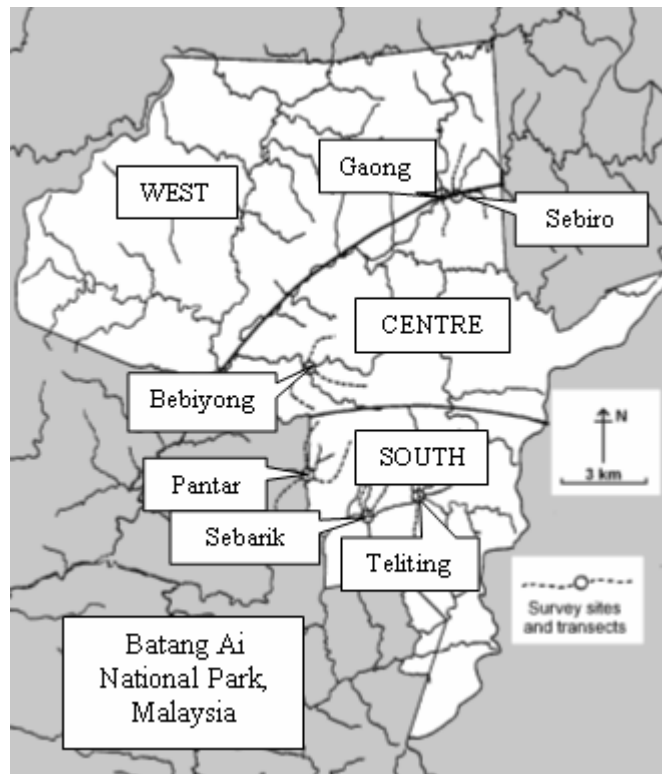


## Distance data for orang utan nests

### A. Background

For this exercise, we'll use line transect data for orang utan nests collected in Batang Ai NP (see map below) by June Rubis. The data set has the advantage of having a lot of data points and that distances from the transect could be measured accurately as the nests did not run away before June got the tape measure out.

Map of Batang Ai NP where data on orang utan nests used in this exercise were collected.



Like most great apes, orang utan build a simple nest of broken branches to sleep at night, and sometimes build a day-time nest too if they take a siesta. If we assume that there is a simple relationship between the number of nests and the number of orang utan in our study area, we can use the number of nests to compare populations between sites or to detect declines or increases in the population over time.

Open the file "O-u\_nests\_DISTANCE.csv" in MS Excel® (or other spreadsheet program) and check what it contains.

June worked from 6 camps in the park, with 2 or 3 transects leading out from each camp. This is shown in the 'Transect' column. 'Length (km)' is the length of the transect: all the transects were 2km long.

The park was divided into 3 zones, because local people reported seeing lots of orang utan in the south of the park, and fewer in the rest of the park. Surveys in 1992 sighted more animals in the South zone, few animals but many nests in the Centre, and hardly any animals or nests in the West (see map).

## ***B. Getting started with DISTANCE***

Go to the DISTANCE web site (<http://www.ruwpa.st-and.ac.uk/distance/>), and fill in the registration form. Then download the installer for the latest version of the software, currently DISTANCE 5.0 release 2 in 'd50setup.exe'. As with any .exe file, it's wise to download it, run a virus check with an up-to-date virus scanner, and create a Windows Restore Point (go to 'Start > Programs > Accessories > System Tools > System Restore') before running the setup program. Close all programs (except File Manager) as you will have to restart the computer during installation, then run d50setup.exe. Follow the on-screen instructions.

Finally, check the 'Distance support page' for any additional patches that may be needed.

If you use the results from DISTANCE in any report or paper, please cite it as:

Thomas, L., Laake, J.L., Strindberg, S., Marques, F.F.C., Buckland, S.T., Borchers, D.L., Anderson, D.R., Burnham, K.P., Hedley, S.L., Pollard, J.H., Bishop, J.R.B. and Marques, T.A. 2006. Distance 5.0. Release 2. Research Unit for Wildlife Population Assessment, University of St. Andrews, UK. <http://www.ruwpa.st-and.ac.uk/distance/>

## ***C. Importing the data into DISTANCE***

Check that the data in the file 'O-u\_nests\_DISTANCE.csv' is sorted so that all the data for each transect are together and all the transects for each zone are together. (If necessary, use Data > Sort... and sort by Zone first, then by Transect.)

In this case, nests were recorded on all of the transects. In some surveys, there may be transects where you saw no animals: in that case the file **MUST** contain a row with the name and region for the transect, with the perpendicular distance column left blank.

Close the file without saving it (unless you sorted the data, in which case you need to make sure it is saved as a ".csv" file)..

Start DISTANCE, go to 'File > New project...' and select the folder where you want to save your project - DISTANCE will start off with the Samples folder, but that probably isn't the best place to save your own projects. Name the project something like "O-u Nests".

DISTANCE saves all your work as you go along, you do not need to save frequently. (On the other hand, if you screw up and do **not** want to save your work, don't just close DISTANCE but select 'File > Revert to Backup Copy' and DISTANCE will use the backup copy created when you started the session.) If you check the target folder now, you will see three new items, files named 'O-u Nests.dst' and 'O-u Nests.ldb' and a folder named 'O-u Nests.dat'.

The Data Import Wizard will start immediately.

Step 1 : Select 'Analyze a survey that has been completed' and click 'Next'

Step 2 : This is an information page; read through it then click on 'Next'

Step 3 : The nest survey was a Line transect survey with a Single observer, where Perpendicular distances were measured to Single objects.

Step 4 : The distances from the transect are in Metres, the transect length is in Kilometres and we want the results for nest density per Square kilometre.

Step 5 : We don't need to use any Multipliers: all these check boxes should be blank.

Step 6 : Select 'Proceed to Data Import Wizard' and click on 'Finish'.

DISTANCE will create the necessary file structure, and then the Data Import Wizard will start.

Step 1 : This is an information page; check through it then click 'Next'.

Step 2 : In the 'File containing data to import' box, browse to the file 'O-u\_nests\_DISTANCE.csv' and click 'Open'.

DISTANCE uses a hierarchical structure with

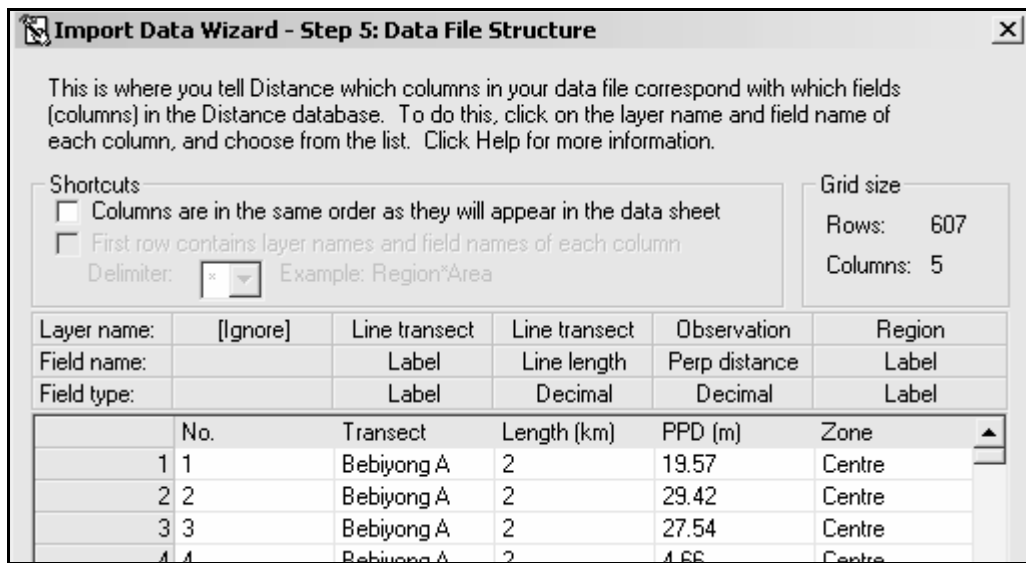
- a Global layer = study area or topic (in our case Batang Ai nests), containing...
  - Regions (our 3 Zones, West, Centre and South), containing...
    - Transects, containing...
      - Observations.

Step 3 : The 'Lowest data layer' in our data file is Observation and the 'Highest data layer' is Region. We want to Add all new records under the first record in the data layer and to Create one new record for each line of the import file.

Step 4 : In our file, the Delimiter is Comma, and we do not want to import the first row, as that is just the column titles.

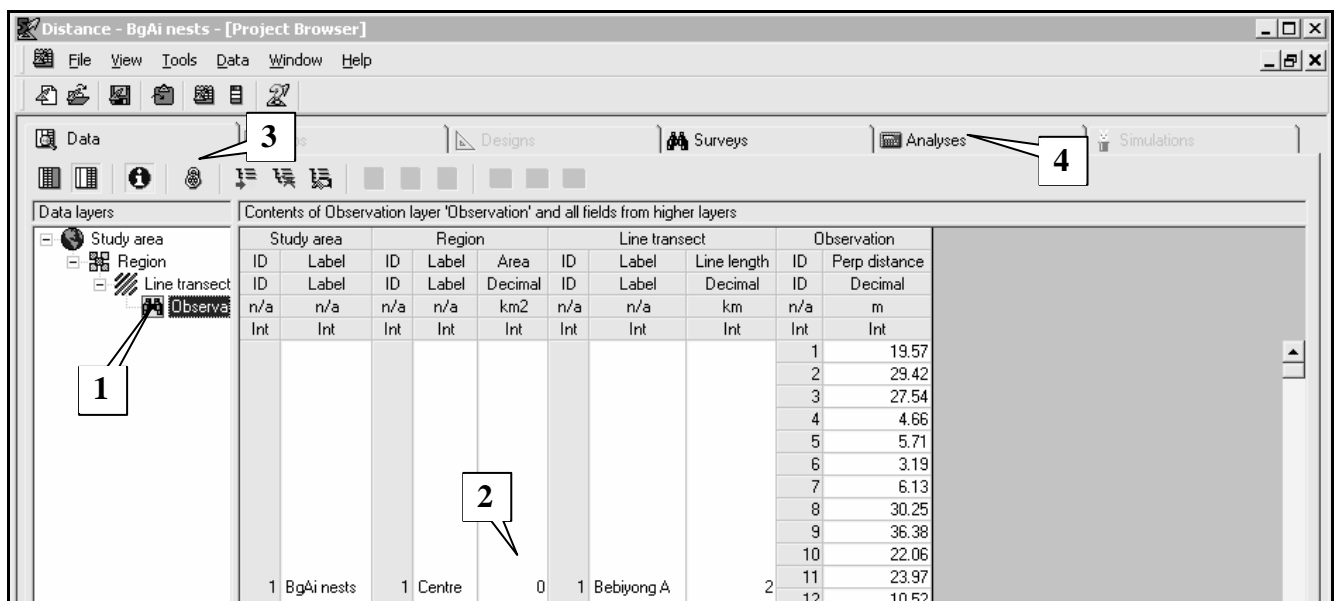
Step 5 : In our case, the columns are not in the same order as in DISTANCE, so we need to tell DISTANCE what each column means (see screen shot next page).

Step 6 : Select Overwrite existing data and click 'Finish'!



DISTANCE will take a little while to read your data file and get the data into its own database format, then the Data browser will appear (see screen shot below).

Click on "Observations" **1** in the left-hand window to display all the data and scroll down to check the contents.



You will notice that the transects all have 2km as the length, which is correct, but area for each 'Region' is "0" **2**. This will cause problems later on when we need to calculate densities for each area, so we'll enter that data now. (We could have included a column in the .txt file with zone areas, but since only 3 numbers are involved it is easier to do it like this, directly in DISTANCE.)

Click on 'Region' in the 'Data layers' panel on the left to see all three Regions conveniently. Double click on the first '0', next to 'Centre'.

You may see a box pop up with the title "Cannot edit data" explaining that the data sheet is locked. Click on the padlock icon **3** to unlock the data sheet.

Double-click on each of the 3 zeros in the Area column and enter the Areas as follows:

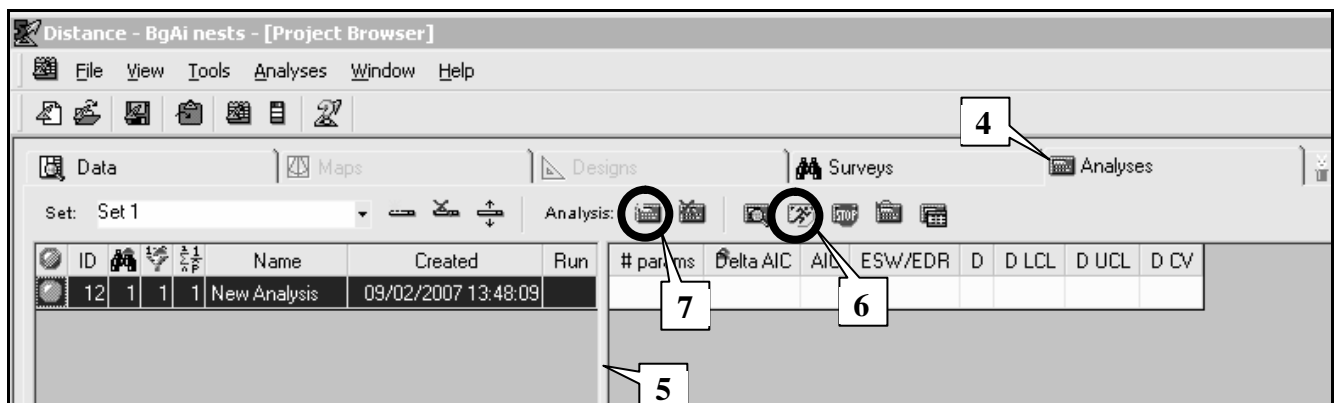
- o Centre: 74
- o South: 56
- o West: 128

Now click on the padlock icon **3** to lock the data sheet and prevent any changes being made accidentally.

### D. A first run through an analysis

We'll first run through a complete analysis just with the default settings, to get an overview of how it works.

Click on the "Analyses" tab **4** to open the Analysis Browser window. You may have to drag the divider between the windows **5** to the right to see the full names. DISTANCE has already created a "New Analysis" with the default settings, and we can run that by clicking on the 'runner' icon **6**.



The analysis will take a minute or so to run, during which time your screen may go black – don't panic! When it's finished, the grey bullet on the far left will turn to an orange color and some numbers will appear in the columns in the right-hand panel of the browser. Check that you have these results:

# params	1	These are used to compare analyses, so we'll discuss them later, when we have more than one analysis.
DeltaAIC	0.00	
AIC	4179.28	
ESW/EDR	19.02	Effective Strip Width (or Effective Detection Radius for point transects)
D	531.792	Estimate of the density of animals (or, in our case, nests)
D LCL	420.135	Lower and upper 95% confidence limits of D
D UCL	673.123	
D CV	0.112	Coefficient of Variance for D

That all looks reasonable so far. Let's have a look at the details of the analysis.

Double-click the orange bullet on the left to go to the analysis window.

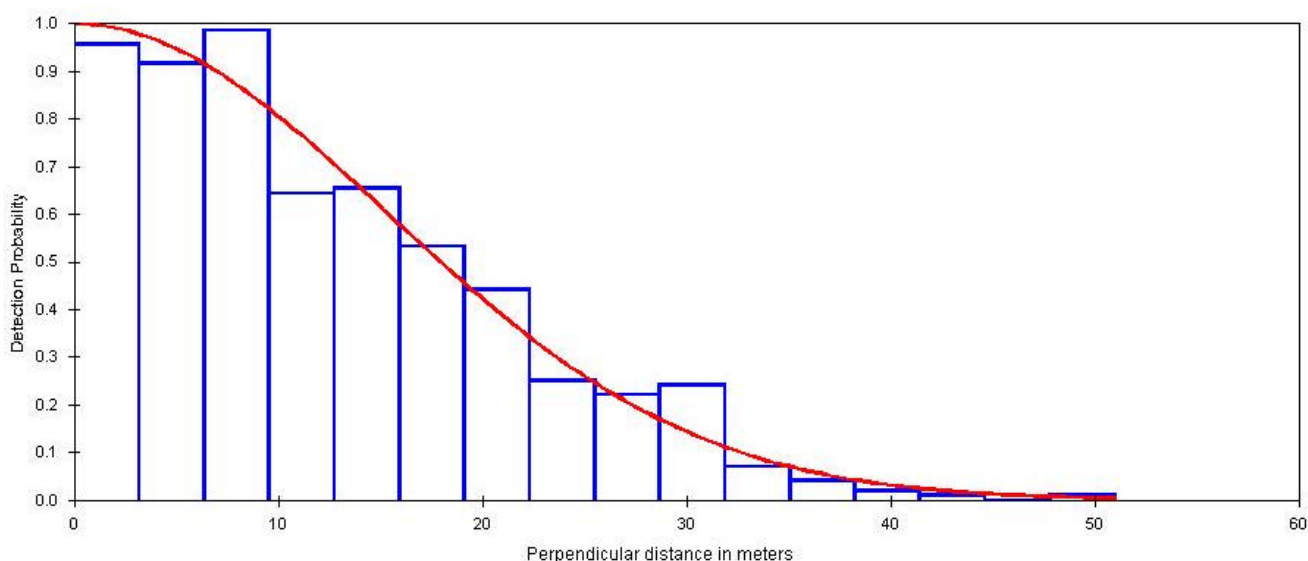
The analysis window opens with the ‘Log’ tab, which is orange, like the bullet you just clicked. That’s because the analysis produced warnings, but in this case it’s only a warning that we have too many points for all to be included on the q-q plot (more on the q-q plot later). Nothing here to worry about.

Now select the ‘Results’ tab. Click the “Next” and “Back” buttons at the top right of the window to move between the various results pages.

The first page has general information about the default analysis. Some of this is self-explanatory, some items we’ll come back to later. The bottom half of the page has a Glossary of terms and symbols which you may find useful.

On the second page, check the four numbers top left: effort = total length of the transects (15 x 2km); # samples = number of transects (15); width = in this case, the maximum perpendicular distance observed (51m); # observations = number of observations (duuh!) (607).

Flip through the pages until you find a graph that looks like the one below:



The histogram (blue in DISTANCE) shows the distribution of the observations. It shows that most observations are in the groups below 10m, with fewer and fewer observations as the distance gets larger. There is a clear ‘shoulder’ to the histogram: the observer seems to have detected all nests out to about 10m and then a decreasing proportion beyond that. The curve (red in DISTANCE) is the fitted detection function, in this case a ‘normal’ or Gaussian curve – or the right-hand half of a normal curve, to be exact – adjusted to fit the observed frequencies (the histogram) as closely as possible. If you go back to the first page, you will see under the heading “Estimators” the entry “Key: Half-normal”; this is the default for a new project. DISTANCE plots three graphs like this with different widths for the histogram bars but the same detection curve, and these sometimes show up problems. For example, the second graph has two tall bars and two short bars on the left: if the tall bars corresponded to 5m and 10m we might suspect that “heaping” was going on, i.e., the observer was actually estimating distances to the nearest 5m. All three look okay in this case.

The fit *looks* pretty good, now let’s look at the various measures of goodness-of-fit that DISTANCE supplies. You’ll see all these if you flick back and forth among the pages of the Results tab:

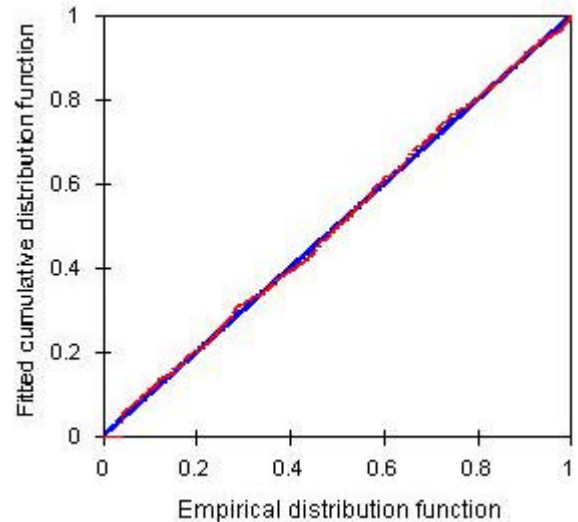
- o chi-squared tests: each of the histograms is followed by a chi-squared table, with actual and expected values for each of the bars of the histogram, with the chi-squared statistic<sup>1</sup> calculated from the difference. For the first histogram, the probability of getting a greater chi-squared value if the fitted curve is the correct model is fairly low, i.e. our data are rather a long way from the modelled values. Not a very good fit. But a point to note is that some of the expected values are pretty low: we should not be doing a chi-squared test with expected values < 4, and we should be

<sup>1</sup> The chi-squared statistic is named from the Greek letter chi ( $\chi$ ). For each bar of the histogram, calculate  $(\text{Observed} - \text{Expected})^2 / \text{Expected}$

then add up the results for all the bars.

pooling the values beyond a distance of 40m into a single group for this calculation. In fact we'll adopt a different solution to this problem, as we'll discuss below.

- qq-plot (below): this compares the actual observations (red dots in DISTANCE) with the expected values predicted by the detection function (blue line). If the fit was perfect, all the red dots would lie exactly on the blue line.
- the Kolmogorov-Smirnov and Cramér-von Mises tests are measures of how far the red dots are from the blue line, K-S using the biggest difference and C-vM taking a weighted average. Both then see how likely such a distance is if in fact the model is correct. A high likelihood (close to 1) means a good fit. The cosine-weighted C-vM test gives greater weight to the fit nearest to the transect centre line, which for us is the most important part of the curve (our density estimate depends on where the curve cuts the distance=0 axis). In this case, the K-S value is just about acceptable, and the C-vM results are okay.



The estimates of densities (and total number of nests) are calculated from two elements: the encounter rate ( $n/L$ ) and the effective strip width (ESW).

Go on to the “Estimation Summary – Encounter rates” page. This gives the encounter rate ( $n/L$ ), which is the number of nests per kilometer of transect, plus its confidence interval.

$n/L$  : number of objects (nests) observed per unit length (the encounter rate),  $607/30 = 20.233$  nests per km.

Now go on to the “Estimation Summary – Detection probability” page. This begins with some information on the fit of the Half-normal/Cosine model used to compare models: number of parameters, loglikelihood ( $\ln L$ ), AIC, and two alternatives to AIC which we'll discuss when we've run a few models to compare. You should recognize  $f(0)$  and  $p$  from the theory of distance measurements. The figure we want here is:

ESW : the effective strip width, we'd see the same number of nests if we'd spotted *all* the nests within 19.024m of the line, and none beyond.

Remember that ESW is the width of the area surveyed (51m in this case) x detection probability,  $p$ .

The last page is “Estimation Summary – Density & Abundance”,

D : the density of nests, 532 per sq km.

N : the number of nests in the whole park, based on the density x the areas we put in for the Regions, 137,200 km<sup>2</sup>.

DISTANCE also estimates variances: you will see that the %CV for the ESW is quite low, and most of the variance in the density (and number of nests) is due to variance in the encounter rate. We'll return to that point later to see if we can reduce it.

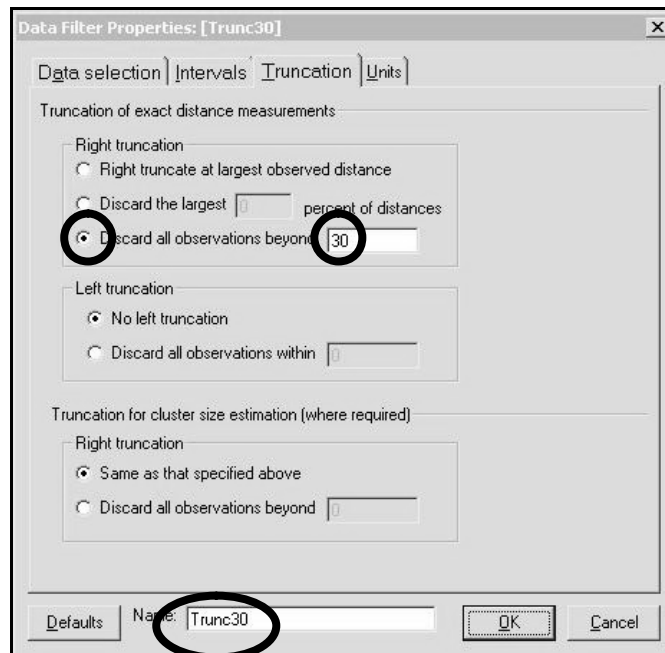
Before we try fitting different models, let's go back to the problem we saw with the chi-squared test.

### ***E. Truncating the data set***

If you look back at the histogram-plus-fitted-curve shown above, you'll see there are just a few observations beyond 40m. These we saw caused a problem with the chi-squared test, but that is really only a symptom of a bigger problem: fitting the curve to these few outlying points can distort the results which we are really interested in, which is the area close to the line. The solution is to be brutal and discard these distant observations. In fact, looking at the chi-square table, you'll see there are only a few observations beyond 30m, so we'll truncate at 30m.

Select 'View > Analysis components' from the pull-down menus to open the Analysis Components window, which has 2 pages: click on the first button on the toolbar (with the funnel icon) to see Data Filters.

The 'Default Data Filter' should be highlighted. Click on the yellow star (third icon on the lowest toolbar) to create a new data filter. Now click on the "properties" button (the 5th on the toolbar, with a hand icon) to open the Data Filter Properties Box. Select the Truncation tab (see screen shot below).



Select the "Discard all observations beyond" button and enter '30' in the field. Change the name of the filter to something short but informative, such as 'Trunc30', then click OK.

Now go back to the Project Browser (use the View menu) and the Analyses tab. Look for the New Analysis button on the toolbar (📄) in the first screen shot in section D. "A first run through an analysis") and click to create another analysis. Double-click on the grey bullet next to the new analysis.

The Inputs tab of the new analysis has several panels. At the bottom is the Model definition with the default model. Next is the Data filters panel with the old default filter and the new filter we just created. Click on the 'Trunc30' filter to highlight it. Skip the Surveys panel as we only have one survey. At the top change the name to something informative such as "Trunc30 + default model".

Then click on "Run" and wait... Ignore the warning about the q-q plot and look at the Results tab.

Check first the number of observations: 580 instead of 607, so we have discarded 27 data points. The curve still looks okay, and although the K-S and C-vM and the first chi-square tests are still not brilliant, the last two chi-square tests look more reasonable.

In the Results Browser, we can see that the ESW is about 30cm less than for the untruncated data, and the density is somewhat lower. Note that we cannot compare the AIC for these two, as they are based on analysis of different data sets.

Now let's go on to compare different models.

## ***F. Constructing and comparing models***

DISTANCE offers a range of models with different mathematical equations for the detection function – the red line in the histogram diagrams in DISTANCE. Look at the graphs in the Appendix to get an idea of what these are like. Each model is based on one of three key functions (uniform, hazard rate or half-normal), each of which has a fairly flat area near the transect line and then drops off<sup>2</sup>. To obtain a better fit, DISTANCE tries adjusting the curve by adding in other terms, either cosines or simple

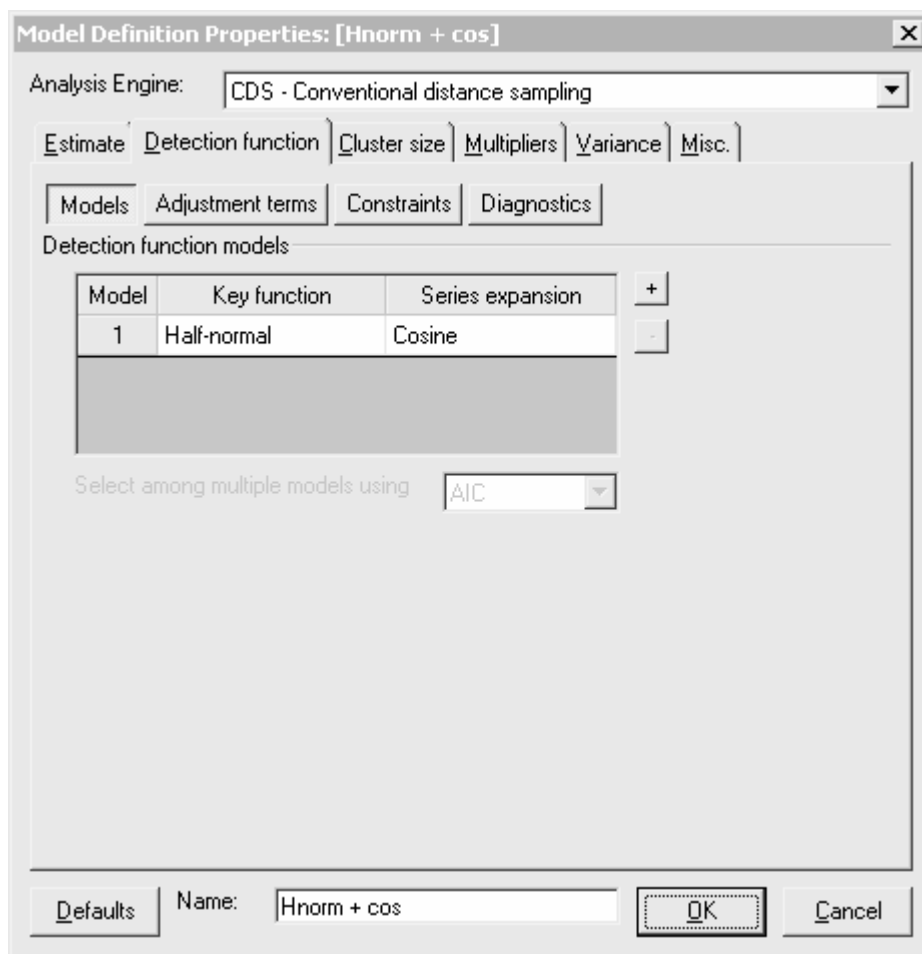
<sup>2</sup> The software can also fit a fourth key function, the negative exponential, but this does not have a flatish section at the left end. That means it gives very imprecise estimates *even if* it is a good fit to the data, and it is not recommended.

polynomial terms (of the form  $y = a + bx + cx^2 + dx^3 + \dots$ ) or Hermite polynomials, the last being similar to a polynomial but more suitable for adjusting the half-normal curve. The number of extra terms to use is decided automatically; in the example we just used, adding terms didn't improve the fit, so just the plain half-normal curve was used. (Look at the second page in the Results tab and you'll see that DISTANCE tried Model 1 with no adjustments and Model 2 with 1 adjustment and found that Model 1 was better.)

Four combinations are recommended:

- uniform + cosine (also known as the Fourier series)
- hazard rate + polynomial
- half-normal + cosine (the default option we used above)
- half-normal + Hermite polynomial

Go back to the Analysis components window, but now we want the Model Definitions page. The 'Default Model Definition' should be highlighted. Now open the Model Definition Properties box (see screen shot below), go to the 'Detection function' tab and press 'Models'.



This model is already a half-normal key with cosine adjustment; just change the name at the bottom to something like "Hnorm + cos". While we're here, press the 'Diagnostics' button and change 'Maximum num points in qq plots' to something bigger than 600, say 700. Now click on OK. A box will pop up telling you that you'll have to rerun an analysis if you change the model definition – that's not a problem, so click 'Yes'.

Click 3 times on the yellow star (third icon on the lowest toolbar) to create 3 new model definitions.

Open the properties box for the second Model Definition and change this to a Uniform key function with a Cosine series expansion, and change the name at the bottom to something like "Uni + cos". In the same way, change the other two Model Definitions to "Haz + poly" and "Hnorm + Herm".

Now go back to the Project Browser and the Analyses tab. Highlight the "Trunc30 + default model" analysis and click 3 times on the New Analysis button to create 3 more analyses.

Double-click on the grey bullet of the “Trunc30 + default model” analysis and go to the Inputs tab. At the bottom of the page you’ll see the 4 models definitions, with the “Hnorm + cos” model highlighted – that’s the default model that we renamed. Check that “Trunc30” is highlighted in the Data filter panel, then change the name of the analysis to, say, “Trunc30 + Hnorm + cos”, to indicate which model it uses.

Open another new analysis, but now in the Model definition panel at the bottom click on “Uni + cos” to highlight it, then change the name of the analysis to “Trunc30 + Uni+ cos”. Do the same with the remaining two analyses, changing the Model Definitions so that we have one analysis for each model.

Run all the analyses with grey buttons: highlight all of them with shift-click or ctrl-click, then click the “run” button.

Let’s begin with the results in the Analysis browser, ignoring for the moment the first analysis with the untruncated data, which you can now delete if you wish (use the “Delete selected analysis” button on the toolbar).

The four should be in order with the one with the lowest AIC at the top (if it isn’t, click on the heading of the AIC or deltaAIC column). The “Trunc30 + Uni + cos” model is the best, with an AIC of 3822.05, but the two Hnorm models are close behind. In fact, they are identical: the number of parameters (1) tells us that one model is the Hnorm model with no cosine terms and the other is the Hnorm model with no Hermite polynomial terms! The Haz model is more than 2 AIC units behind, so we don’t need to consider it seriously any further.

If you don’t get the same results as this, check the analysis setup. This is a bit complicated and it’s easy to get the different filters, key functions and adjustments mixed up. To check, look at the second page of the results, headed “DetectionFct/Global/Model Fitting”.

- o Near the top, the Width should be 30.000 and the # observations 580 if you have truncated the data at 30m.
- o Further down is the heading “Model 2” and the next two lines tell you the key being used (Uniform, Half-normal or Hazard Rate) and the type of adjustment (Cosine, Hermite polynomial or Simple polynomial).

Check that these correspond to the name of the analysis.

The difference in the density estimates for the two best models is small, about 2% of the value. The coefficient of variance (“D CV”), however, is still high, though about the same for all the models. Remember that we saw earlier that most of the variance in the density estimate is due to the variance of the encounter rate ( $n/L$ ) between the different transects, which is the same whatever model is chosen for the detection function. In the next section we’ll look at a possible way to reduce this, but first you should take a look at the detailed results for the three “Trunc30 + ...” analyses.

## ***G. Differences between zones***

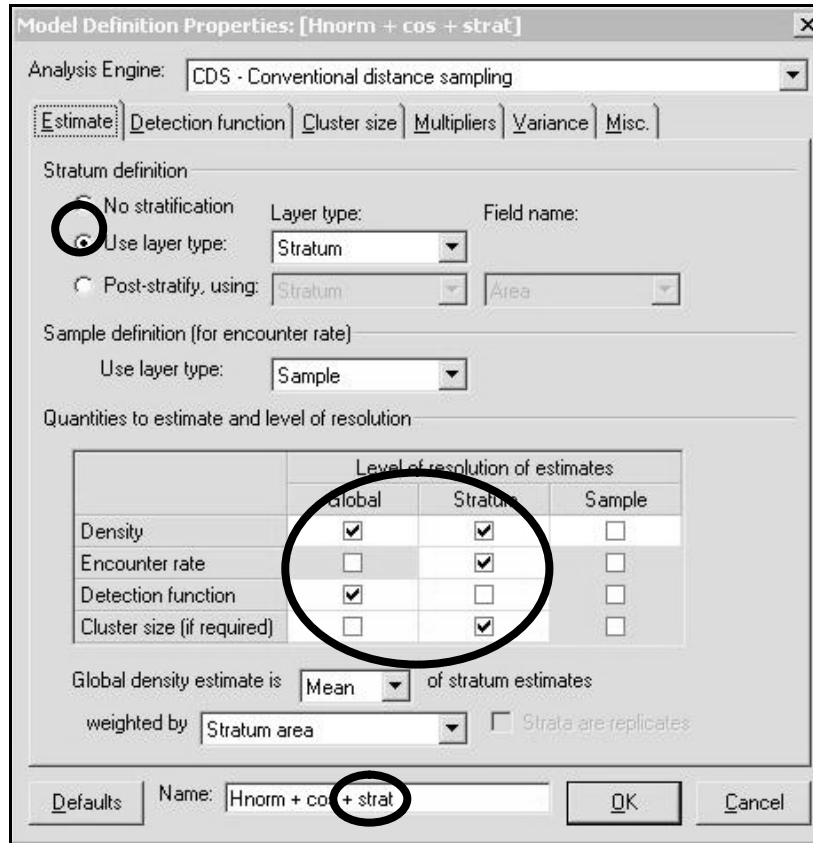
The park was divided into 3 zones on the basis of information we had before doing this survey.

Ideally we should do separate analyses for the data for each zone (i.e., stratifying by zone). However, we only have 34 observations from the West zone, not enough to calculate a detection curve properly. So our stratification will only affect the encounter rate. We saw above that most of the variance (CV) in the density estimate was coming from the encounter rate, and calculating that separately for each stratum might help to reduce the CV.

We need to create new Model Definitions for the stratified analysis:

Open the Analysis Components window and select the Model Definition page. Highlight one of the models and click on the “new” button. Open the Properties box for the new model and select the “Estimate” tab (see screen shot below). Click on the “Use layer type:” button and make sure that Stratum is selected as the Layer type. Then, under “Quantities to estimate...” check the boxes as shown: the Encounter rate

should be calculated separately for each stratum, but the detection function should be the same for all strata, i.e., it should be global. Change the name, putting "strat" at the end, then click OK.



Repeat this with each of the other three models, so that you have a set of four "...+ strat" models.

Now go back to the Analysis Browser, create 4 new Analyses, and set them up with "Trunc30" as the filter and one of the "...+strat" models. Run all four.

When they have run, take a look at the Results Browser. You'll see that stratification makes no difference to the AIC value. That's because AIC only reflects the detection curve fitting process, and the stratified model uses the same detection curve. (This would not be the case if we had asked DISTANCE to calculate separate curves for each zone.)

Double-click on the green button next to the "Trunc30 + Uni + cos + strat" model and go to the Results tab.

Compare the results for the stratified analysis with the unstratified results.

## H. Getting finished

DISTANCE automatically saves each of the Filters, Model Definitions and Analyses and the results as you create them or run them, so there is no need to save your work before closing. If you do **not** want to save the work you have done during the current session, you can select 'File | Revert to Backup Copy' and DISTANCE will use the backup copy created when you started the session.

To close DISTANCE, select 'File | Exit' or click on the  button in the top right corner of the main DISTANCE window. When the "Are you sure..." box appears, click on 'Yes'.

Detection functions for the orang utan nests data set

	Uniform + Cos	Hazard rate + polynomial	Half-normal + cosine	Half-normal + Hermite
0				
1				
2				
3				
4				
5		<p>● = Generated an error message.</p> <p>▲ = Generated a warning message</p>		