

# Analyzing camera trap data with PRESENCE

## A. Camera trap data for Golden Cats

From December 1997 to October 1999, a major camera trapping survey was carried out at 9 sites in Peninsular Malaysia (Lynam et al 2007). (See map below.) The main purpose of the survey was to ascertain the status of tigers, but of course the camera traps picked up many other species. In this unit we will look at the camera trap results for golden cats (*Catopuma temminckii*), and see if there is a difference in occupancy for different habitats.

Open the data file, "Golden\_cats\_PRESENCE.xls", in Excel or other spreadsheet software and look at the first worksheet, "raw data".

The "raw data" worksheet gives details of 171 camera trap sites in 9 regions ('zones').

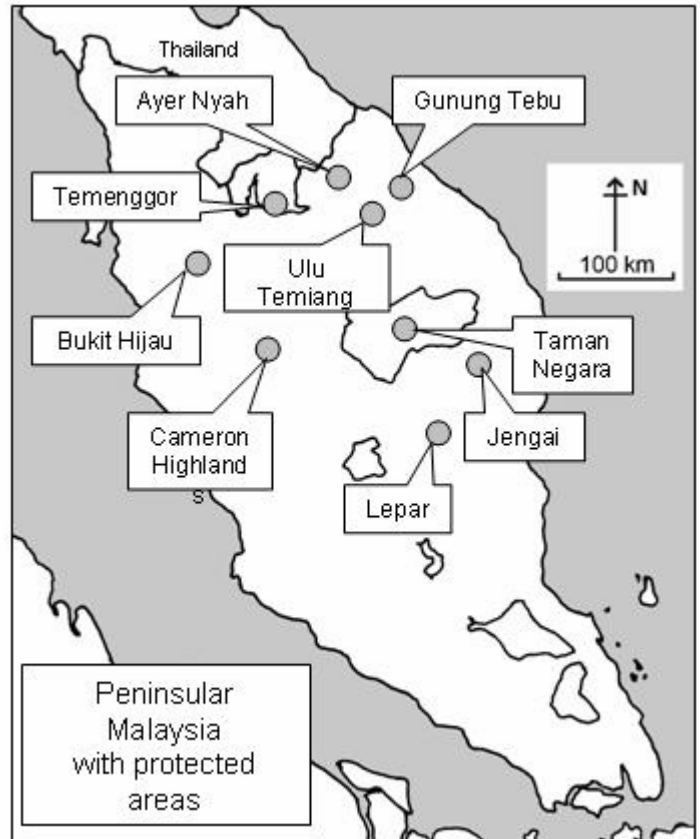
The habitat column indicates for each site if it is logged forest (with the date of logging if known), primary forest (neither logged nor cultivated in the past, with an assessment of quality), or a plantation.

The record of camera trapping has "1" or "0" for each day that the camera was deployed: "1" if a golden cat was photographed, and "0" otherwise. In several cases, the date-time recording system failed, and data from those cameras are not included in the spreadsheet. Cameras were generally operating for a month, but in one case 86 days, and in another case the camera failed after just one day. Fortunately PRESENCE does not need equal numbers of observations from each site.

We need to 'clean up' the data before passing it to PRESENCE.

The "raw data" are results for consecutive days at each site, so we can't regard them as independent observations. We can improve things by grouping the results into 5-day periods, so that if a golden cat visits a site on consecutive days (as at site T10) it scores one "1" instead of two. (The same applies if golden cats fail to visit a site on consecutive days.) On the "data for PRESENCE" spreadsheet, the results have been grouped into 5-day periods. A couple of sites with less than 5 nights of data have been eliminated from the data set.


We do not have the date of logging for all sites, or an indication of the quality of the primary forest at all sites, so we will simply group sites as "logged" or "primary." Once the sites are sorted by habitat, it's soon obvious that no golden cats were recorded at any of the plantation sites. This could be because golden cats never venture into plantations, or because the probability of photographing one is too low. With no data to go on, PRESENCE will not be able to tell you, so we exclude those sites from the analysis. We'll come back to the plantations data later, when using simulations.



This leaves us with 162 camera sites, split between primary and logged habitats, with detection data for up to 17 occasions (each 5 days). This is in the worksheet “data for PRESENCE” within the file “Golden\_cats\_PRESENCE.xls.”

## B. Importing the data into PRESENCE

Download and install PRESENCE as described on the wcsmalaysia.org web site ([http://www.wcsmalaysia.org/stats/Software\\_summary.htm#PRESENCE](http://www.wcsmalaysia.org/stats/Software_summary.htm#PRESENCE)).

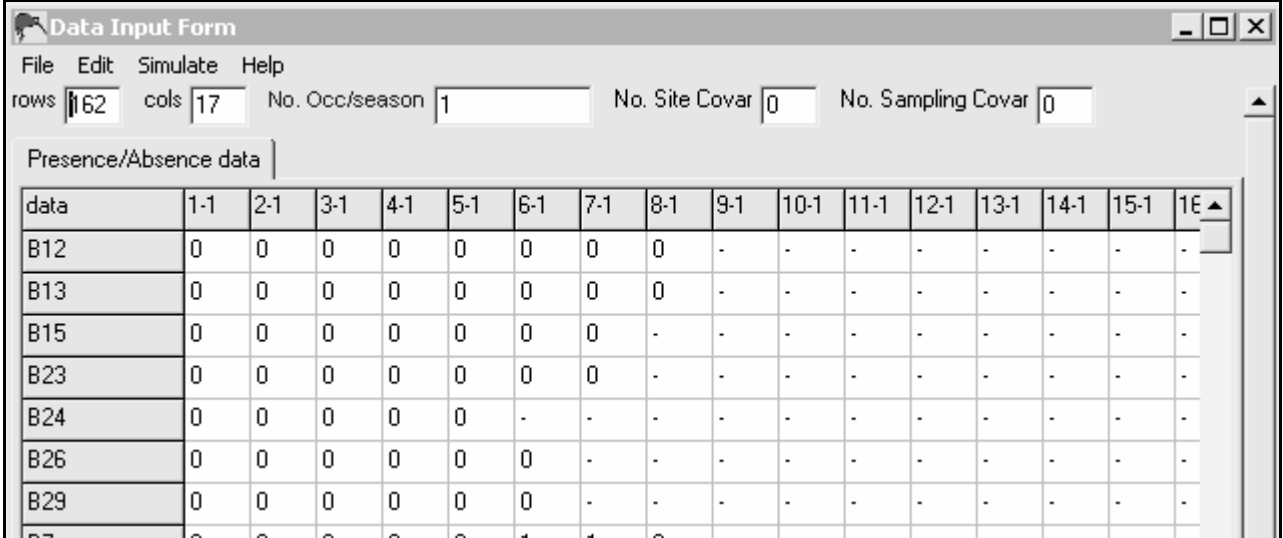
 PRESENCE is still being actively developed, with modifications being made every few months. The date and time of the version is given at the top of the main window in PRESENCE. This lab guide has been updated for version 2.1 dated 080116.1316. There have been changes in the way data are entered, and the results for model fit are completely different (and much better). If you don't have this version (or a later version), you should upgrade.

Open PRESENCE and select 'File > New project' from the pull-down menus.

Type “Golden cats” (or something similar) in the “Title for this set of data” box.<sup>1</sup>

Click on the 'Input Data Form' button.

The data input form looks like a spreadsheet with 20 rows and 4 columns. The number of columns and rows will be adjusted when we copy and paste the data from Excel into the data input form.



data	1-1	2-1	3-1	4-1	5-1	6-1	7-1	8-1	9-1	10-1	11-1	12-1	13-1	14-1	15-1	1E
B12	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-
B13	0	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-
B15	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-
B23	0	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-
B24	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-	-
B26	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-
B29	0	0	0	0	0	0	-	-	-	-	-	-	-	-	-	-
B37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Copy and paste the data from the “data for PRESENCE” spreadsheet into the ‘Presence/Absence Data’ form:


- o In the “data for PRESENCE” sheet in Excel, select all the camera data in columns A to R, including the first column of site names, but not the column headings. Press Ctrl-C to copy to the clipboard.
- o Back in PRESENCE, paste in the site names as well as the data: use Edit > Paste > Paste w/sitenames (you can't use Ctrl-V to paste in PRESENCE).

Change the ‘No. Site Covar’ to 2 (see screen-shot below); an additional tab named ‘Site Covars’ appears. Click on this and you'll see it has 2 columns.

- o In Excel, select columns T and U, including the column headings, and copy to the clipboard.
- o In PRESENCE, use ‘Edit > Paste > Paste with covnames’.

<sup>1</sup> In version 2.1 you need to do this *before* entering the data.

siteco	Primar	Logge
B12	1	0
B13	1	0
B15	1	0
B23	1	0
B24	1	0
B25	1	0

 Save the data to a file before you close the Input Data Form, or you may have to start over! Select 'File > Save as' from the pull-down menus and save it as a .pao file (eg. 'Golden cats.pao').


Now close the form (select 'File > Close' or click on the  button in the top right corner.)

Back at the Specification window, you will see that the values for “No. Sites” = 162, “No. Occasions” = 17, “No. Site Covariates” = 2 and “No. Sampling Covariates” = 0 have been updated to match the data we have just entered. The box “No. Occasions/season” is unchanged, as we do not have data for more than one season.

Now use the 'Click to select file' button and browse to the .pao file you just saved. The name of the output file (the same name with a .pa2 extension), should then appear.

Click on 'OK'.

PRESENCE will read your data file and a Results Browser window will appear – with no results in it until we run some analyses.

 Do not close the Results Browser until you have finished the session, as the main PRESENCE window will also close without warning. If you do close down by mistake, don't panic: your work will be saved automatically and you can restart PRESENCE and reload the project.

## C. A simple analysis

Select 'Run > Analysis:single-season' from the pull-down menu to open the 'Setup Numerical Estimation Run' window.

We'll begin with the default model, ie. we won't change anything, but we will look at the setup first:

Click on Custom in the Models section and the Design Matrix will appear.

This looks like a small spreadsheet with tabs for Occupancy and Detection. The Occupancy matrix has one row, labeled 'psi', and one column, labeled 'a1'. The probability that a site is occupied is usually represented by  $\Psi$ , the Greek letter 'psi'. a1 is a parameter that PRESENCE will calculate. This model assumes that all the sites have the same probability of occupancy, so PRESENCE only needs to calculate one number and then  $\text{psi} \sim 1 * a1$ . The Detection matrix also has one column, labeled 'b1', which is again a parameter which PRESENCE will calculate. It has rows for each of the 17 survey occasions, and the column is filled with 1's. This model sets all the detection probabilities for all the surveys to the same value,  $p \sim 1 * b1$ .

Note that we use a '~' sign, meaning 'varies as', instead of '='. PRESENCE is going to juggle the variables a1 and b1 to get the best fit to the data, but p and psi are probabilities, and can only vary between 0 and 1. To make sure p and psi stay in the right range, a 'logit link' is used:

$$\log_e(\text{psi} / (1-\text{psi})) = 1 * a1$$

$$\log_e(p / (1-p)) = 1 * b1$$

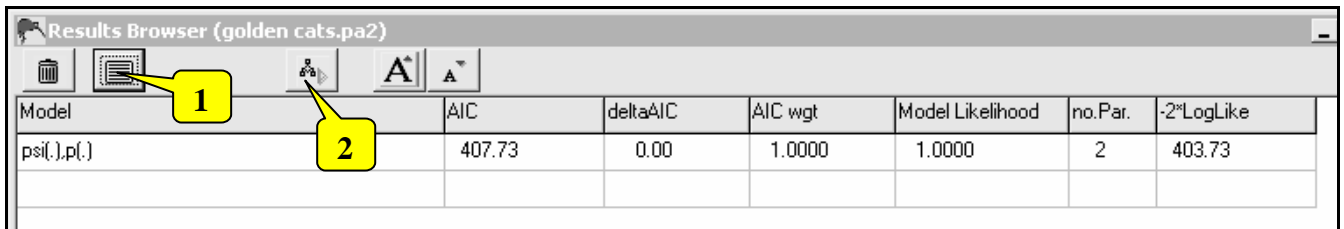
$\psi = 0$  corresponds to  $1 * a_1 = -8$  and  $\psi = 1$  corresponds to  $1 * a_1 = +8$ , and the same for  $p$ .

PRESENCE works with the terms on the right hand side of the equation (called the 'linear predictor') and then translates the results back into values for  $\psi$  and  $p$  in the model output.

Return to the 'Setup Numerical Estimation Run' window, and you'll see that the name ' **$\psi(\cdot), p(\cdot)$** ' has been entered for the default model. We'll see what that means later. Make sure there's something in the "Title for analysis" box: type in "Golden cats" if necessary. Leave all the tick-boxes blank and click 'OK to run'.

PRESENCE now juggles around with  $a_1$  and  $b_1$ , finding the likelihood of getting this particular set of presence / absence data for each combination, and selecting the combination which gives the maximum likelihood.

When a small box appears asking if you want to add the results to the Results Browser, click 'Yes'.



Model	AIC	deltaAIC	AIC wgt	Model Likelihood	no.Par.	-2*LogLike
<b>psi(.),p(.)</b>	407.73	0.00	1.0000	1.0000	2	403.73

The  $\psi(\cdot), p(\cdot)$  model results appear in the Browser. On the right is the Likelihood of the values of  $a_1$  and  $b_1$  selected based on this set of presence / absence data, expressed as  $-2 * \log(\text{Likelihood})$ ; it's 403.73.

If you're not familiar with the concepts of "likelihood" and "maximum likelihood estimation", take a look at "Frogs in ponds – maximum likelihood estimators" on the [www.wcsmalaysia.org/stats](http://www.wcsmalaysia.org/stats) web site.

Next to the  $\log(\text{Likelihood})$  is the Number of Parameters used: it's 2 ( $a_1$  and  $b_1$  in the Design Matrix). On the left is the AIC or Akaike Information Criterion, which is

$$-2 * \log(\text{Likelihood}) + 2 * \text{No. of Parameters}$$

The AIC and the other numbers in the table are useful when we want to compare models, so we'll run a few more before we explain what they are.

And you thought PRESENCE was going to tell you how many of the camera trap sites were occupied by golden cats? Well it does, but almost as an after-thought!

Click on the model name in the Results Browser to highlight it, then click on the 'View model output' button (1 in the screen-shot above).

A Notepad window opens with all the gory details. The first part summarizes what you put in, including the Design Matrices. Scroll down to the section headed 'Custom model', where you'll find:

Naive estimate = 0.2160

Golden cats were photographed at 35 of the 162 camera trap locations, so the "naïve estimate" of occupancy is  $35/162 = 0.2160$ . But that assumes that they were absent from the other 127 locations, when there's a distinct chance that one or more may have golden cats which went undetected.

Towards the bottom of the file, you'll find:

Individual Site estimates of Psi:

Site	Survey	Psi	Std.err	95% conf. interval
1	B12	1	1-1:	0.3472 0.0657 0.2315 - 0.4842

Individual Site estimates of p:

Site	Survey	p	Std.err	95% conf. interval
1	B12	1	1-1:	0.1394 0.0276 0.0935 - 0.2028

Don't be fooled by the "Individual site estimates..." bit! We told presence to calculate one value of psi and one value of p for all sites, and that's what it's done. These are the results for site "B12" – which happens to be first on the list – and the other sites are all the same.

PRESENCE's best estimate of occupancy (psi) is 0.347, with 95% confidence interval (95% CI) of 0.232 to 0.484. This is higher than the naïve estimate, as some sites where no golden cats were detected were probably occupied.

The probability of detection (p) is 0.139 with 95% CI of 0.094 to 0.203; this refers to the probability of photographing a golden cat at least once in 5 days (if the site is occupied), and that's quite low.

With just 2 parameters, this is a bit too simple, but it gets more interesting with more complex models, which we'll now run.

### D. More complex models

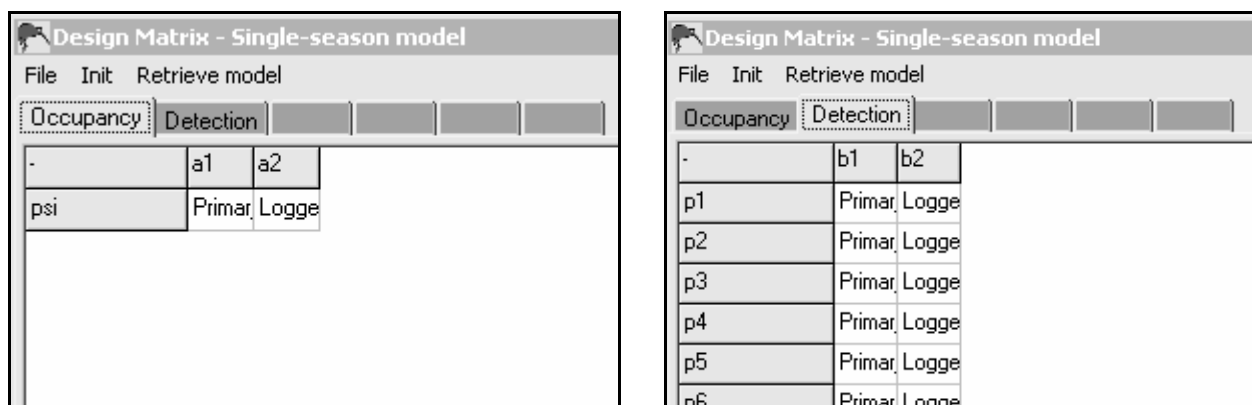
Click on the 'Run' button in the Results Browser window (2 in the screen shot) to open the 'Setup Numerical Estimation' window again. If Custom is selected, the Design Matrix should also open, but it might be hidden behind the 'Setup...' window.

We want to tell PRESENCE to calculate separate occupancy parameters and separate detection probability parameters, one of each for 'primary' and 'logged' habitats.

In the Occupancy tab of the Design Matrix, make an extra column: right-click anywhere in the window and select Add col from the context menu.

Click in the first cell in the matrix (it already has a "1"), and select 'Init > \*Primary' from the pull-down menus. In the same way select 'Init > \*Logged' for the second cell.

Do the same in the Detection tab; note that 'Init' will fill the whole column each time.



We're now asking PRESENCE to model psi using two parameters, a1 and a2, where

$$\text{psi} \sim a1 * \text{Primary} + a2 * \text{Logged}$$

and also to model p using two parameters, b1 and b2, where

$$p \sim b1 * \text{Primary} + b2 * \text{Logged}$$

If you look back at the columns in the Excel spreadsheet, you'll see that Primary = 1 for the primary forest sites, and 0 otherwise, so a1 and b1 apply only to the primary forest sites. Similarly, a2 and b2 apply only to the logged forest sites.

PRESENCE will thus calculate separate psi and p values for each of the two habitat types, just as if we had pulled out the data for each habitat and run a separate analysis for each.

Go back to the 'Setup ...' window and change the model name (which is still 'psi(.),p(.)') to **psi(habitat),p(habitat)**.

This name indicates that we have put in habitat covariates for both psi and p. In the first model, which was called 'psi(.),p(.)', there were no covariates.

Click on 'OK to run' and let's see what we get.

Once the model has run, which may take a few moments, it will appear in the Results Browser below psi(.),p(.). PRESENCE automatically puts the model with the lowest AIC at the top of the list. The best model in the list has the lowest AIC, i.e. the lowest value for  $-2 \cdot \log(\text{Likelihood}) + 2 \cdot \text{No. of Parameters}$ . If you aren't familiar with AIC (Akaike's Information Criterion), take a look at "Frogs in Ponds – AIC and likelihood" on the WCS Malaysia website.

Left-click in the psi(habitat),p(habitat) row to highlight it, then click on the 'View model output' button (1 in the screen-shot).

The naive estimate is of course the same, 0.2160.

Now we really do have Individual Site estimates for Psi and p, but the first 69 are all the same, as are the last 93:

Individual Site estimates of Psi:

	Site	Survey	Psi	Std.err	95% conf. interval
1	B12	1 1-1:	0.2991	0.0967	0.1473 - 0.5131
2	B13	1 1-1:	0.2991	0.0967	0.1473 - 0.5131
3	B15	1 1-1:	0.2991	0.0967	0.1473 - 0.5131
	...				
70	A10	1 1-1:	0.3801	0.0886	0.2269 - 0.5616
	...				

Individual Site estimates of p:

	Site	Survey	p	Std.err	95% conf. interval
1	B12	1 1-1:	0.1463	0.0528	0.0697 - 0.2817
	...				
70	A10	1 1-1:	0.1367	0.0323	0.0847 - 0.2131
	...				

Let's try the remaining two possibilities, a single estimate of Psi for both habitats but separate p's, and a single estimate of p with separate Psi's.

Click on the 'Run' button in the Results Browser window (2 in the screen shot). Custom should already be selected and the Design Matrix should appear (if it doesn't, click on 'Pre-defined' then on 'Custom' again.)

Select 'Retrieve Model > Refresh model-list from results' from the pull-down menu at the top of the Design Matrix window, then again 'Retrieve Model > %psi(habitat),p(habitat)': the setting for that model will appear in the design matrix.

Go to the Occupancy tab, right-click and select Del Col, so there is only one column left. Then use 'Init > Constant' and the column will fill with "1"s. Now return to the "Setup..." window, name the model psi(.),p(habitat), and run it.

Repeat the above steps for the Detection tab, leaving just a column of "1"s. This is the psi(habitat) p(.) model.

We have four models in the Results Browser, which should now look like the screen shot below.

Model	AIC	deltaAIC	AIC wgt	Model Likelihood	no.Par.	-2*LogLike
psi(.),p(.)	407.73	0.00	0.4975	1.0000	2	403.73
psi(habitat),p(.)	409.31	1.58	0.2258	0.4538	3	403.31
psi(.),p(habitat)	409.63	1.90	0.1924	0.3867	3	403.63
psi(habitat),p(habitat)	411.28	3.55	0.0843	0.1695	4	403.28

Check the detailed output from the two new models – there don't seem to be any problems with the analysis. Summarize Psi and p (and put their 95% confidence intervals in brackets) for all four models in the table below:

Model	Psi		p		deltaAIC	AICwgt
	primary	logged	primary	logged		
psi(.),p(.)						
psi(habitat),p(.)						
psi(.),p(habitat)						
psi(habitat),p(habitat)						
<b>weighted mean</b>						

Looking at the screen shot above, the psi(.),p(.) model is the winner, but the other models are not far behind. The AIC-weights can be used to calculate weighted model average parameters, and entered in the last line of the table above.

### E. Assessing model fit

AIC and its relatives in the Results Browser only tell us which of the models we have run is the best. But it might be the best of a bad bunch: it is possible that none of the models is anywhere near a good fit. We need to check the fit of the model which contains all the covariates [in our case, psi(habitat),p(habitat)]; all the models with fewer covariates are 'nested' within this one. If the top-level model fits the data well, the others will also be okay.

PRESENCE has an option to "Assess Model Fit" in the 'Setup...' window.

Run the psi(habitat),p(habitat) model again, but this time check the box next to "Assess Model Fit". You will need to give it a different name, eg. add "GOF" to the end; PRESENCE won't let you overwrite the old model results.

A window pops up with the number of bootstrap replications. The default is 100, which takes a minute or so to run, and we'll use that to begin with – click on OK. It would be better to run more, up to 10,000, but that could take up to 1 hour – try it when you have time. (If PRESENCE is taking a long time, you can get a 'progress report' by selecting the icon on the Windows task bar; you can also terminate the run and close PRESENCE by pressing Ctrl-C.)

Add the "GOF" model to the Results Browser and delete the old psi(habitat),p(habitat) model: left-click to highlight it, then click on the trashcan icon at the top of the Results Browser window. If you don't do this,

PRESENCE will calculate the AIC-weights for all five models, without realizing that the last two are the same.

Look at the results for the new model in Notepad.

Right at the bottom of the results page you will see:

```
Test Statistic      =          6.2651
Probability of test statistic >= observed
from 100 parametric bootstraps = 0.2926
```

(The bootstraps use random numbers, so you will get a different results for that.)

Assessing model fit is a two-stage process. First we calculate the chi-squared test statistic, which is a measure of how far the observed data are from the values calculated from our  $\psi(\text{habitat}), p(\text{habitat})$  model – it’s a “badness of fit” measure; we want this to be small. It’s not obvious what’s ‘small’ and what’s ‘big’, so the second stage is to simulate some data based on the model and random numbers, and see what range of values we get.

### **Stage 1**

Find the “Assessing Model Fit” heading in the results file. Just below this is the calculation of the test statistic for the observed data.

The left column has the capture history, the row of zeros and ones as in the original data. The first row refers to sites where cameras were out for 8 five-day periods and photographed no golden cats. There were 11 such sites, as shown in the ‘Observed’ column. The model with the estimated values for  $\Psi$  and  $p$  predicts that, on average, we would get rather more, 15.1284, as shown in the ‘Expected’ column. The ‘Chi-square’ value is:

$$(\text{Observed} - \text{Expected})^2 / \text{Expected}$$

which for the first row is 1.126602, rounded to 1.13.

Some rows have blanks in the last three columns. This is because the Expected value is small ( $< 2$ ), and small Expected values on the bottom line of the equation can give huge values for Chi-square, even if Observed and Expected are very close. So these sites are grouped together (‘pooled’), Chi-square is calculated from the sum of the Observed and Expected values and reported in the rows labeled “\*\* pooled \*\*”.

The output from stage 1, Test Statistic of 6.2651, is the sum of the ‘Chi-square’ column.

### **Stage 2**


Immediately below the 4 columns, you will see 100 rows beginning “Test statistic = ”. These correspond to the 100 bootstraps.

For each bootstrap, PRESENCE uses random numbers to produce a simulated data set based on the parameters calculated – the two values of  $\Psi$  and the two values of  $p$ . The simulated values are compared with the Expected values in exactly the same way as the actual Observed values, and a chi-square test statistic is calculated.

Taken together, the 100 bootstraps give us an idea of the kind of scatter we might get just by chance, when the model and the values for  $\Psi$  and  $p$  are correct. About 30% of the bootstrap values are higher than 6.2651, so this value is not especially big, and the difference between the actual Observed values and the Expected values could well be due to random scatter. (If only a small proportion of the bootstrap values (say  $< 5\%$ ) were bigger than the observed value, we’d conclude that the model was not a good fit, and the calculated values of  $\Psi$  and  $p$  were likely to be badly in error.)

## F. Getting finished

PRESENCE automatically saves all the results of analyses when you run them in a .pa2 file. You do not need to save results manually.

To exit PRESENCE, select 'File > Exit' or press Alt-F4 or click on the  button at the top right of the window.

You can re-open the project by starting PRESENCE again and selecting File > Open Project.



Opening a second project with 'File > Open project' when one is already open results in a total muddle.

To open a second project when one is already open, start PRESENCE again (eg. from the Start > Programs menu or from a desktop icon) so that you have a separate PRESENCE main window. Use File > Open Project to open the second project here. The bar at the top of the window indicates which project file is open in each window.

## G. Simulating data for study design

PRESENCE allows you to investigate the likely results of different study designs and to compare them with the 'true' values underlying the simulated data.

Open PRESENCE and select 'Run > Simulations' from the pull-down menus. Just to see how it works, run a simulation with the default values: click on "Simulate".

After a few moments, a Notepad window opens with the simulation results. The first section of the file recaps the elements of the study design:

```
Total number of sites sampled:200
Number of sites sampled more intensively:100
Number of visits to intensively sampled sites:9
Number of visits to other sites:5
```

This is a pretty intensive study, involving  $900 + 500 = 1,400$  surveys.

Next is a report on the simulation carried out:

```
Number of simulations:100
Number of times species was not detected at any site:0
Number of times convergence not achieved:0
```

You would clearly be in trouble if your study failed to detect any animals at all, or if there was a high chance of not getting enough data for PRESENCE to be able to calculate the parameters (which is what failure to converge means in practice).

Then come the real detection probability (0.2) and occupancy rate (0.7) used to generate the simulated data, and finally the results of the analysis. Since the process is based on random numbers, your results will be somewhat different from the following:

```
Average naive estimate from single visit:0.143300
Average estimate of occupancy probability:0.706045
Simulation based estimate of standard error:0.050459
Average estimate of the standard error:0.052302
```

In this case the average estimate of occupancy is very close to the true value, but that will almost always be the case if you do a high enough number of simulations. Since your study will only have one 'run', you need an indication of how far from the true value you could be. The 'simulation based estimate' of standard error indicates the range of occupancy probabilities obtained during the simulation – approx. 95% of the results fall within  $2 \times \text{SE}$  of the average, in this case  $0.7 \pm 0.1$ . The 'average estimate of the SE' is an indication of the SE you are likely to get when you analyze the data from your single run.

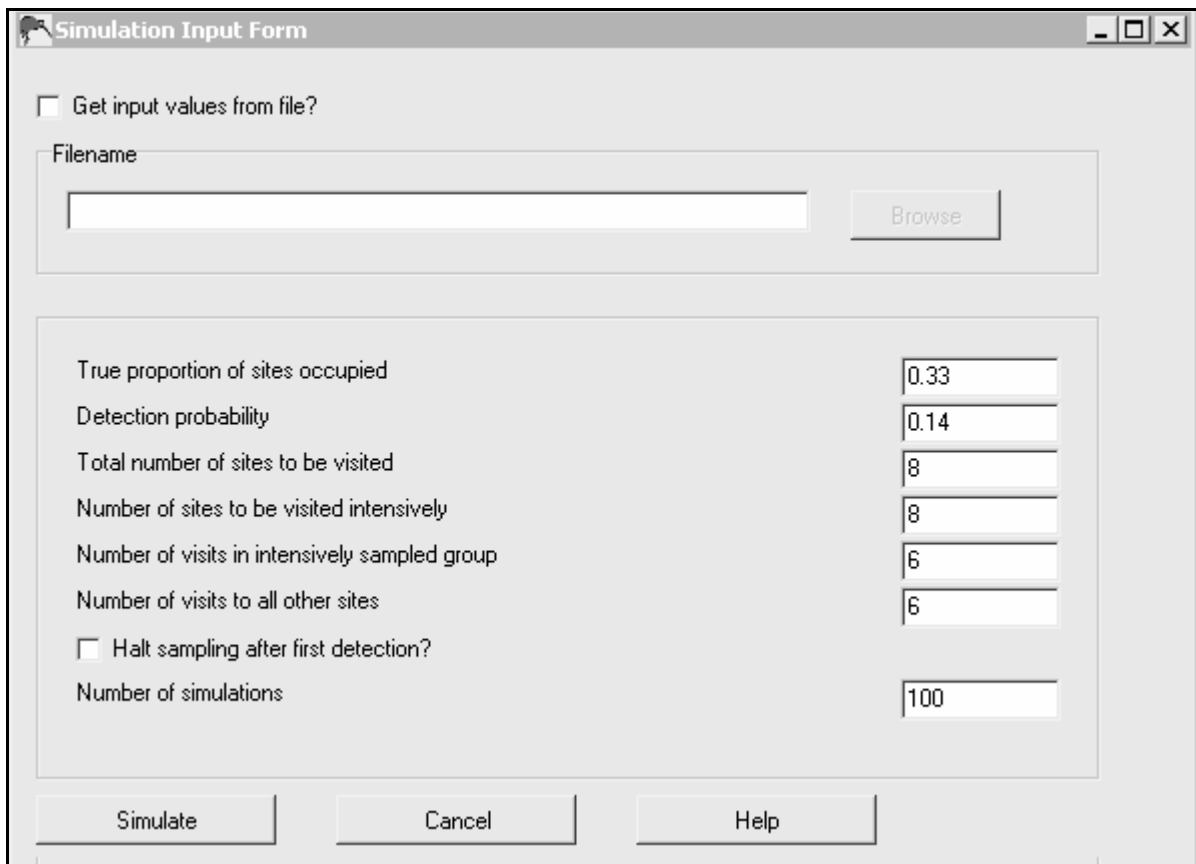
Now let's run a simulation of our own. In particular, let's use simulations to investigate golden cats in plantations.

To begin with we need plausible estimates of the true occupancy and probability of detection, and that may be a major difficulty unless a similar study has been done before or you have other data sets which

you can work on, e.g., going through camera trapping records or line transect surveys to get an idea of detection probability. For golden cats in plantations, let's just assume that the true occupancy and detection rates are about the same as those for forests,  $\Psi = 0.33$  and  $p = 0.14$ .

The camera-trapping study included 8 sites in plantations, and cameras were out for about 30 days = 6 occasions of 5 days each.

Put these values into PRESENCE:



Simulation Input Form

Get input values from file?

Filename

Browse

True proportion of sites occupied: 0.33

Detection probability: 0.14

Total number of sites to be visited: 8

Number of sites to be visited intensively: 8

Number of visits in intensively sampled group: 6

Number of visits to all other sites: 6

Halt sampling after first detection?

Number of simulations: 100

Simulate Cancel Help

The exact results will vary between runs, as they are based on random numbers. First check the line Number of times species was not detected at any site:

How many simulations had no detections in plantations, even though the true occupancy rate was the same as that we found for forests? What does this tell us about the power of this design to estimate occupancy in plantations?

Increase the number of sites to 80 (ie. roughly the same as for the Primary and Logged habitats) and try again.

## Literature cited

**Lynam, A J; R Laidlaw; Wan Shaharuddin Wan Noordin; S Elagupillay; E L Bennett.** 2007. Assessing the conservation status of the tiger *Panthera tigris* at priority sites in Peninsular Malaysia. *Oryx* **41**:454-462.