were to conduct a survey such as this, you would want to investigate this assumption more thoroughly.

Pay a visit to http://www.wcsmalaysia.org/analysis/Nest_dung_decay.htm where Mike Meredith of Wildlife Conservation Society in Malaysia thoroughly describes an analysis to estimate decay rates for animal nests or dung.

Analysis Exercises

Use the Distance project **Deer pellets.zip** for the following analyses.

1. Adjust the multipliers in the project (replacing the place-holders in the project, with values provided in the previous section of this exercise).
2. Fit the usual series of models (uniform, half normal, and hazard rate) models to the data.
3. Select the Multipliers button in the Model Definition Properties to specify the layer and the field in the project database for the multipliers you wish to employ (along with their measure of precision).
4. Produce estimates using the woodland as strata, pooling data across strata for fitting the detection function, but using woodland-specific encounter rate to produce woodland-specific estimates of density.
5. Produce an overall estimate of density as mean of woodland-specific densities weighted by the effort allocated within each woodlot.
6. Make special note of the components of variance (contribution of detection function, encounter rate, decay rate, and what happened to defecation rate component?) in each of the strata.

Introduction to Distance Sampling

Exercise 9b: Cue Counting Analysis Exercise

This practical involves analysing an aerial cue counting survey of minke whales in the Atlantic. Minke whales tend to occur singly. An estimate of mean cue rate and its coefficient of variation have been obtained from tagging studies on a number of minke whales in the area.

The sample size is relatively small for a cue counting survey (which require larger sample sizes for reliable estimation of the detection function than line transect surveys), but this is the sample that was generated by the (expensive) survey, so you just need to do the best you can with it.

The data are stored in the distance project **CueCountingExample.zip**. Open the project, and click on the **Data** tab to see how the data are stored. The species code for minke whales is "W" in this project; "bss" is Beaufort sea state code. A simple analysis has been set up but not run in which data filters are used to subset the data so as to use only the data we desire. Have a look at the model definition, in particular, the 'Multipliers' tab.

Question 1: what is $\hat{\eta}$ (see presentation overheads for its meaning) and its coefficient of variation for these data?

Question 2: what is ϕ (see presentation overheads for its meaning) for this data?

Question 3: Find and fit a suitable detection function model to these data and from this estimate minke whale abundance in the survey region, together with a 95% confidence interval.

We do not describe how you ought to go about selecting a suitable model and assessing its fit (you are becoming experienced using goodness-of-fit statistics and model selection criteria). Note that there was some evidence on the survey of poor-quality distance estimation, so it is worth conducting an analysis on grouped distance data.