

PRESENCE User Manual

Darryl I. MacKenzie

November 12, 2012

Program PRESENCE version 4.4 <120515.1509> by James E. Hines

File View Run Tools Help

Program PRESENCE 4.4

Start a new analysis by clicking File/New Project
Open an old analysis by clicking File/Open Project

Recent Modifications:

- Ver 4.4 (15May2012) - fixed multi-season-multi-state general parm. model
- Ver 4.3 (13Apr2012) - Modified spatial-correlation model to allow robust design
- Ver 4.2 (17Feb2012) - added parameters to 2 species multi-season model
- Ver 4.1 (7Feb2012) - fixed output for multi-season model w/ combined DM
- Ver 4.1 (31Jan2012) - fixed conf. interval in Royle/Nichols output
- Ver 4.0 (6Oct2011) - Added models: multi-season-het, multi-season-2species
- Ver 3.1 (14Oct2010) - Added model: single-season-false-positive-detections
- Ver 3.0 (27Jan2010) - changed results file structure...
Results AIC table and model output files are now stored in a folder, instead of a single pa2 file. New results AIC table is saved as pa3 file. PRESENCE will automatically convert old results pa2 file to new pa3 file and save it in the same folder.
e 'Recent changes' in Help menu for more information

** New: 'Recent changes' in Help menu **

Proteus
wildlife research consultants

USGS
science for a changing world

Program PRESENCE <<<< New: Model averaging revised >>>>

Preface

This User Manual has been developed as a learning tool for PRESENCE. It is intended to be complimentary to the information already contained within the PRESENCE help files and to the primary literature. It should not be regarded as a replacement for any of these other sources.

It is a dynamic document that is sure to evolve over time. Currently (November 2012) it only covers the basic single- and multi-season models, although the intent is that more will be added as time and resources allow. Feedback on the User Manual would be much appreciated, particularly on errors or mistakes in the document. Requests for added features or sections will be considered, but may not be actioned if they will take a substantial time to develop. Anyone that is interested in contributing to the manual should contact me.

Acknowledgements

Getting the manuscript to this point would not be possible without the assistance of the following groups and people;

- Jim Hines, USGS Patuxent Wildlife Research Center
- Rob Sutter, Enduring Conservation Outcomes
- Clark County Desert Conservation Program, Las Vegas, Nevada

Conditions of use

This document may be freely distributed in its entirety provided it is not for commercial gain. It may not be partly reproduced or distributed by any means without the written permission of the author.

Darryl MacKenzie
Proteus Wildlife Research Consultants
<http://www.proteus.co.nz>

Foreword

Part of the impetus and support for this Presence User Manual came from the Clark County, Desert Conservation Program (the state of Nevada, US), which implements the counties' Multiple Species Habitat Conservation Plan. One of the more significant species covered by the plan is the Mojave desert tortoise (*Gopherus agassizii*), an US Federally-listed Threatened species. While inhabiting what most people think is a pretty inhospitable and desolate desert environment in California, Nevada, Arizona and Utah, the species' populations have been declining. The reasons for the decline range from habitat loss and degradation from urbanization (Las Vegas, Nevada, St. George, Utah), energy development (solar facilities), grazing, invasive species, and fire to mortality by vehicles, disease, and predation. A primary need for any conservation effort for a declining species is to assess the status and trends of the species.

The US Fish and Wildlife Service has developed a robust sampling method using line-distance sampling to estimate abundance and density throughout the species' range. However, the lack of precision in this methodology over short timeframes and the inability to use existing data to make management and conservation decisions have led to suggestions for different monitoring techniques. In addition, the line-distance sampling does not lend itself to collecting environmental covariates that would allow a better understanding of finer-scale habitat preferences for the species. It is in this context that the Desert Conservation Program has initiated a project to use occupancy sampling to assess the proportion of habitat within an area that contains evidence of the Mojave desert tortoise and to collect environmental covariates. In the process of developing the occupancy sampling monitoring protocol, Darryl assisted in the sampling design and analysis sections. Needing a greater understanding of how to analyze the data, we contracted with Darryl to develop this manual.

We thank Darryl for all the support that he has given our project to test occupancy sampling for the Mojave desert tortoise. We look forward to implementing the project (Spring 2013) and using the single season models after the first year's data are collected, and the multiple season models over the 3 to 5 year timeframe of testing the methodology. We know that this manual will be useful to many researchers and conservationists.

Robert D. Sutter
Conservation Ecologist
Enduring Conservation Outcomes
Science Advisor for the Clark County, Nevada, US, Desert Conservation Program
November 2012

Glossary

.out file extension for PRESENCE output file with results of a single model that has been fit to the data.

.pa3 file extension for project results file that stores a summary of the models that have been fit to the data.

.pao file extension for PRESENCE input data file.

AIC Akaike's Information Criterion.

ASCII American Standard Code for Information Interchange, i.e., a standard character-encoding scheme for storage of text characters to computer files.

detection history a sequence of 1's and 0's indicating whether the species was detected or not (respectively) in each of the surveys of a sampling unit.

epsilon probability of species going locally extinct from a sample unit between two seasons.

eps abbreviation for 'epsilon'.

ε see 'epsilon'.

gamma probability of species colonizing a sample unit between two seasons.

gam abbreviation for 'gamma'.

γ see 'gamma'.

GUI Graphical User Interface, i.e., a point and click window on the computer.

h see 'detection history'.

lambda growth rate or rate of change in occupancy.

lam abbreviation for 'lambda'.

λ see 'lambda'.

MLE Maximum Likelihood Estimate.

OR odds ratio.

p probability of detecting the species in a survey given the species is present at the sampling unit.

psi probability the species is present at a sampling unit.

ψ see 'psi'.

R a free statistical computing package, that does much more than statistics <http://www.r-project.org>.

sampling unit basic landscape unit at which the presence or absence of the species is being determined. Could be naturally occurring or arbitrarily defined (e.g., pond, habitat patch, grid cell, etc.).

season the applicable timeperiod for which statements about the presence or absence of a species is biologically meaningful. It may not necessarily corresponding to a biological season (e.g., breeding season) or climatic season.

survey a single opportunity for the detection of the target species. Multiple opportunities for detection (i.e., multiple surveys) may exist in a single 'visit' to a sampling unit in some situations.

Chapter 1

Introduction

PRESENCE is Windows-based software that has been primarily developed to fit occupancy models to detection/nondetection data. The basic sampling situation envisioned for most of these models is that within a region, appropriately defined sampling units are surveyed for a target species to establish its presence/absence within one, or multiple, sampling seasons. However, due to imperfect detection, the species will not always be detected when present leading to false absences which, unaccounted for, could lead to misleading conclusions about the occurrence or distribution of the species. To address the detection issue, within each season repeat surveys of the sampling unit are conducted. The surveys may be temporally or spatially replicated, and may occur within a single or multiple visit to each sampling unit. The detection/nondetection of the species in each of the multiple surveys is recorded providing the necessary information to reliably separate out false and true absences. For more details on the basic sampling requirements see MacKenzie et al. (2006).

It has been purposely designed to resemble Program MARK to lessen the learning curve for users familiar with this popular mark-recapture software. A user can fit multiple models to their data and PRESENCE stores the results for each model and presents a summary of how well the models rank according to a model selection metric (Akaike's Information Criterion, AIC, is used as the default). Models are fit using maximum likelihood techniques MacKenzie et al. (2006), hence parameter estimates are known as maximum likelihood estimates.

PRESENCE consists of 2 main windows; 1) a typical Windows graphical user interface (GUI); and 2) a black-box command window (reminiscent of a DOS-window) that performs the number crunching. Typically most users, especially beginners, will conduct analyses using the GUI to set up data files, specify and fit models and examine the results. More advanced users, however, may choose to use PRESENCE via the command line and call it directly, or from software packages such as R to enable further manipulation of results. This User Manual focuses on using PRESENCE via the GUI.

The results of an analysis are stored as a project. Each project consists of series of files, all stored within a project folder; a project file (*.pa3); a data file (*.pao); and an output file for each model fit to the data (*.out). The project file stores information relevant to the analysis such as the data file name, number of models fit to the data and a summary table of results. It is just an ASCII text file so can be opened with a text editor for inspection. The data file is a tab-

Table 1.1: Single-season models available in PRESENCE as of June 2012

Model type	Description	Key References
Standard	Fit standard models with and without covariates. Also allow for dependent surveys	MacKenzie et al. (2002); Hines et al. (2010)
Multi-method	Repeat surveys and multiple detection methods used on each survey	Nichols et al. (2008)
False positive detections	Accounts for potential misidentification of species	Miller et al. (2011)
Multi-state	Extension to allow for 2 or more occupied states	Nichols et al. (2007)
Two-species	Examine patterns of co-occurrence for 2 species while accounting for imperfect detection	MacKenzie et al. (2004)
Heterogeneity (Royle/Nichols)	Account for 'abundance' induced heterogeneity in detection	Royle and Nichols (2003)
Staggered entry	Relaxes the closure assumption of the standard single season model allow	Kendall et al. (in press)
Repeated Count Data (Royle Biometrics)	Repeated counts of unique individuals rather than detection/nondetection	Royle (2004)

delimited text file with a specified format (described in the PRESENCE help files). These can be easily created external to PRESENCE, and then imported when a new project is begun, or inputted directly into PRESENCE using a spreadsheet-like interface. Each output file contains the results for a specific model that has been fit to the data, including which covariates had been used and the resulting parameter estimates.

PRESENCE has the capability of fitting a wide range of occupancy models; Tables 1.1 and 1.2 contain lists of the available models as of June 2012. Different model types enable different aspects of the biology of the system to be investigated, with models becoming more complex further down the list in each table. However, there is a great deal of similarity in the mechanics of how models are specified in terms of which covariates are include in a particular model that is fit to the data.

The basic steps required for an analysis within PRESENCE are:

1. Identify the general class of analysis to use to address the questions of interest (e.g., single- vs multi-season analysis, standard or accounting for false positive detections,

Table 1.2: Multi-season models available in PRESENCE as of June 2012

Model type	Description	Key References
Standard	Standard multi-season model enabling estimation of occupancy, colonization, extinction and persistence	MacKenzie et al. (2003)
False positive detections	Account for potential misidentification of species	Unpublished
Heterogeneous detections	Allow heterogeneity in detection with finite mixtures	Unpublished
Multi-state	Extensions to allow for 2 or more occupied states, and changes between them	MacKenzie et al. (2009)
Integrated occupancy	habitat- Simultaneous modeling of changes in occurrence and discrete habitat types	MacKenzie et al. (2011)
Two-species	Examine what effect one species may have on the dynamic processes of another (i.e., competition)	Bailey et al. (2009)

etc.).

2. Develop a list of models that will be fit to the data (known as a candidate set). Each model may have different combinations of covariates on each parameter type (e.g., occupancy and detection) and should represent different questions of interest for the species.
3. Prepare the detection data and any necessary covariates in software outside of PRESENCE (e.g., in a spreadsheet or database software).
4. In PRESENCE, start a new project and enter the data.
5. Define a model to be fit to the data, identifying which covariates are to be included.
6. Interpret the output.
7. Fit all models in the candidate set.
8. Draw overall conclusions.

This User Manual will not cover all of the different types of models available within PRESENCE, and will concentrate on the standard single-season and multi-season models. It is also beyond the scope of the manual to provide a lot of detail on topics like comparing models, theory on model selection and model selection strategies; some brief explanation may be given but the expectation is that users will use the references provided for additional reading. The main focus of the User Manual is to lead users through the mechanics of using PRESENCE, setting up models and interpreting the PRESENCE output. It is presumed that you have already identified what type of model (i.e., general class of analysis) you should be using and the specific models in your candidate set.

Chapter 2

First Steps

2.1 Before going to the field

There are many points associated with the analysis that need to be considered while designing the data collection protocols. Issues such as defining a sampling unit and what constitutes a repeat survey are very relevant to the correct interpretation of the results, but these fundamental issues are beyond the scope of this User Manual and have been addressed elsewhere (MacKenzie and Royle, 2005; MacKenzie et al., 2006). More relevant to the analysis is consideration of the potential covariates or predictor variables that might be included in the models; that information will need to be collected while in the field if it is not readily available by some alternative means (e.g., from remote sensing). What also needs to be considered is how the results are ultimately going to be used.

If the models are going to be used to make predictions, or create maps, of species occurrence at places that will not be surveyed, then the types of covariates that could be used for that predictive modeling will be limited to those that can be collected without visiting the other locations. Hence collecting detailed information on covariates that can not be used for that predictive modeling could be wasted effort better utilized elsewhere. However, if the intent is to gain a better understanding of what covariates might be important factors for species occurrence (regardless of whether predictions to other places will be made or not), then information on those covariates should obviously be collected.

Another relevant point with respect to covariates, and how they are used in PRESENCE, is that you can only have missing covariate values if the corresponding observation is also missing, otherwise that data will need to be sub-setted or that covariate can not be used. Therefore, all practical steps should be taken to avoid missing covariate values. Further discussion on covariates is given in Section 2.3.2.

In short, prior to collecting the data, careful thought should be given to the expected method of analysis with due consideration of its data requirements and limitations.

2.2 Downloading and installing PRESENCE

PRESENCE is Windows-based free-ware developed by Jim Hines of the US Geological Survey and a zip file can be downloaded from <http://www.mbr-pwrc.usgs.gov/software/presence.html>. The zip file contains a self-extracting executable file that once opened will begin the typical Windows installation process. In order for the PRESENCE file type of be successfully registered with the Windows operating system administrator privileges are required. However, PRESENCE can also be installed by users without administrator privileges provided it is installed to a directory where the user has write privileges (e.g., My Documents). The disadvantage is that the PRESENCE file types will not be registered with the Windows registry, but PRESENCE should still run.

Linux and Mac users have successfully used PRESENCE via Windows emulators such as Wine. Alternatively, if you have access to a Windows installation disk, a virtual machine could be created (e.g., with VirtualBox) and the Windows operating system installed so PRESENCE can run under a virtual Windows environment.

2.3 Preparing your data

There are two aspects of the data used by PRESENCE; 1) the detection/nondetection data; and 2) covariates or predictor variables. This information is stored in a PRESENCE data file (*.pao), which is just an ASCII text file with a specific format. You can create this file yourself directly either through code or with a text editor, or PRESENCE has a built in spreadsheet-like interface that enables you to copy and paste your data in from spreadsheet software and then save the data in the required format. Basically, PRESENCE requires data in a row and column format where each row of data represents a single sample unit, and each column represents a survey occasion. Sample data sets are installed with PRESENCE as spreadsheet files.

2.3.1 Detection/nondetection data

The detection/nondetection data is, for most models, a sequence of 1's and 0's indicating whether the species was detected or not in a particular survey of a particular sampling unit. For example, the history:

0101

indicates the species was detected in the second and fourth survey of the unit, but not detected in surveys one and three. We refer to this sequence as a *detection history*. A detection history may also include missing values which can be represented as either a dash (-) or a dot (.), e.g., 01-1. These missing values represent survey occasions when no survey of the unit was actually conducted, which may be used to align surveys across all units into a meaningful chronological order or to denote occasions when a survey was planned, but was not completed for some reason. Missing values may also be used to 'pad out' a detection history if an unequal number of surveys was conducted at different units as PRESENCE expects all histories to be the same length, e.g.,

0101

10--

000-

01--

For most of the basic models, it is therefore required that the outcome of each survey can be determined in terms of either a detection (1), nondetection (0) or a missing value (- or .). For more complicated models, the detection history may include other values (e.g., 2, 3, 4, ...) though the interpretation of the values depends on the model being used. If the data is being setup in a spreadsheet, then the outcome of each survey should be placed in a separate cell.

For multi-season data it is important that the first survey of each unit each season are in the same column of data. Missing values can be inserted to pad out the data if need be.

2.3.2 Covariates

Continuous and categorical covariates

There are two general types of covariates that could be included in PRESENCE, continuous or categorical variables. Continuous variables could potentially take any value between $\pm\infty$, e.g., elevation, rainfall, latitude, temperature etc. While not strictly necessary, often it is highly advisable to standardize continuous covariates prior to entering them into PRESENCE to improve the performance of the software; this is particularly true for covariates that are a relatively long way from zero, or include large positive or negative values. A general formula for standardizing continuous covariates is

$$x_i^* = \frac{x_i - a}{b} \quad (2.1)$$

where x_i is an observed covariate value, and a and b are constants. If a was the average of the covariate values and b the standard deviation, then Equation 2.1 would be known as the z-transformation, which is a completely legitimate approach. However, in practice it may be preferable to use alternative values for a and b to ensure interpretable results (discussed further below). Typically it is a good idea to use values that result in zero on the standardized scale being somewhere near the middle of the range of standardized covariate values (i.e., select a to be close to the mean or median of the unstandardized values), and the standardized values are between ± 5 (or thereabouts). It is recommended that b should be chosen so that a one unit change of the standardized covariate is on a convenient scale. For example, for an elevation covariate initially measured in meters, b might be chosen to be 100 so the standardized elevation covariate is on the scale of per 100m.

Categorical variables can only take a limited number of discrete values, where the values do not necessarily relate to any particular ordering. In PRESENCE these need to be represented as a series of indicator (or dummy) variables, which are binary, 0-1 variables used to represent the different levels of a categorical covariate. For example, if you have three habitat categories A, B, and C, you could have three indicator variables; one for each type (Table 2.1).

Table 2.1: An example of defining indicator variables

Habitat Type	HabA	HabB	HabC
A	1	0	0
B	0	1	0
C	0	0	1

Many statistics packages will allow you enter the categorical covariate directly, and perform the conversion to a series of indicator variables automatically (technically these indicator variables are defined by something called a *contrast*). PRESENCE does not currently include that functionality therefore it must be done manually before entering the data into PRESENCE. This can be easily accomplished within a spreadsheet using the `if` function. Table 2.2 provides an example of a portion of a data set illustrate what the indicator variables may look like in practice.

Table 2.2: An example of using indicator variables in practice

Unit	Habitat Type	HabA	HabB	HabC
1	A	1	0	0
2	C	0	0	1
3	B	0	1	0
4	B	0	1	0
5	C	0	0	1
6	A	1	0	0
7	A	1	0	0
8	B	0	1	0
9	C	0	0	1
10	C	0	0	1

If there is a natural order to the covariate categories, rather than creating a series of indicator variables, an ordinal categorical covariate can be created by using appropriate numeric values. For example, if Habitats A, B and C represent low, medium and high quality habitat, the values A, B and C in Table 2.2 could be replaced with the values of 1, 2 and 3. One consideration of creating an ordinal categorical covariate is that the relative difference in the numeric values reflects that relatively difference in the ordering of the categories. So by using the values of 1, 2 and 3 for low, medium and high quality habitats, it is presumed that the difference between low and medium quality habitats is similar to the difference between medium and high quality habitats as the ordinal covariate value is different by 1 in each case.

Even though a distinction has been made here between continuous and indicator variables, mathematically, both are treated in the same way when fitting models; they are both simply a number and interpretation of effect sizes is very similar.

Transformations of covariates

A separate issue from the standardization of continuous covariates, is the need to consider the general form of the relationship between the response variable (e.g., occupancy or detection) and the continuous covariate or predictor variable. Just using the standardized covariate presumes a linear relationship. That is, as the value of the covariate increases, occupancy or detection continues to increase (or decrease) at a constant rate (Figure 2.1). If some other functional relationship is expected (e.g., Figure 2.2), then a transformation of the covariate values would have to be performed and the transformed covariate values be included in the PRESENCE data file. An example of this is given in the third example in the chapter on single-season models.

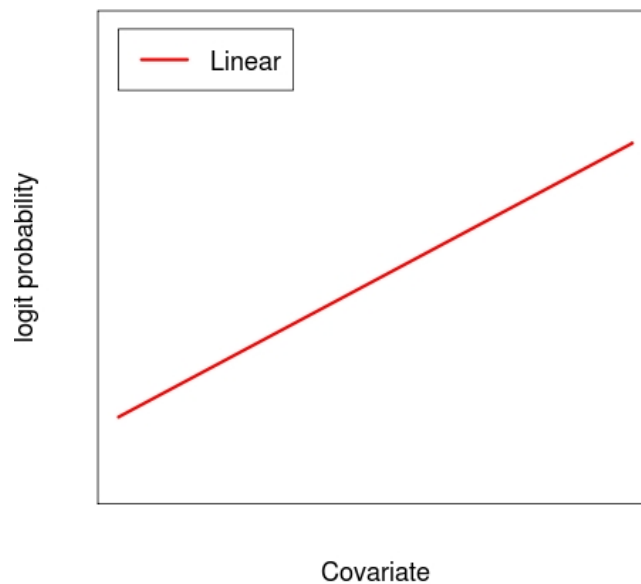


Figure 2.1: Example of linear relationship between a probability on the logit scale and a covariate

Site-specific and sampling-occasion covariates

Regardless of whether a particular covariate is a continuous or indicator variable, a distinction is made within PRESENCE between site-specific and sampling-occasion covariates. A site-specific covariate is typically some characteristic of a sampling unit (also referred to as a site) that may be different for different units, and whose value does not change during a season, but may change between seasons. Habitat type, elevation and distance from nearest water source may all be examples of site-specific covariates. The value of a sampling-occasion covariate (sometimes called a survey-specific covariate) could be different for every survey of any unit, e.g., air temperature, time of day, observer and cloud cover.

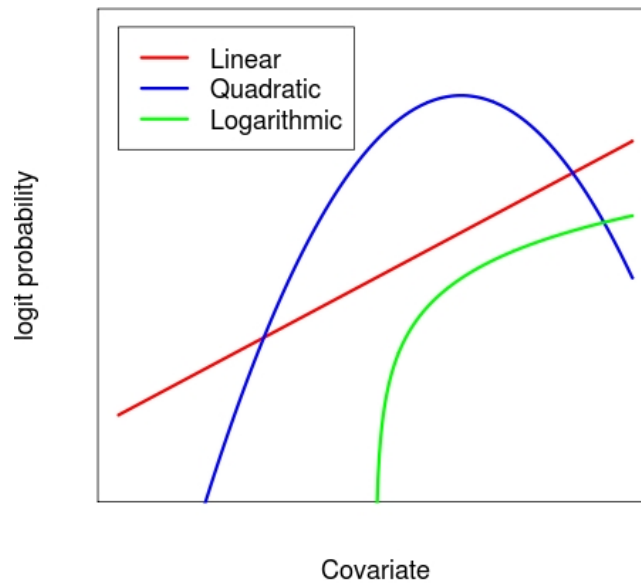


Figure 2.2: Example of linear, quadratic and logarithmic relationships between a probability on the logit scale and a covariate

The distinction is important for two reasons. Firstly, site-specific covariates can be used for both occupancy and detection probabilities whereas sampling-occasion covariates can only be used for detection probabilities. As a unit is presumed to be either occupied or not for all surveys within a season (MacKenzie et al., 2006), it does not make sense to attempt analyzes where the value observed during a particular survey affects the probability of an event that has already occurred. It may however be reasonable to summarize a sampling-occasion covariate and use the summary as a site-specific covariate. For example, average air temperature may provide an overall indication as to whether the unit typically experiences warmer or cooler temperatures, which may influence occupancy probabilities. Secondly, the different types of covariates have a differing number of observations. Site-specific covariates have one value for each sampling unit and get entered into PRESENCE as a single column, while sampling-occasion covariates have one value for every survey hence will require multiple columns and will have the same dimensions (i.e., number of rows and columns) as the detection/nondetection data.

Missing covariate values

Missing sampling-occasion covariate values (again, denoted with a '-' or '.') are allowed provided that they correspond to a detection/nondetection survey occasion that is also missing (e.g., a particular unit was not surveyed on day three as planned, hence the water temperature at that unit on that day was not collected either). PRESENCE will provide a warning if a covariate value is missing, but the corresponding detection survey data is non-missing, but

will proceed with an analysis otherwise using a default covariate value of -999 hence results may be incorrect. Site-specific covariates can not be missing. In order to relate an observed outcome to a measured covariate, you need to have both the outcome and the covariate value. This is a basic requirement of most statistical analyses. If there are missing covariate values for non-missing data then it may be necessary to subset your data and conduct different sets of analyses on the different subsets (with some common models being fit in both to enable some comparison of results).

Automatic variable selection

PRESENCE is not designed to enable some form of automatic selection procedure to identify what variables appear to be 'important'. While such an approach is conceptually possible, there are no plans to attempt to implement such a procedure in PRESENCE. Our philosophy is that users should be able to justify why a particular set of covariates has been included in a model that has been fit to the data by using sound scientific judgment around the objective of why the data was collected in the first place (MacKenzie et al., 2006). Furthermore, by fitting a large number of models to the data (as in a automated procedure), there is a strong potential for finding a spurious combination of covariates that happen to explain the observed data very well, but have little biological meaning. Such a model would provide a very good description of the data, but have limited usefulness for prediction to other data sets. As noted earlier, topics such as variable selection and model comparison are detailed subjects that are beyond the scope of this manual.

Data preparation check-list

- Detection data is formatted with 1 row per unit and 1 column per survey and only contains whole number values or the '-' or '.' characters
- Categorical covariates have been converted to a series of indicator variables, or to ordinal categorical covariates
- Continuous covariates have been standardized to an appropriate scale, particularly if zero is well outside the unstandardized range
- Site-specific covariates are each formatted into a separate column, with each row for a different unit
- Sampling-occasion covariates are in the same format as the detection data, and any missing values correspond to missing surveys
- The same row ordering is used for the detection data and all covariates

2.4 Creating a PRESENCE data file

As noted previously, a PRESENCE data file is simply a tab-delimited ASCII text file with a certain format. It can either be created directly by the user in a text editor or with computer code, or from within PRESENCE by copying and pasting data from a spreadsheet. Rather than detail how to do so here in abstract, the necessary steps for creating the data files are given in the examples in the following chapters.

Chapter 3

Single-season model

A single-season model can be used to look at the level or patterns in occupancy at a single point in time. It is essentially a snapshot of the presence and absence of a species for a given time period. The single season model could be used to produce numerical summaries of the situation (i.e., estimated proportion of units occupied or quantifying the effect of certain covariates) or to produce maps indicating areas with higher and lower probabilities of occurrence (e.g., species distribution maps). In this chapter we shall briefly review the underlying statistical methods of the modeling approach, then work through three examples of increasing complexity. The examples used here focus on the standard single-season model of which there are a number of extensions now available in PRESENCE, although these extensions are beyond the scope of this User Manual. However, the basic mechanics of how to setup up models in PRESENCE to fit to data is very similar for all model-types; the main differences is in terms of the range of parameters associated with each model-type.

3.1 Underlying model

The statistical model used by PRESENCE is that developed by MacKenzie et al. (2002), and estimates obtained by using the principle of maximum likelihood. Similar analytic approaches have been developed by others (Tyre et al., 2003; Wintle et al., 2004; Stauffer et al., 2004) although the framework developed by MacKenzie et al. (2002) was the most general. Users are directed to MacKenzie et al. (2002) or MacKenzie et al. (2006) for greater detail. It should be noted that these models can also be used within a Bayesian approach to statistical inference (rather than with maximum likelihood), but it is not currently possible to do so with PRESENCE hence a Bayesian approach is not discussed here. Although it is stressed that the actual underlying model is, conceptually, exactly the same regardless of whether maximum likelihood or Bayesian methods are used; the difference is simply how that model is applied in association with the observed data.

The single-season model has two fundamental processes, occupancy and detection. Sample units within the region of interest are either occupied by the target species or not (i.e., species is present or absent at each unit) and the probability of the species being present at the i th unit is denoted as ψ_i (spelt 'psi'). Given the unit is occupied, the probability of detect-

ing the species in the j th survey of that unit is p_{ij} . If the unit is unoccupied then, using the standard single-season models, the species cannot be detected. In order to reliably separate out occupancy from detection (i.e., where the species is vs where the species is found) repeat surveys within the season are required. During the season it is assumed that units are closed to systematic changes in occupancy, the outcome of each survey of a unit is independent and there is no misidentification of species (i.e., no false detections). Because the species is detected imperfectly there is the potential for false absences in the data (i.e., units where the species was never detected, but it was actually present) which will lead to occupancy being underestimated if unaccounted for. The intent of this modeling is to explicitly correct for detection issues leading to improved inferences about occupancy and the factors that may be influencing it.

The repeated surveys of a unit will yield a *detection history* denoting the sequence of detections and nondetections of the species at that unit. From the detection history, a verbal description of the data can be developed and then translated into a probability statement, which is an expression for determining the probability of observing that particular detection history given the model. For example, consider the detection history (h_i):

$$h_i = 0101.$$

A verbal description would be:

the unit was occupied, was not detected in the first survey, detected in the second, not detected in the third and detected in the fourth survey.

Translating to a probability statement is simply achieved by replacing the relevant phrases associated with certain events in the verbal description by the probability of the event occurring. For example, the unit is occupied with probability ψ_i and the species is not detected in the j th survey with probability $1 - p_{ij}$. Therefore, the probability statement for this detection history is:

$$\Pr(h_i = 0101) = \psi_i(1 - p_{i,1})p_{i,2}(1 - p_{i,3})p_{i,4}. \quad (3.1)$$

Probability statements for units where the species was detected at least once are constructed in a similar manner.

For units where the species was never detected, e.g., $h_i = 0000$, the same approach is used while recognizing that the species may go undetected at a unit for two reasons; due to either a true or a false absence, and that both possibilities must be accounted for in the verbal description and probability statement. The verbal description for the history 0000 is:

*the unit was occupied and the species was not detected in all four surveys (i.e., a false absence) **OR** the unit was unoccupied (i.e., a true absence).*

To account for the multiple options in the verbal description (that cannot be differentiated between from the available data) within the probability statement, the probability of each option is determined and then added together. That is:

$$\Pr(h_i = 0000) = \psi_i(1 - p_{i,1})(1 - p_{i,2})(1 - p_{i,3})(1 - p_{i,4}) + (1 - \psi_i). \quad (3.2)$$

Given the set of detection histories from the s units that were surveyed, the model likelihood is defined as,

$$L = \prod_{i=1}^s \Pr(h_i). \quad (3.3)$$

Once derived, the likelihood equation is used by substituting in numeric values for the ψ and p parameters and finding what combination of values maximizes the value of the likelihood expression. The parameter values that maximize the likelihood are known as *maximum likelihood estimates* (MLE's).

As a PRESENCE user, it is good to have a basic understanding of what the software is doing with the data you input to obtain the resulting output and MLE's. However, from a practical perspective, one does not have to get overly concerned about the fact that PRESENCE is creating these probability statements which contain combinations of ψ 's and p 's. The focus of the user should be how to use the software to focus on questions associated with the separate components of occupancy and detection, although being mindful that the two are inextricably linked.

Further modeling of the ψ 's and p 's (e.g., to investigate what factors are important covariates for occupancy and detection) is facilitated by using something called the logit-link function, which is a non-linear transformation used to rescale probabilities from the 0-1 scale to the $\pm\infty$ scale. With the logit-link, occupancy and detection probabilities can be expressed as a function of site-specific and sampling-occasion covariates, e.g.,:

$$\text{logit}(\psi_i) = \ln\left(\frac{\psi_i}{1 - \psi_i}\right) = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i}, \quad (3.4)$$

$$\text{logit}(p_{ij}) = \ln\left(\frac{p_{i,j}}{1 - p_{i,j}}\right) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 y_{1,ij} + \beta_4 y_{2,ij}, \quad (3.5)$$

where \ln is the natural logarithm, the x 's and y 's are site-specific and sampling-occasion covariates respectively, and the α and β parameters are the regression coefficients to be estimated. These resulting equations are essentially logistic regression equations. The terms of the form:

$$\ln\left(\frac{\theta_i}{1 - \theta_i}\right) \quad (3.6)$$

are the actual transformations applied to the probabilities (denoted with θ here for generality), and is just the log of the ratio of the probability of success to the probability of failure. This ratio is also known as the *odds* hence the logit-link is also referred to as the log-odds link. More detail on the use of the logit-link and its interpretation is given in the examples in this chapter and in MacKenzie et al. (2006).

The key to successfully using PRESENCE is realizing that one is simply setting up a number of logistic regression equations for the different parameters of interest. The regression coefficients for each parameter type are estimated simultaneously through the framework of the probability statements which involve a combination of the ψ 's and p 's, hence are automatically corrected for the effect of the other parameters. How the user specifies the logistic equations to be applied is covered in further detail below.

3.2 Example 1: Blue-ridge two-lined salamander

This first example uses one of the sample data sets that is installed along with PRESENCE. It is data that was collected on the blue-ridge two-lined salamander (*Eurycea wilderae*) in Great Smoky Mountain National Park, USA, where 39 transects were each surveyed 5 times between April and mid-June. Further details are given on page 99 of MacKenzie et al. (2006). In this example we shall go through the process of setting up a new project, inputting the data from a spreadsheet and fitting two relatively simple models to the data. There are no covariates in this example.

3.2.1 Creating a new project and entering the data

Once you have started PRESENCE, in the menu bar, select **File>New Project**. This will bring up a dialog window entitled **Enter Specifications for PRESENCE Analysis**, and you should select the large **Input Data Form** button in the bottom right corner (Figure 3.2). This, in turn, will open a data input interface window that allows you to copy and paste your data into PRESENCE from a spreadsheet package.

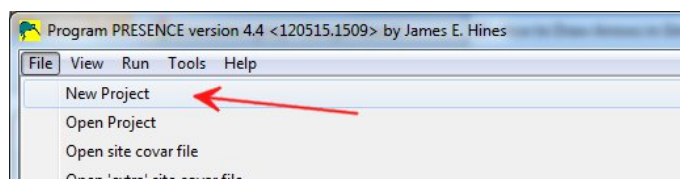


Figure 3.1: Starting a new project

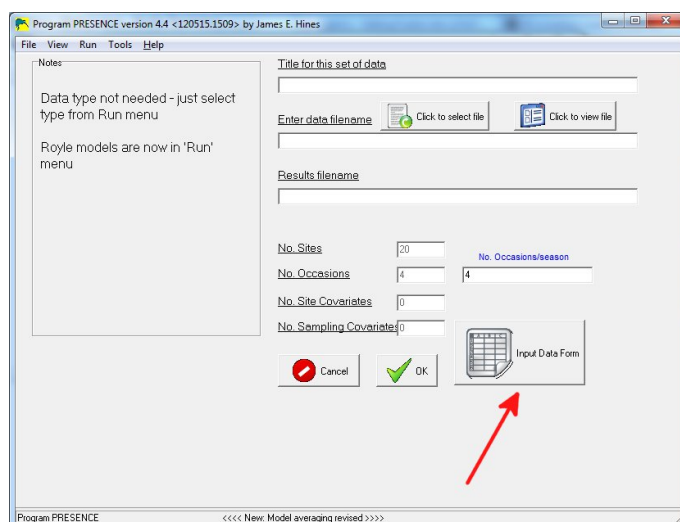
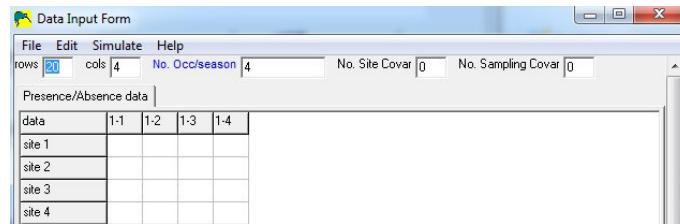


Figure 3.2: Location of **Input Data Form** button

The **Data Input Form** has a spreadsheet-like format. The detection/nondetection data or covariate values are entered into the respective grid cells, with the five text boxes near the

top of the window being used to define the various dimensions of the spreadsheet; the number of rows (sampling units), number of columns (cols; repeated surveys), number of survey occasions per season, number of site-specific covariates and number of sampling-occasion covariates. For this example there are no covariates, so last 2 boxes can left as 0, and PRESENCE will automatically adjust the number of rows and columns when we paste in the data. For single-season models, PRESENCE will ignore the value specified for the number of occasions per season, although it will default to the total number of surveys. This value is required however for other models such as the multi-season models, but it shall be ignored for now.

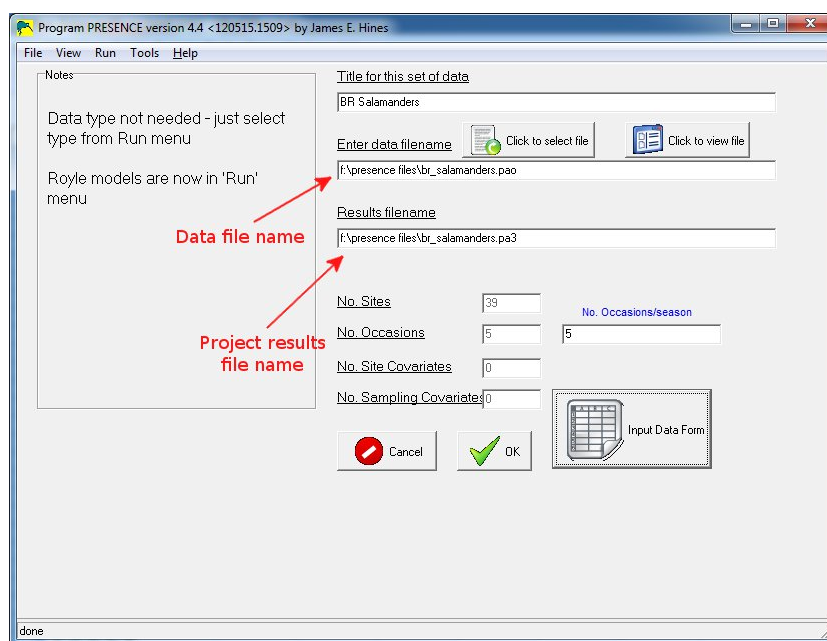


The data for this example is in a file called **Blue_Ridge_pg99.xls** which in the **sample_data** folder that will be created within the PRESENCE installation folder (e.g., `c:\program files\PRESENCE\sample_data`).

1. Open **Blue_Ridge_pg99.xls** using a spreadsheet package. The data should appear as a series of 1's and 0's arranged in 39 rows (1 for each transect) and 5 columns (1 for each survey).
2. Select the entire data range (i.e., from cell A1-E39) and copy to the clipboard, then return to the **Input Data Form** (note you may need to locate the window in your task bar).
3. Select `Edit>Paste>Paste Values` from the `Edit` menu to paste in values into the grid cells.
4. The number of rows and number of columns should automatically update to 39 and 5 respectively.
5. If the data is pasted in the incorrect position for some reason, select `Edit>Paste>Undo Paste`, ensure the top leftmost cell is selected and attempt to `Paste Values` once again.
6. Select `File>Save As` to save the data, specifying an appropriate location and file name. During the saving process you will be asked:
 - whether you want to use the last column of the detection data as frequency data (number of units at which that history was observed). Select **No**.
 - specify a title for the data file which is just to provide a simple description of the project for your own reference.

7. Once the data is saved, close the **Data Input Form** which will return the user to the **Enter Specifications** window, with the data file and project name being automatically updated.

At this point the data file has been saved, but the set up of the project has not been completed hence no attempt should be made to run a model yet. However, before proceeding some explanation should be given. The middle text box should contain the full pathway and file-name for the data file that has just been saved (I called it **br_salamanders.pao**). The third text box contains the full pathway and project filename, which PRESENCE specifies automatically¹. The naming convention for the project file is to use the same base filename as the data file, but to change the extension to .pa3. This is important as if you want to reanalyze the same data set, but retain the results of a previous analysis, you first want to create a copy of the data file otherwise PRESENCE will attempt to overwrite the previous project file (although you will be prompted as to whether you want to overwrite the project first). The project title in the first text box is simply for your own reference and can be anything. The raw .pao file can be viewed by clicking on the **Click to view file** button. Once the data file has been selected (either by entering and saving the data through the spreadsheet interface, or by browsing for a data file that has already been created, **Click to select file**), summary information is presented on the number of sites, number of surveys, and number of covariates. These can not be modified here, although, the number of occasions per season can.



In order to create the project file and proceed with the analysis, you must hit **OK**. After doing so, the **Enter Specifications** window should close and be replaced with the **Results**

¹As of November 2012 the project pathway is not correct as all files related to a project are placed within a new folder called `datafilename_project`. For example, here the full pathway and filename for the project should be `f:\presence files\br_salamanders_project\br_salamanders.pa3`

Browser window (Figure 3.3). If the **Results Browser** window does not appear, you have not created the project file and must repeat the above steps. An important point is that when the project is created a copy of the data file is placed within the project folder, hence there are two copies of your data file; one inside and one outside of the project folder. The data file actually used by PRESENCE for an analysis is the copy inside of the project folder, therefore if the data file needs modified (e.g., add in additional covariates) make sure the data file inside of the project folder is replaced.

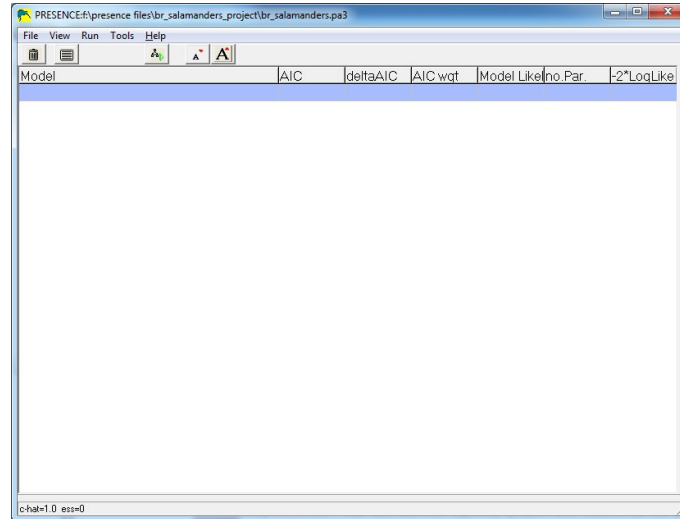


Figure 3.3: Blank Results Browser window which should appear on screen once the project has been successfully set up.

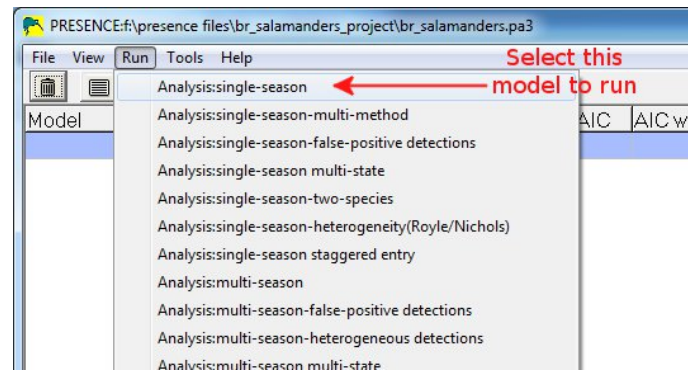
Congratulations! Your data has been inputted, project file created and now you are ready to begin your analysis.

3.2.2 A first analysis

Here we are going to fit two single-season models to the blue-ridge salamander data that we have just set up the project for above. If you happened to have closed PRESENCE after creating the project file you can reopen the project by selecting **File>Open Project** from the main PRESENCE window. Note that recent projects are also listed down the bottom of the **File** menu.

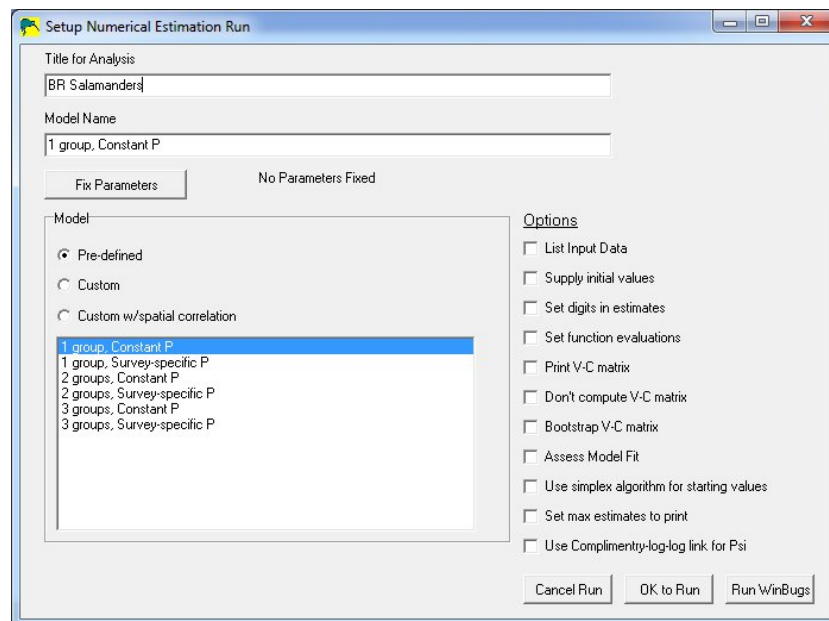
To begin a single-season analysis of this data, in the main PRESENCE window select **Run>Analysis:single-season**. This will bring up the **Setup Numerical Estimation Run (SNER)** window. This same window is used for all analyses, only the **Model** box varies between the different types of analyses. The other parts of the window include the analysis title (which should be carried over from the title given when the project was created), model name, the ability to fix real parameter values, and various options that shall be ignored for now.

Within the model box, there are 3 options; **Pre-defined**, **Custom** or **Custom with spatial correlation**. By default the pre-defined option is selected which provides a list of six pre-



defined models that users can select to learn how to use PRESENCE, but generally will be of limited use for a full-blown data analysis (see the PRESENCE Help documentation for a definition of these models). In this example, we are going to fit the first two pre-defined models to the data, then fit the same two models using the **Custom** model and compare the outputs. The **Custom with spatial correlation** model is an extension of the model described above (Section 3.1) and was developed for a specific situation where the repeat survey information was collected with spatial replicates (segments of a transect) rather than temporal replication (e.g., surveying the same transect more than once, Hines et al. (2010)). The spatial correlation aspects relates to potential lack of independence in the detection of the species in the segments and not spatial correlation in terms of occupancy probabilities. Similar correlation issues can arise with temporal repeat survey hence this particular model type could be useful for a wide range of practical applications, however, the details of this model are not discussed further in this User Manual.

The first pre-defined model **1 group, Constant P** assumes that all transects have the same occupancy and detection probability, and that detection probability is also constant for all



survey occasions. Only two parameters will be estimated for this model; one for occupancy and one for detection. The second model **1 group, Survey-specific p** again assumes that all transects have the same probability of occupancy and detection, but now detection will be estimated separately for each of the five surveys, allowing them to be all different (which might be due to changes in weather conditions for example). Under this second model, six parameters will be estimated; one for occupancy and five for detection.

To run the first pre-defined model;

1. Ensure the **1 group, Constant P** model is highlighted.
2. Hit the **OK to Run** button.
3. After a brief pause, a message box will appear with some summary information about the results, asking you to confirm that you want these results added to the project file. Click **Yes**. Once confirmed, the results for that model will be saved and added to the project for later retrieval, and a summary of the model will appear in the **Results Browser**
4. Left-click on the model name to select it, then right-click to open a pop-up menu (as indicated above), and then select **View model output**. Alternatively, once the model has been selected you can click on the text icon to view the output. Figures 3.4 and 3.5 display the output that should appear in your default text editor (e.g., Notepad).

The output begins with some summary information about the data file and internal coding used by PRESENCE to define which model is being fit to the data. Following that there is some more summary information about the data file; number of sites (transects in this case), sampling occasions and covariates. The **Data checksum** is a single-value representation of the detection data which PRESENCE checks with each model run to ensure the detection data has not been changed as the user has access to the data file. The **Naive occupancy estimate** is the estimate of occupancy ignoring detection, i.e., the fraction of units where the species was detected at least once. Next in the output is the design matrices that were used to fit the model to the data. An explanation of them is deferred at this point to the next section.

Further down the output are the number of parameters used to fit the model, twice the negative log-likelihood value evaluated at the maximum likelihood estimates (which is used to calculate AIC values or could be used in likelihood ratio tests), and the AIC value for the model (discussed below). The log-likelihood is the value obtained by substituting in the values for ψ and p that are estimated from the data into the probability statements that were discussed earlier, then taking the natural logarithm of the evaluated probability statements. Next in the output are the estimated *beta parameters*, which are associated with the design matrices to which we shall return in the next section.

At this point, it is the estimated *real parameters* that are of most interest. The real parameters are the probabilities that are being estimated, occupancy and detection. Even though the output states **Individual site estimates of ...**, only estimates for site 1 are given. That is because there are no covariates in this example and the predefined model that was selected does not use covariates even if there were any available. For the real parameters, the estimate, its standard error and a 95% confidence interval are given in the output.

```

pres_1_group_Constant_P.out - Notepad
File Edit Format View Help
PRESENCE - Presence/Absence-Site occupancy data analysis
Mon Jun 11 23:06:07 2012, Version 4.4_120515
-----
==>i=br_salamanders_pao
==>l=pres_1_group_Constant_P.out
==>name=1_group, Constant P
==>model=100
==>j=f:\presence_files\br_salamanders_project\br_salamanders.dm
==>lmt=200
Varcov: nsig=6 eps=1.000000e-002
model=100 N,T-->39,5
modtype-->1

***** Input Data summary *****
Number of sites = 39
Number of sampling occasions = 5
Number of missing observations = 0
Data checksum = 29349

NSiteCovs-->0
NSampCovs-->0
Primary periods=1 Secondary periods: 5
Naive occupancy estimate = 0.4615

-----
BR Salamanders
-----
modtype=1 N=39 T=5 Groups=1 bootstraps=0

-->1-5
Matrix 1: rows=2, cols=2
      -,a1,
psi      1
=====
Matrix 2: rows=6, cols=2
p1      -,b1,
p2      1
p3      1
p4      1
p5      1
=====

```

Internal coding and command-line information

Data summary

Naive estimate

Occupancy design matrix

Detection design matrix

Figure 3.4: Beginning of the output for the **1 group, Constant P** model

The last part of the output is a *derived parameter* (parameters that are estimated using secondary calculations of real parameters) called conditional ψ^c , which is the probability of the species being present at a sampling unit, given the observed detection history for that unit (MacKenzie et al., 2006). This probability will be 1.0 for any unit at which the species was detected at least once (if the species has been seen at least once at a unit, then its presence at that unit has been confirmed) and will be some value between 0 and \hat{psi} for any unit where the species was never detected.

From Figure 3.5, $\hat{\psi}$ (i.e., 'psi' = 0.60 indicating that, on average, blue-ridge salamanders will be present on 6 out of every 10 randomly selected 50m transects, although from the 95% confidence interval, the value might be between 3.5-8 transects. The estimated detection probabilities suggest during each survey on an occupied transect, the probability of detecting blue-ridge salamanders in a single survey is about 0.26. The conditional ψ estimates are either 1.0 or 0.25, hence for those transects where blue-ridge salamanders were never detected after 5 surveys, the probability of salamanders being present is estimated to be 0.25. Had more surveys been conducted, and salamanders still had not been detected, then conditional ψ would get closer to 0.

The next step is to fit the second predefined model **1 group, Survey-specific p** using the same procedure as before, but selecting the second model in the pre-defined list and hitting **OK to Run**. After confirming this second model, your **Results Browser** should look like Figure 3.6.

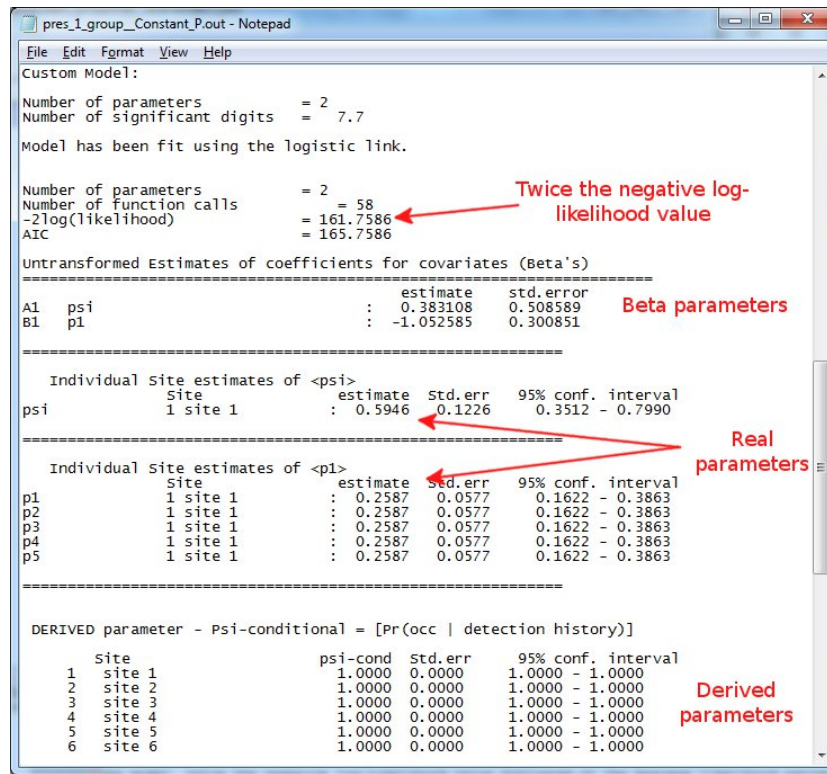


Figure 3.5: End of the output for the **1 group, Constant P** model

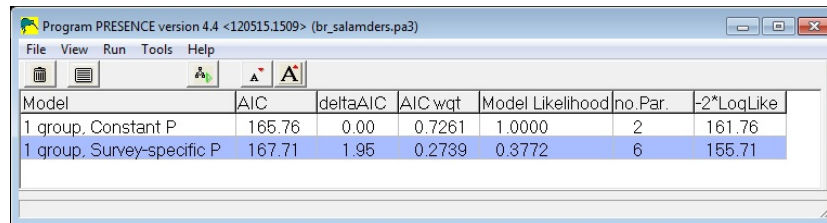


Figure 3.6: Results Browser after fitting first two pre-defined models

The Results Browser should now contain a summary for two models and the default is for PRESENCE to automatically rank them in terms of Akaike's Information Criterion (AIC). An in-depth discussion about AIC is not provided here (see Burnham and Anderson (2002); MacKenzie et al. (2006); or Google it), other than to say it is a method for comparing models based upon their relative distance from some unknown 'truth'. It's not possible to determine exactly how far each model is from this 'truth', but it is possible to determine which model is closest, second closest, etc. AIC is an approximation of this relative distance and for the m^{th} model fit to the data is calculated as:

$$AIC_m = -2l(\theta_m) + 2par_m$$

where $-2l(\theta_m)$ is twice the negative log-likelihood for the model evaluated at the maximum likelihood estimates (MLEs) for θ_m , θ_m is the set of parameters in the model that are estimated

and par_m is the number of parameters in model m .

There are seven columns in the Results Browser, each presenting a piece of summary information about the model, and how it ranks compared to others in the set.

1. **Model** is the name that was specified for each model
2. **AIC** is the AIC value for each model
3. **deltaAIC** is the relative difference in AIC values between each model and the currently top-ranked model (the one with smallest AIC)
4. **AIC wgt** is the AIC weight which is a measure of support for each model being the ‘best’ model (Burnham and Anderson, 2002)
5. **Model Likelihood** is the ratio of each models AIC weight over the model weight for the top-ranked model²
6. **no. Par** is the number of parameters in the model (par_m)
7. **-2*LogLike** is twice the negative log-likelihood evaluated at the MLEs ($-2l(\theta_m)$)

Examining the results, a difference of only 1.95 AIC units between these two models indicates that even though it is not ‘the best’, the second model still has a reasonable level of support and there is further evidence of this with the second model having a substantial amount of AIC weight. Thus, while most of the evidence points towards the probability of detection being constant (based upon the available data), the evidence is not overwhelming and there is some indication that detection probability may vary between surveys.

Alternatively, one could use these results to perform a likelihood ratio test of the null hypothesis that detection probability is equal on all five days, with the alternative hypothesis being that detection probability varies between surveys. The test statistic for this is $161.76 - 155.71 = 6.05$, which we could compare to the chi-square distribution with $6 - 2 = 4$ degrees of freedom (values can be read from the two rightmost columns in Figure 3.6). Doing so (with another piece of software) provides a p-value of 0.195 so there is insufficient evidence to reject the null hypothesis of constant detection.

Opening the output for the second model, note that it has the same layout as for the first model (Figures 3.4 and 3.5). The only major difference is towards the middle of the output where the real parameters are listed, rather than the same estimated value for detection probability being reported five times, there are now five different estimates. Note there is a large degree of variation in the estimated values of p , and that the likely reason this model does not have greater support is the (statistically) small sample size.

Finally, note that for both models, the estimated occupancy probability is very similar; 0.595 and 0.581 from the first and second models respectively. Hence, even though there is no clear indication of which model is the ‘best’, when interest is on estimating the probability of occupancy, both models give essentially the same results, and both are 26% larger than the naive estimate suggesting that the blue-ridge salamander was never detected at 1 in every 4 occupied transects.

²this likelihood is different from the likelihood discussed earlier derived from the probability statements of the observed data

3.2.3 Fitting a custom model

Continuing with this example, we shall now fit the same models as above, but using the **Custom Model** option. The **Custom Model** option provides the user with a high degree of flexibility for the types of models they would want to fit to the data as it provides the ability to include covariates on all parameter types and also create relationships and constraints between some parameters. The flexibility is achieved through the *design matrix*.

To perform a custom model analysis, begin a single-season analysis from the Run menu, or alternatively we can repeat the previous type of analysis by clicking on the icon with the green triangle towards the top of the **Results Browser** (this icon will be greyed out if no previous analysis has been conducted in this session). Selecting the **Custom** radio button in the Model box in the **Setup Numerical Estimation Run window** brings up the **Design Matrix** window. The design matrix is used to specify what factors you want to include in the model for each parameter type. Unlike Program MARK, in PRESENCE, the design matrix for each parameter type is given a separate sheet, and in the single-season model there are two parameter types: Occupancy and Detection probabilities (Figure 3.7).

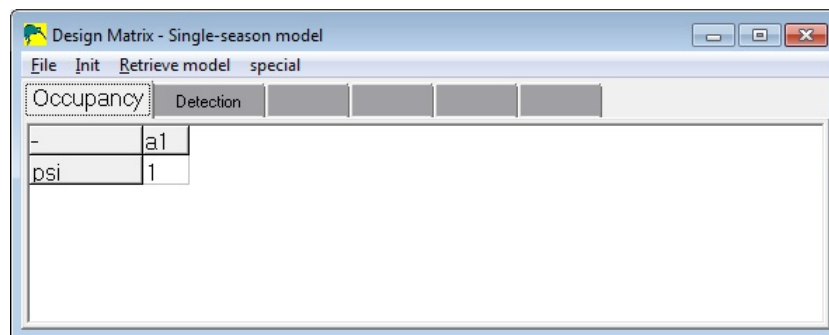


Figure 3.7: Design Matrix window for single-season model. Occupancy tab is visible.

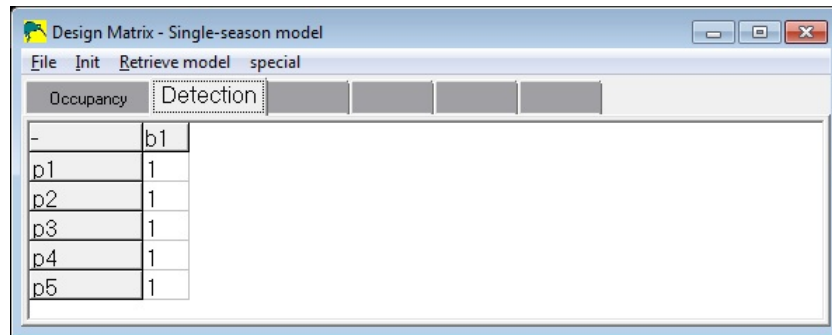
Each row of the design matrix represents a real parameter (e.g., occupancy and detection) which are the parameters that are used to construct the probability statements. The real parameters associated with each row can also be considered as the response variables which are going to be modeled. The columns of the design matrix represent the beta parameters; these are the parameters that are actually estimated and are the regression coefficients associated with the model that is defined by the cell entries in the design matrix. It should be noted that there is a third dimension to the design matrix that drills back into the computer screen, namely the sampling units. Therefore, if a numeric value is entered into the design matrix, that value is applied to all sampling units. If a covariate name is entered into a grid cell, the value of that covariate for each sampling unit is used in the regression equation.

A design matrix is read by moving along each row and summing the terms that result from multiplying the value in the cell by the corresponding beta parameter. Typically in PRESENCE where the real parameter (or response variable) is a probability, the logit-link function is implicitly assumed. For example, the design matrix in Figure 3.7 defines the following

logistic regression equation for the occupancy probability (the cell entry is in bold font):

$$\begin{aligned} \text{logit}(\psi_i) &= a1 \cdot \mathbf{1} \\ &= a1 \end{aligned} \quad (3.7)$$

That is, the probability that a sampling unit (transects in this example) is occupied is the same for all sampling units, the logit of which will be estimated as $a1$. The value of $a1$ is unknown and will be estimated from the data. For the remainder of this example the design matrix for occupancy will be left unchanged as there are no covariates in this data set. Note that the '.' symbol is equivalent to the multiplication symbol '×'.



	b1
-	b1
p1	1
p2	1
p3	1
p4	1
p5	1

Figure 3.8: Design Matrix window for single-season model. Detection tab is visible.

Next, select the tab labeled **Detection** and note there are now five rows in the design matrix (Figure 3.8, each one corresponding to the probability of detection in surveys 1 to 5 respectively (in general there will be one row in the detection design matrix for each survey)). For this design matrix there are five corresponding logistic regression equations:

$$\text{logit}(p_{i,1}) = b1 \cdot \mathbf{1} = b1 \quad (3.8)$$

$$\text{logit}(p_{i,2}) = b1 \cdot \mathbf{1} = b1 \quad (3.9)$$

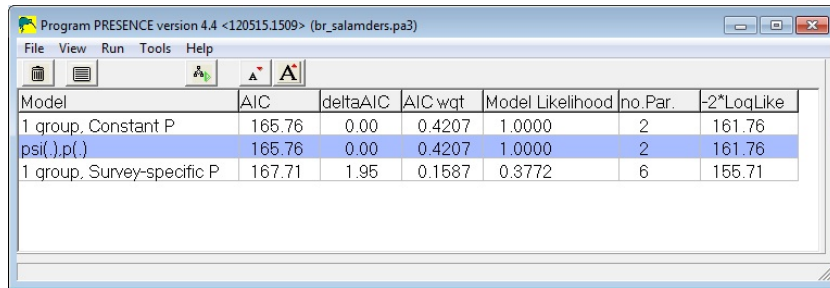
$$\text{logit}(p_{i,3}) = b1 \cdot \mathbf{1} = b1 \quad (3.10)$$

$$\text{logit}(p_{i,4}) = b1 \cdot \mathbf{1} = b1 \quad (3.11)$$

$$\text{logit}(p_{i,5}) = b1 \cdot \mathbf{1} = b1 \quad (3.12)$$

So in effect, this design matrix represents a model where detection probability is the same for all five surveys, the logit of which equals an amount $b1$. Therefore, this model is equivalent to the first predefined model; there is a single occupancy probability for all transects, and detection probability is constant both in time and across transects. To run this model, return to the **Setup Numerical Estimation Run** window (you may need to navigate through your task bar to locate the correct window), rename the model if desired³ and hit **OK to Run**. After confirming the results, the **Results Browser** should now look as in Figure 3.9.

³the model naming convention used here is to list the model parameter types with the factors included for that parameter given in parentheses. A dot '.' is used to denote a parameter that is constant.



Model	AIC	deltaAIC	AIC wgt	Model Likelihood	no.Par.	-2*LogLike
1 group, Constant P	165.76	0.00	0.4207	1.0000	2	161.76
psi(.).p(.)	165.76	0.00	0.4207	1.0000	2	161.76
1 group, Survey-specific P	167.71	1.95	0.1587	0.3772	6	155.71

Figure 3.9: Results Browser window after fitting the $\text{psi}(\cdot)\text{p}(\cdot)$ custom model.

Note that the results summary for the custom model that has just been fit (denoted as $\text{psi}(\cdot)\text{p}(\cdot)$) is identical to the first predefined model. This is because both are exactly the same model; the first predefined model is simply a shortcut for fitting this simple custom model. Opening the output for the $\text{psi}(\cdot)\text{p}(\cdot)$ model and comparing it the output for the first predefined model shows that exactly the same results have been obtained. Below the initial summary information in the output, the design matrices that have been used to fit the particular model is given which is useful as the exact interpretation of the beta parameters (i.e., regression coefficients) depends upon the design matrices that were specified.

Following the design matrices there is some more summary information, then a table titled **Untransformed Estimates of coefficients for covariates (Beta's)** which presents the estimated beta parameters (or regression coefficients), with the label for each beta parameter in the leftmost column corresponding to those used in the design matrices. So here $a1$ is estimated to be 0.383 with a standard error of 0.509, and $b1$ is estimated to be -1.053 with standard error 0.301.

The real parameter estimates (occupancy and detection probabilities) are given below the beta parameter estimates. As no covariates were used in this model, the values given apply to all transects. Recall that the real parameters are calculated using the design matrices and the logit link, so for occupancy the design matrix provides the equation (where the 'hats' indicate we now have estimated quantities):

$$\begin{aligned} \text{logit}(\hat{\psi}_i) &= \hat{a}1 \\ &= 0.383 \end{aligned}$$

$$\begin{aligned} \hat{\psi}_i &= \frac{e^{\hat{a}1}}{1 + e^{\hat{a}1}} \\ &= \frac{e^{0.383}}{1 + e^{0.383}} \\ &= 0.595 \end{aligned}$$

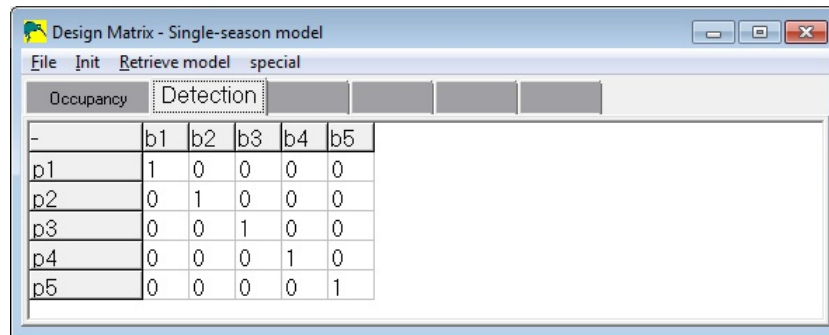
The standard error for $\hat{\psi}_i$ is obtained using a technique known as the delta method (e.g., MacKenzie et al. (2006)), which for this model is relatively simple because the transformed

quantity of interest is only a function of one beta parameter. Detail of the delta method is not given here as PRESENCE performs those calculations for the user, but that is the underlying method.

$$\begin{aligned} SE(\hat{\psi}_i) &= \hat{\psi}_i(1 - \hat{\psi}_i)SE(\widehat{a1}) \\ &= 0.595 \cdot 0.405 \cdot 0.509 \\ &= 0.123 \end{aligned}$$

A similar procedure is used to calculate the detection probabilities. Remember that the model fit here assumed that detection probability was the same for all five surveys so the same value is calculated five times. The final piece of the output for the custom model is a derived parameter: a quantity that is not directly in the model structure, but can be calculated from parameters that are. For the single-season model the derived parameter is the probability that a unit is occupied given that the species is never detected there. So in this example, if the salamander was not detected in any of the 5 surveys, what is the probability of blue-ridge salamanders being present at a transect (see MacKenzie et al. (2006) for details)? Here that is estimated to be 0.247 with a standard error of 0.147.

To fit the equivalent of the **1 group, Survey-specific P** model, again select to run a custom, single-season model. Leave the design matrix for occupancy as the default option (as in the last model), and set the design matrix for detection probability to be as in Figure 3.10.



-	b1	b2	b3	b4	b5
p1	1	0	0	0	0
p2	0	1	0	0	0
p3	0	0	1	0	0
p4	0	0	0	1	0
p5	0	0	0	0	1

Figure 3.10: Detection tab of the Design Matrix window for fitting the model $\psi_i(\cdot)p(\text{Survey})$ custom model.

A shortcut to creating this design matrix is to select the menu item `Init>Full Identity` in the **Design Matrix** window. This creates the identity matrix which has as many columns as rows, with 1's on the main diagonal and 0's elsewhere. From this matrix we can write the following sets of equations.

$$\text{logit}(p_{i,1}) = b_1 \cdot \mathbf{1} + b_2 \cdot \mathbf{0} + b_3 \cdot \mathbf{0} + b_4 \cdot \mathbf{0} + b_5 \cdot \mathbf{0} = b_1 \quad (3.13)$$

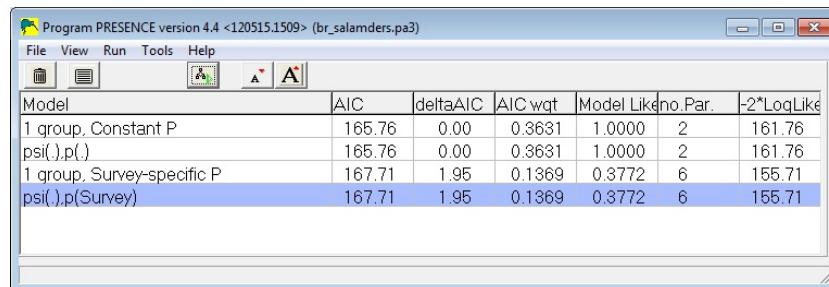
$$\text{logit}(p_{i,2}) = b_1 \cdot \mathbf{0} + b_2 \cdot \mathbf{1} + b_3 \cdot \mathbf{0} + b_4 \cdot \mathbf{0} + b_5 \cdot \mathbf{0} = b_2 \quad (3.14)$$

$$\text{logit}(p_{i,3}) = b_1 \cdot \mathbf{0} + b_2 \cdot \mathbf{0} + b_3 \cdot \mathbf{1} + b_4 \cdot \mathbf{0} + b_5 \cdot \mathbf{0} = b_3 \quad (3.15)$$

$$\text{logit}(p_{i,4}) = b_1 \cdot \mathbf{0} + b_2 \cdot \mathbf{0} + b_3 \cdot \mathbf{0} + b_4 \cdot \mathbf{1} + b_5 \cdot \mathbf{0} = b_4 \quad (3.16)$$

$$\text{logit}(p_{i,5}) = b_1 \cdot \mathbf{0} + b_2 \cdot \mathbf{0} + b_3 \cdot \mathbf{0} + b_4 \cdot \mathbf{0} + b_5 \cdot \mathbf{1} = b_5 \quad (3.17)$$

Note that the values associated with each beta parameter in each equation (i.e., the 1's or 0's) correspond with the entries in the design matrix, and result in a model where the probability of detecting the salamanders in each of the five surveys will be estimated by a different beta parameter (i.e., b_1 - b_5). Going back to the **Setup Numerical Estimation Run** window, the model name needs to be changed before running the model. Model names are completely arbitrary and the main point is that the model name should be meaningful to the user. The suggested name here is **psi(.),p(Survey)** indicating that occupancy is the same for all transects and detection probability is different for each survey. After renaming the model, hit **Ok to Run**, and after confirming the results, the Results Browser will appear as in Figure 3.11 where the summary information for the model **psi(.),p(Survey)** is the same as the predefined model, **1 group, Survey-specific P**.



Model	AIC	deltaAIC	AICwgt	Model Lik	no.Par.	-2*LogLike
1 group, Constant P	165.76	0.00	0.3631	1.0000	2	161.76
psi(.),p(.)	165.76	0.00	0.3631	1.0000	2	161.76
1 group, Survey-specific P	167.71	1.95	0.1369	0.3772	6	155.71
psi(.),p(Survey)	167.71	1.95	0.1369	0.3772	6	155.71

Figure 3.11: Results Browser window after fitting the **psi(.),p(Survey)** custom model.

Opening the output for this model we again see the initial summary information followed by the design matrices. Note the more complex design matrix for detection probability. Below, the table of beta parameter estimates is presented and given the defined design matrix, the estimated beta parameters b_1 - b_5 represent the absolute probability of detecting salamanders at an occupied transect in each survey (on the logit-scale). The real parameter estimates are calculated in the same way as for the previous custom model.

Given the duplicate models that have been fit to the data, attempting to interpret the Results Browser window would not be very meaningful, particularly with respect to AIC model weights and Model Likelihoods.

3.2.4 More on interpretation of estimates

The correct interpretation of the estimated occupancy probability would be that it is the probability of blue-ridge salamanders being present at a randomly selected transect from within the

region of sampling. For example, from the $\mathbf{psi}(\cdot), \mathbf{p}(\cdot)$ model, $\hat{\psi}_i = 0.595$, implying that if a transect was randomly selected then the probability of blue-ridged salamanders being at that transect would be, approximately, 0.6. Therefore, if 10 transects were randomly selected then it would be expected (on average) that 6 of those transects would be occupied by blue-ridge salamanders and 4 unoccupied.

Detection probabilities relate to the probability of finding the species in a single survey, using the field methods that were employed, given the species was actually present at the sampling unit. From the $\mathbf{psi}(\cdot), \mathbf{p}(\cdot)$ model, the estimated detection probability was 0.247, suggesting that if blue-ridge salamanders were present on a transect, they would only be detected there once out of every 4 surveys, i.e., 25%.

3.3 Example 2: Mahoenui Giant Weta

In this example we shall fit models that involve covariates for both occupancy and detection probabilities to data that has been collected on the Mahoenui giant weta (*Deinacrida mahoenui*) in a scientific reserve by the Department of Conservation in the King Country district of New Zealand. Seventy-two 3m-radius circular plots were surveyed between 3-5 times for weta. Each plot was assessed as to the level of browsing by feral goats and each survey was conducted by 1 of 3 observers. The level of browsing is an indicator of habitat condition at each plot; browsed bushes have denser foliage while unbrowsed bushes tend to more open. It was believed that the level of browsing at each plot would influence whether weta were present, with browsed bushes being preferred as they provide better refuge from introduced mammalian predators. It was also believed that the observers would differ in their ability to find weta due to previous experience. Therefore, the level of browsing and observer are going to be considered as site-specific and sampling occasion covariates respectively. This data set was examined on page 116 of MacKenzie et al. (2006), see there for additional details.

The data is included in the sample data folder that is installed along with PRESENCE in the spreadsheet **Weta_pg116.xls**. This file consists of 5 sheets containing the detection-nondetection data (on the sheet called 'detection_histories'), whether a plot was browsed or unbrowsed ('site_covar'), and which observer conducted which survey ('Obs1', 'Obs2' and 'Obs3'). Each of the covariates are dummy variables that =1 if the covariate is of the value indicated by the covariate name, and =0 otherwise. In this example the number of surveys is not constant for all plots hence the detection-nondetection data includes missing observations that are indicated with a '-'. Note that the observer covariates contain missing values too, and that these exactly correspond with the missing values in the detection-nondetection data. Covariate values are allowed to be missing, but only if the detection-nondetection data for the corresponding survey of the sampling unit is also missing. This assumption is required by most statistical methods. Missing values for site-specific covariates are not allowed.

3.3.1 Starting a new project and entering the data

Complete the following steps to create the PRESENCE project:



Figure 3.12: Photo credit: Mahoenui giant weta, Amanda Smale, Department of Conservation, New Zealand.

1. Begin PRESENCE, start a new project and open the data input form.
2. Open the file **Weta_pg116.xls** from the sample data folder using a spreadsheet package.
3. Copy and paste the detection-nondetection data from the spreadsheet into PRESENCE in the same manner as for the previous example.
4. On the **Data Input Form** change the number of site-specific covariates to **2** and number of sampling occasion covariates to **3**. Note that additional tabs appear for sheets on which to enter the covariate data (Figure 3.13).
5. Select the spreadsheet labeled **site_covar**, highlight the range of the covariate data including the covariate names (i.e., cells A1:B73) and copy the selected cells.
6. Return to the **Data Input Form** and select the tab labeled **Site Covar**.
7. Click on the top-left grid cell and make sure the cell border becomes a dotted line (not a flashing cursor), then select Edit>Paste>Paste w/covnames (Figure ??). This paste option pastes the covariate values into the grid cells and automatically renames the covariates.
8. Select the spreadsheet labeled **Obs1** and copy the values (cells A1:E72)
9. Return to the **Data Input Form** and select the tab labeled **SampCov1**. Make the top-left cell (for site 1, survey 1) active and paste the values by selecting Edit>Paste>Paste values.

10. Rename the covariate by selecting **Edit>Rename covariate** (Figure ??) and in the text box type **Obs1** then hit **OK**.
11. Repeat the previous 3 steps for the covariates Obs2 and Obs3.
12. Once you have entered all 3 observer covariates, save the data file using an appropriate name (e.g., **weta**). Select **No** when prompted about the last column being frequency data, and specify a meaningful title.
13. After successfully saving the data file, close the **Data Input Form** which will return you to the **Enter Specifications for PRESENCE Analysis** window.
14. Check there are 72 sites, 5 surveys, 2 site covariates and 3 sampling occasion covariates. If satisfied select **OK**.
15. After a couple of seconds a blank results browser should appear. Remember, if you do not see the results browser, you have not successfully set up your project file.

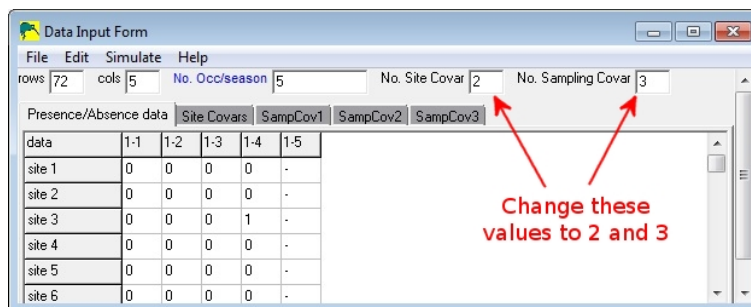


Figure 3.13: Adjust the number of site-specific and sampling occasion covariates.

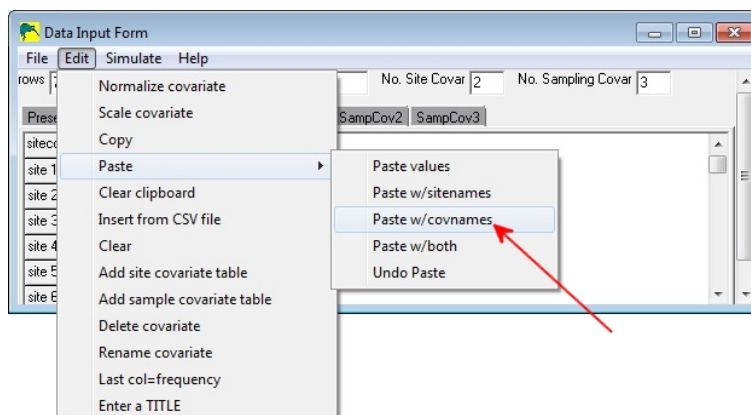


Figure 3.14: Pasting with covariate names for site-specific covariates.

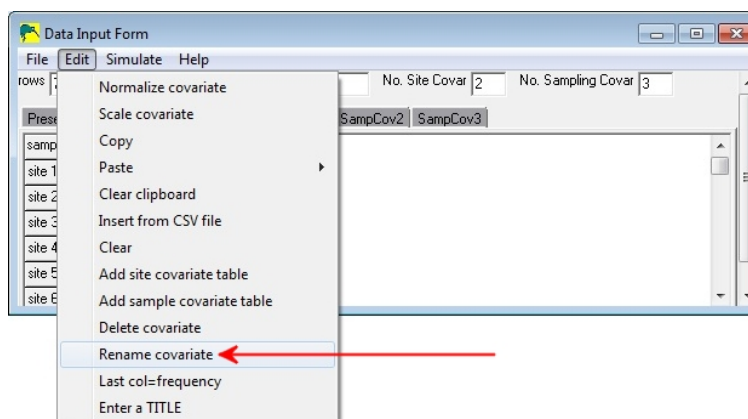


Figure 3.15: Menu option for renaming a covariate.

3.3.2 Including covariates for occupancy

Here, focus shall be upon fitting models that include covariates on occupancy and/or detection. Obviously, models without covariates, as in the previous example, could also be considered and fit to the data. Doing so would enable an assessment to be made about the importance of the potential covariates either on the basis of AIC or likelihood-ratio tests. This will not be done in this example, although users are encouraged to attempt such models with this data set.

The first model to fit supposes the probability of occupancy varies between plots according to the level of browsing, and detection probability varies among days but is the same for all 3 observers. This model could be called **psi(Browsed),p(Day)**. First, we shall fit the model, then work back through it explaining what has actually been done. After setting up the project:

1. Select **Run>Analysis:single-season** and click on the **Custom** radio button. The **Design Matrix** window should appear.
2. On the occupancy design matrix, right-click anywhere on the design matrix to open a pop-up menu and select **Add Col** (the 5th item in the menu) to add an empty column on the right-hand side of the design matrix.
3. Left-click on the new column, then select **Init>*Browsed** and the covariate name *Browsed* will be inserted into the column (Figure 3.16). All available covariates are listed in the **Init** menu, prefixed with a '*'.
 - Alternatively the covariate name could be typed directly into the cell, although the name must be 100% correct and **PRESENCE** is case sensitive.
4. Select the detection design matrix, then select **Init>Full Identity** to allow daily detection probabilities (Figure 3.16).
5. Select the **textbfSetup Numerical Estimation Run** window, rename the model (e.g., **psi(Browsed),p(Day)**) then hit **OK to Run**.
6. Confirm the results when prompted.

The figure consists of two screenshots of a software window titled "Design Matrix - Single-season model".

The top screenshot shows the "Occupancy" tab selected. The table below it has two columns labeled "a1" and "a2", and one row labeled "psi" with values "1" and "Browsed".

	a1	a2
psi	1	Browsed

The bottom screenshot shows the "Detection" tab selected. The table below it has five columns labeled "b1" through "b5" and five rows labeled "p1" through "p5". Each row has a single "1" in a different column, representing a design matrix for detection probability.

	b1	b2	b3	b4	b5
p1	1	0	0	0	0
p2	0	1	0	0	0
p3	0	0	1	0	0
p4	0	0	0	1	0
p5	0	0	0	0	1

Figure 3.16: Design matrices for the model $\text{psi}(\text{Browsed}), p(\text{Day})$.

So what has happened? Lets start with the simple one; the design matrix for detection probability. Recall that this design matrix has already been used in the blue-ridge salamander example and that to read the design matrix, move along each row summing the terms produced by multiplying the values in the grid cell with the corresponding beta parameters for each column. The design matrix for detection probability represents the set of equations:

$$\text{logit}(p_{i,1}) = b1 \cdot 1 + b2 \cdot 0 + b3 \cdot 0 + b4 \cdot 0 + b5 \cdot 0 = b1 \quad (3.18)$$

$$\text{logit}(p_{i,2}) = b1 \cdot 0 + b2 \cdot 1 + b3 \cdot 0 + b4 \cdot 0 + b5 \cdot 0 = b2 \quad (3.19)$$

$$\text{logit}(p_{i,3}) = b1 \cdot 0 + b2 \cdot 0 + b3 \cdot 1 + b4 \cdot 0 + b5 \cdot 0 = b3 \quad (3.20)$$

$$\text{logit}(p_{i,4}) = b1 \cdot 0 + b2 \cdot 0 + b3 \cdot 0 + b4 \cdot 1 + b5 \cdot 0 = b4 \quad (3.21)$$

$$\text{logit}(p_{i,5}) = b1 \cdot 0 + b2 \cdot 0 + b3 \cdot 0 + b4 \cdot 0 + b5 \cdot 1 = b5 \quad (3.22)$$

That is, detection probability on each day is represented by a different value $b1$ - $b5$. Graphically, this could be presented as in Figure 3.17.

It is important to note that at this stage the values for $b1$ - $b5$ are unknown (they will be estimated by PRESENCE from the data), the graph is just to conceptualize the relationship

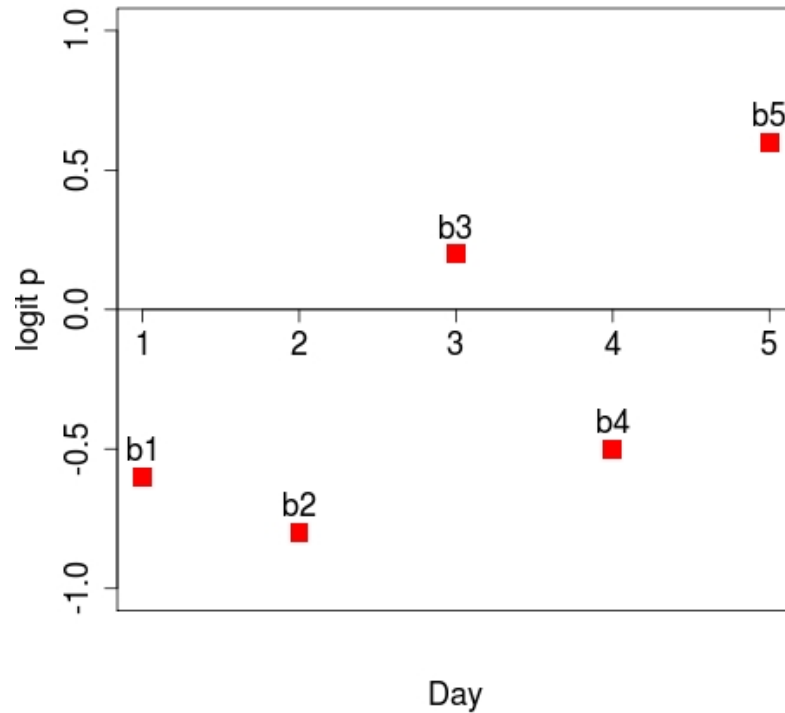


Figure 3.17: Graphical representation of daily detection probabilities.

between detection probability and day for the model that has just been fit to the data. In this case, there is no relationship or constraints on how detection probability might vary among days; each day is free to have a unique value that is different to the others. That is, this design matrix represents the model that allows detection probability to be day specific, with all survey plots having the same probability on each day (as no site-specific covariates have been included for detection).

Next, consider the design matrix for the occupancy probability, which is a representation of the following equation.

$$\begin{aligned} \text{logit}(\psi_i) &= a_1 \cdot \mathbf{1} + a_2 \cdot \mathbf{Browsed}_i \\ &= a_1 + a_2 \cdot \mathbf{Browsed}_i \end{aligned} \quad (3.23)$$

Recall that *Browsed* is a covariate that was defined to =1 if plot *i* showed signs of browsing, and =0 otherwise. Therefore, for an unbrowsed plot (where $\mathbf{Browsed}_i = 0$),

$$\begin{aligned} \text{logit}(\psi_i) &= a_1 \cdot 1 + a_2 \cdot 0 \\ &= a_1 \end{aligned} \quad (3.24)$$

and for a browsed plot (where $Browsed_i = 1$),

$$\begin{aligned} \text{logit}(\psi_i) &= a_1 \cdot 1 + a_2 \cdot 1 \\ &= a_1 + a_2 \end{aligned} \quad (3.25)$$

Note that a_2 is therefore the difference in occupancy between a browsed and unbrowsed plot (on the logit scale), or alternatively what effect browsing has on occupancy compared to plots with no browsing (Figure 3.18).

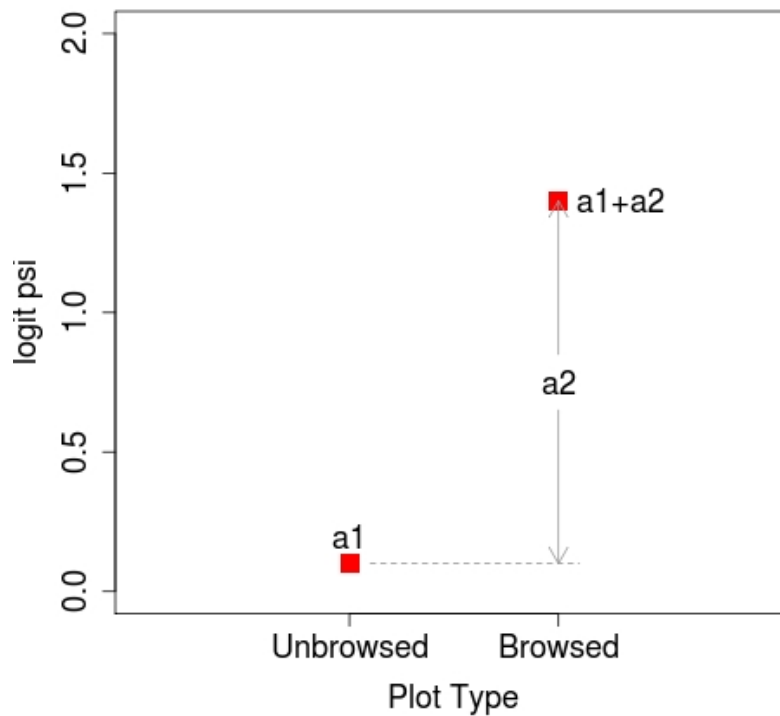


Figure 3.18: Graphical representation of difference in occupancy at unbrowsed and browsed plots.

Looking at the output from this model, $\widehat{a_2} = 1.24$ which, as it is >0 , indicates that the probability of occupancy is higher at browsed plots. In this fairly simple model, $\widehat{a_2}$ can be interpreted in a fairly straightforward manner, although for more complicated models it can be useful to interpret the effect of a covariate in terms of an odds-ratio. Recall that the logit-link is that natural logarithm of the odds of a 'successful' event (in this case, the presence of the weta), and that to calculate an odds-ratio, we can take the inverse-logarithm of the beta parameter. That is:

$$\begin{aligned}\widehat{OR}_{Browsed} &= e^{\widehat{a}^2} \\ &= e^{1.24} \\ &= 3.44\end{aligned}$$

Interpreting this odds-ratio, for every plot where weta are absent, weta would be present in 3.44 times more plots that had been browsed than had not been browsed. That is, if for every unbrowsed plot where weta were absent there were 1.50 unbrowsed plots where weta were present, then for every browsed plot where weta were absent there would be 5.16 (1.50×3.44) browsed plots where weta were present. An approximate 2-sided 95% confidence interval for the odds-ratio would be:

$$\begin{aligned}&= (e^{1.24-2 \cdot 0.787}, e^{1.24+2 \cdot 0.787}) \\ &= (e^{-0.27}, e^{2.74}) \\ &= (0.76, 15.49)\end{aligned}$$

3.3.3 Including covariates on detection

As noted previously, both site-specific and sampling-occasion covariates may be used to model detection probabilities. The mechanics of including them is essentially the same as above as one can regard a site-specific covariate as a sampling-occasion covariate whose value may be different in different sampling units, but is unchanging over time. For example, suppose the model $\text{psi}(\mathbf{Browsed}), \mathbf{p}(\mathbf{Browsed})$ was to be fit to the data, i.e., allowing the level of browsing to effect both occupancy and detection probabilities (Figure 3.19).

The occupancy design matrix is unchanged from before, so lets focus on the detection probability design matrix. To include the *Browsed* covariate for detection a column needs to be added to the default design matrix (recall; right-click on the design matrix to open the pop-up menu), then left-click on a cell in the new column and choose `Init>*Browsed`. The word *Browsed* is repeated in each row of the design matrix indicating that for all surveys of plot i , the value of $Browsed_i$ will be used. As defined here, the design matrix implies the following series of equations:

$$\text{logit}(p_{i,1}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Browsed}_i = b1 + b2 \cdot Browsed_i \quad (3.26)$$

$$\text{logit}(p_{i,2}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Browsed}_i = b1 + b2 \cdot Browsed_i \quad (3.27)$$

$$\text{logit}(p_{i,3}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Browsed}_i = b1 + b2 \cdot Browsed_i \quad (3.28)$$

$$\text{logit}(p_{i,4}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Browsed}_i = b1 + b2 \cdot Browsed_i \quad (3.29)$$

$$\text{logit}(p_{i,5}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Browsed}_i = b1 + b2 \cdot Browsed_i \quad (3.30)$$

As *Browsed* is a site-specific covariate that has the same value for all surveys, and there are no other covariates or time effects included in the model, the same equation is repeated

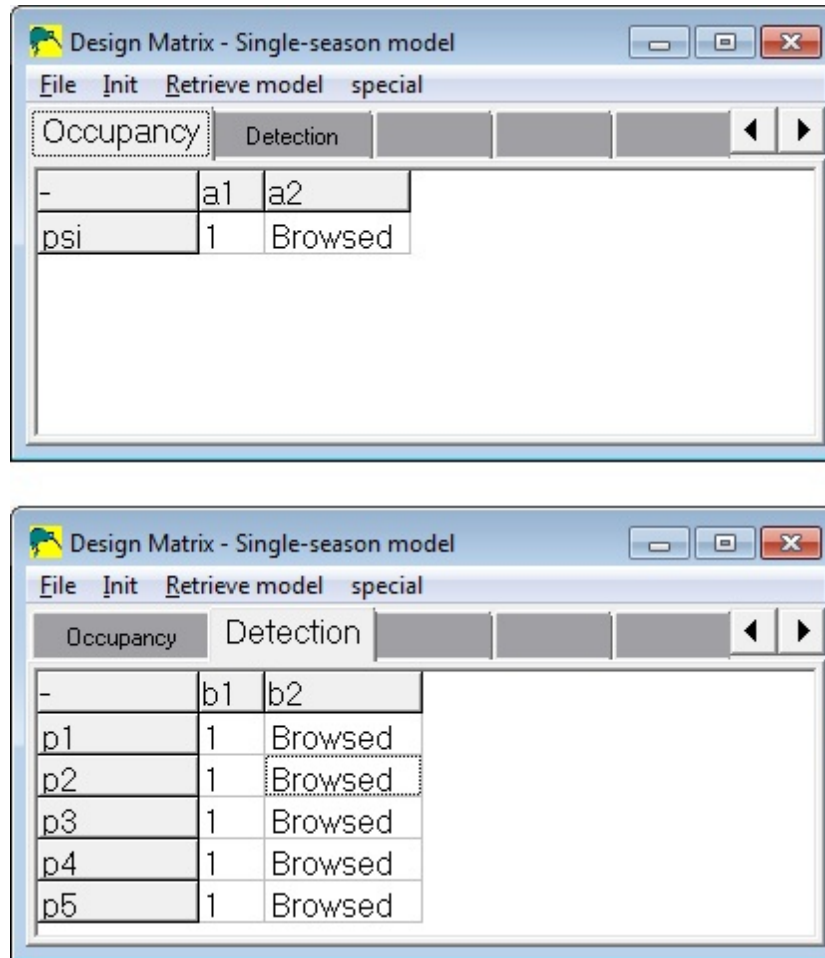


Figure 3.19: Design matrices for the model $\text{psi}(\text{Browsed}), p(\text{Browsed})$.

five times. For unbrowsed plots ($\text{Browsed}_i = 0$), the equation becomes

$$\text{logit}(p_{i,j}) = b1 \quad (3.31)$$

and for a browsed plot ($\text{Browsed}_i = 1$),

$$\text{logit}(p_{i,j}) = b1 + b2 \quad (3.32)$$

hence $b2$ indicates how different detection of weta is (on the logit-scale) in a browsed vs unbrowsed plot.

Including a sampling-occasion covariate for detection follows essentially the same process, although now the value of a sampling-occasion covariate is, potentially, different for each survey of the same plot. For example, for this study the observers were rotated around the different plots hence the probability of detecting weta in a particular survey could be different depending on which observer conducted the survey. In the data file, the $Obs1$, $Obs2$ and $Obs3$ covariates indicate which plot was surveyed by which observer on a particular day. So, $Obs2_{i,j} = 1$ if observer 2 surveyed plot i on day j , or $= 0$ if the survey was conducted by

observers 1 or 3 (i.e., not by observer 2). The model $\text{psi}(\text{Browsed}), p(\text{Obs})$ can be fit to the data with the design matrices given in Figure 3.20.

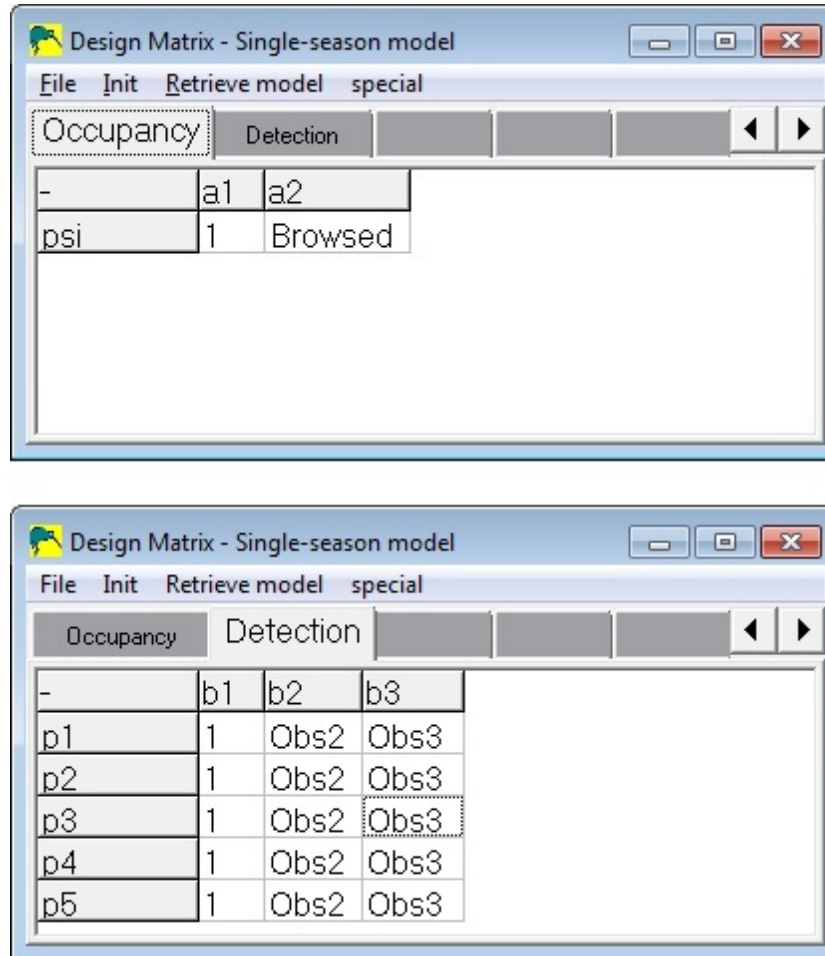


Figure 3.20: Design matrices for the model $\text{psi}(\text{Browsed}), p(\text{Obs})$.

To include the *Obs2* and *Obs3* covariates, add 2 columns then select the respective covariates from the *Init* menu. The *Obs1* covariate does not appear in the design matrix as observer 1 is going to be treated as the standard or control to which the other observers will be compared. The design matrix represents the following series of equations for detection:

$$\text{logit}(p_{i,1}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Obs2}_{i,1} + b3 \cdot \mathbf{Obs3}_{i,1} = b1 + b2 \cdot \text{Obs2}_{i,1} + b3 \cdot \text{Obs3}_{i,1} \quad (3.33)$$

$$\text{logit}(p_{i,2}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Obs2}_{i,2} + b3 \cdot \mathbf{Obs3}_{i,2} = b1 + b2 \cdot \text{Obs2}_{i,2} + b3 \cdot \text{Obs3}_{i,2} \quad (3.34)$$

$$\text{logit}(p_{i,3}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Obs2}_{i,3} + b3 \cdot \mathbf{Obs3}_{i,3} = b1 + b2 \cdot \text{Obs2}_{i,3} + b3 \cdot \text{Obs3}_{i,3} \quad (3.35)$$

$$\text{logit}(p_{i,4}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Obs2}_{i,4} + b3 \cdot \mathbf{Obs3}_{i,4} = b1 + b2 \cdot \text{Obs2}_{i,4} + b3 \cdot \text{Obs3}_{i,4} \quad (3.36)$$

$$\text{logit}(p_{i,5}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{Obs2}_{i,5} + b3 \cdot \mathbf{Obs3}_{i,5} = b1 + b2 \cdot \text{Obs2}_{i,5} + b3 \cdot \text{Obs3}_{i,5} \quad (3.37)$$

If the *j*th survey of a plot was conducted by observer 1, hence both $\text{Obs2}_{i,j}$ and $\text{Obs3}_{i,j} = 0$,

the equations would reduce to:

$$\text{logit}(p_{i,j}) = b_1, \quad (3.38)$$

or if observer 2 had conducted the survey, where $Obs_{2i,j} = 1$ and $Obs_{3i,j} = 0$;

$$\text{logit}(p_{i,j}) = b_1 + b_2, \quad (3.39)$$

and

$$\text{logit}(p_{i,j}) = b_1 + b_3, \quad (3.40)$$

if observer 3 had conducted the j th survey of the plot, hence $Obs_{2i,j} = 0$ and $Obs_{3i,j} = 1$. Therefore b_2 and b_3 in this model indicate how different detection probability is (on the logit scale) for observers 2 and 3, respectively, compared to observer 1; a negative value would indicate they are less effective at detecting weta than observer 1, while a positive value would suggest they detect weta more frequently (Figure 3.21).

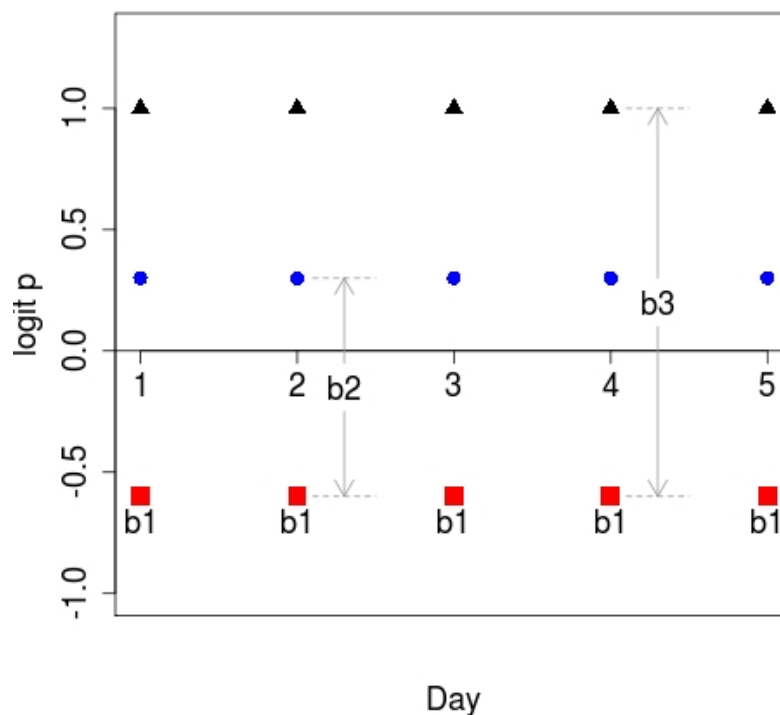


Figure 3.21: Graphical representation of observer effect on detection probabilities. Detection for observers 1, 2 and 3 are indicated by the red squares, blue circles and black triangles respectively.

Note that users are not restricted to only fitting univariate models (i.e., models with single covariates). For example, it may be reasonable to consider that both the level of browsing

at a plot and observers have an effect on the detection probability. It is relatively straight forward to fit such models in PRESENCE by adding additional columns to the design matrix and inserting the required covariates.

3.3.4 Bringing it together

For this final model that shall be fit in this weta example, the entire process for fitting a model to the data shall be worked through from conceptualizing the desired model, to construction of the design matrices, to interpretation of results.

The single-season occupancy model has two components, occupancy and detection. In PRESENCE, there are separate design matrices enabling the user to focus on each individual component when constructing models to fit to the data. This can be useful to compartmentalize the problem, although the user must recognize and understand that the two components are inextricably linked and that the results suggested for one component will often be influenced by what has been specified for the other component.

Here a model will be fit where the presence of weta is affected by the level of browsing at a plot, while detection varies by survey day (e.g., due to variation in weather conditions) and may also be different for each observer. Further, the differences between observers was thought to be relatively consistent over the week of survey because any difference is due to previous experience, i.e., it would be expected that one observer would consistently be the best at finding weta and another to consistently be the worse, rather than a situation where the relative abilities of the observers is changing with each day. In terms of the model notation used here, the model we wish to fit to the data could be called **psi(Browsed),p(Day+Obs)**; the '+' is indicating an *additive* effect between day and observers implying the consistency of the effect across the other factors.

The psi(Browsed) portion of the model has been used previously so the same design matrix can be reused here. In order to define the p(Day+Obs) component, the form of the design matrix needs to be determined. For users with some experience in linear modeling (e.g., linear regression or generalized linear models), construction of the design matrix from an equation that represents the model of interest may be relatively straight forward, but for those with little or no experience it can often help to step through the process while learning these techniques. Often, however, roughly sketching the expected relationship between the factors of interest and occupancy or detection probabilities can be a useful starting point to visualize the type of model that is desired, use that to define a series of equations for each real parameter, then use the equations to construct the design matrix.

Figure 3.22 is a graphical representation of the desired model for detection probability. There is clearly daily variation in detection, and note the consistent difference between each observer. The sketch has also been labeled, with the logit-detection probabilities for observer 1 denoted b_1 - b_5 , the difference between observers 1 and 2 on each day labeled b_6 and the difference between observers 1 and 3 each day labeled b_7 . From this sketch we need to create a series of equations, one for each row of the detection design matrix (i.e., one for the detection probability on each survey day) utilizing the covariates that are available. Note that for each day, there are currently three detection probabilities (one for each observer), so the first step is

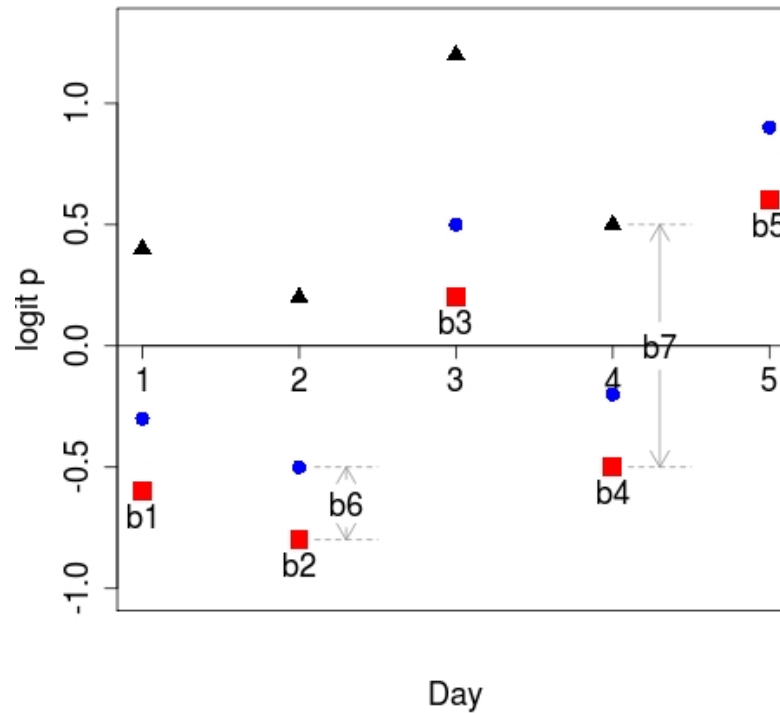


Figure 3.22: Graphical representation of daily and observer effects on detection probabilities. Detection for observers 1, 2 and 3 are indicated by the red squares, blue circles and black triangles respectively.

to find a form that allows the three probabilities to be expressed as a single equation. Focusing on the first survey day, the three points could be defined by the following equations:

$$\text{logit}(p_{i,1,Obs_1}) = b1 \quad (3.41)$$

$$\text{logit}(p_{i,1,Obs_2}) = b1 + b6 \quad (3.42)$$

$$\text{logit}(p_{i,1,Obs_3}) = b1 + b7 \quad (3.43)$$

that is, if the survey was conducted by observer 2, then add an amount of $b6$ on to the value for observer 1, and if the survey was conducted by observer 3, add an amount $b7$ to the value for observer 1. This can be collapsed to a single equation through the observer indicator variables; recall that if a survey was performed by observer 1, both $Obs2_{i,j}$ and $Obs3_{i,j} = 0$; for observer 2, $Obs2_{i,j} = 1$ and $Obs3_{i,j} = 0$; and for observer 3 then $Obs2_{i,j} = 0$ and $Obs3_{i,j} = 1$. Hence, detection probability on day 1 can be expressed by the single equation:

$$\text{logit}(p_{i,1}) = b1 + b6 \cdot Obs2_{i,1} + b7 \cdot Obs3_{i,1}, \quad (3.44)$$

and doing the same for the other days yields the series of equations

$$\text{logit}(p_{i,1}) = b1 + b6 \cdot \text{Obs2}_{i,1} + b7 \cdot \text{Obs3}_{i,1} \quad (3.45)$$

$$\text{logit}(p_{i,2}) = b2 + b6 \cdot \text{Obs2}_{i,2} + b7 \cdot \text{Obs3}_{i,2} \quad (3.46)$$

$$\text{logit}(p_{i,3}) = b3 + b6 \cdot \text{Obs2}_{i,3} + b7 \cdot \text{Obs3}_{i,3} \quad (3.47)$$

$$\text{logit}(p_{i,4}) = b4 + b6 \cdot \text{Obs2}_{i,4} + b7 \cdot \text{Obs3}_{i,4} \quad (3.48)$$

$$\text{logit}(p_{i,5}) = b5 + b6 \cdot \text{Obs2}_{i,5} + b7 \cdot \text{Obs3}_{i,5} \quad (3.49)$$

Note the similarities with each equation, and that the only essential difference is the first term in the equation, i.e., the intercept term. However, in order to construct the design matrix all of the beta parameters (i.e., the regression coefficients) have to be included in each equation. This is achieved very simply by recognizing that any beta parameter multiplied by 0 is still 0, and similarly, any beta parameter multiplied by 1 equals the beta parameter again. Therefore, the series of equations can be expanded as:

$$\text{logit}(p_{i,1}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{0} + b5 \cdot \mathbf{0} + b6 \cdot \mathbf{Obs2}_{i,1} + b7 \cdot \mathbf{Obs3}_{i,1} \quad (3.50)$$

$$\text{logit}(p_{i,2}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{1} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{0} + b5 \cdot \mathbf{0} + b6 \cdot \mathbf{Obs2}_{i,2} + b7 \cdot \mathbf{Obs3}_{i,2} \quad (3.51)$$

$$\text{logit}(p_{i,3}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{1} + b4 \cdot \mathbf{0} + b5 \cdot \mathbf{0} + b6 \cdot \mathbf{Obs2}_{i,3} + b7 \cdot \mathbf{Obs3}_{i,3} \quad (3.52)$$

$$\text{logit}(p_{i,4}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{1} + b5 \cdot \mathbf{0} + b6 \cdot \mathbf{Obs2}_{i,4} + b7 \cdot \mathbf{Obs3}_{i,4} \quad (3.53)$$

$$\text{logit}(p_{i,5}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{0} + b5 \cdot \mathbf{1} + b6 \cdot \mathbf{Obs2}_{i,5} + b7 \cdot \mathbf{Obs3}_{i,5} \quad (3.54)$$

To construct the design matrix (Figure 3.23), the value associated with each beta parameter in each equation (the values in **bold**) are entered into the respective cell of the design matrix. Note that the number of beta parameters used in the equations defines the number of columns required in the design matrix, i.e., in this case, seven.

	b1	b2	b3	b4	b5	b6	b7
p1	1	0	0	0	0	Obs2	Obs3
p2	0	1	0	0	0	Obs2	Obs3
p3	0	0	1	0	0	Obs2	Obs3
p4	0	0	0	1	0	Obs2	Obs3
p5	0	0	0	0	1	Obs2	Obs3

Figure 3.23: Detection design matrix for the model $\text{psi}(\text{Browsed}), \text{p}(\text{Day}+\text{Obs})$.

Set up this design matrix in PRESENCE (and also the occupancy design matrix corresponding to the **(psi(Browsed))** portion of the model as used previously), rename (e.g., $\text{psi}(\text{Browsed}), \text{p}(\text{Day}+\text{Obs})$) and run the model.


```

Untransformed Estimates of coefficients for covariates (Beta's)
-----
      estimate      std.error
A1  psi           :  0.023993    0.459662
A2  psi.Browsed  :  1.169320    0.741302
B1  p1           : -1.299719    0.524637
B2  p2           : -1.454611    0.531320
B3  p3           : -2.242910    0.576175
B4  p4           : -1.370374    0.509675
B5  p5           : -0.264264    0.513130
B6  p1.Obs2     :  0.726461    0.467840
B7  p1.Obs3     :  1.070017    0.460040
-----

```

Figure 3.24: Beta parameter estimates from the model $\psi(\text{Browsed}), p(\text{Day}+\text{Obs})$.

Open the output and locate the table of beta parameter estimates (Figure 3.24). Focusing on occupancy initially, the two estimated beta parameters of interest are $\hat{a}_1 = 0.02$ and $\hat{a}_2 = 1.17$. With these, the equation for estimating the probability of occupancy becomes:

$$\text{logit}(\psi_i) = 0.02 + 1.17 \cdot \text{Browsed}_i \quad (3.55)$$

It is important to realize that these beta parameters, or regression coefficients, have now been corrected for detection so this equation can be used to describe and predict the probability of weta being present at other plots without further regard to detection issues. For this particular example there is only a single indicator variable included in the model, but the principle holds true for more complex situations. The estimated probability of occupancy for each surveyed plot is given as part of the output (starting from immediately below the table of beta parameters), and here those values are either 0.77 or 0.50 for plots that were browsed or unbrowsed by goats respectively. Determining those values come from the following calculations. Firstly, for an unbrowsed plot ($\text{Browsed}_i = 0$):

$$\begin{aligned} \text{logit}(\psi_i) &= 0.02 + 1.17 \cdot 0 \\ &= 0.02 \\ \psi_i &= \frac{e^{0.02}}{1 + e^{0.02}} \\ &= 0.50 \end{aligned}$$

and for a browsed plot ($\text{Browsed}_i = 1$):

$$\begin{aligned} \text{logit}(\psi_i) &= 0.02 + 1.17 \cdot 1 \\ &= 1.19 \\ \psi_i &= \frac{e^{1.19}}{1 + e^{1.19}} \\ &= 0.77 \end{aligned}$$

These values indicate that the probability of weta being present at a randomly selected plot would be 0.50 if the plot was unbrowsed and 0.77 if it was browsed. As was done previously, the effect of goat browsing on weta occupancy can also be interpreted in terms of an odds ratio, which may be preferable when more covariates are included in a model. That is:

$$\begin{aligned}\widehat{OR}_{Browsed} &= e^{\widehat{a2}} \\ &= e^{1.17} \\ &= 3.22\end{aligned}$$

which means that for every plot where weta were absent, weta would be present in 3.22 times more plots that had been browsed than had not been browsed.

Inserting the estimated beta parameters for detection yields a series of equations that can be used to describe and predict detection probabilities for each survey. That is;

$$\begin{aligned}\text{logit}(p_{i,1}) &= -1.30 + 0.73 \cdot \text{Obs2}_{i,1} + 1.07 \cdot \text{Obs3}_{i,1} \\ \text{logit}(p_{i,2}) &= -1.45 + 0.73 \cdot \text{Obs2}_{i,2} + 1.07 \cdot \text{Obs3}_{i,2} \\ \text{logit}(p_{i,3}) &= -2.24 + 0.73 \cdot \text{Obs2}_{i,3} + 1.07 \cdot \text{Obs3}_{i,3} \\ \text{logit}(p_{i,4}) &= -1.37 + 0.73 \cdot \text{Obs2}_{i,4} + 1.07 \cdot \text{Obs3}_{i,4} \\ \text{logit}(p_{i,5}) &= -0.34 + 0.73 \cdot \text{Obs2}_{i,5} + 1.07 \cdot \text{Obs3}_{i,5}\end{aligned}$$

That the coefficients for $\text{Obs2}_{i,j}$ and $\text{Obs3}_{i,j}$ are both positive indicates the detection probabilities for both observers 2 and 3 are greater than the probability of detection for observer 1.

Table 3.1: Estimated detection probabilities for each observer on each day for the Mahoenui weta example

Observer	Day 1	Day 2	Day 3	Day 4	Day 5
1	0.21	0.19	0.10	0.20	0.43
2	0.36	0.33	0.18	0.34	0.61
3	0.44	0.41	0.24	0.43	0.69

Table 3.1 contains the estimated detection probabilities for each observer on each day, obtained by inserting the appropriate covariate values into the above equations and converting the results from the logit to probability scale. According to this model, there is a good deal of daily variation in detection probability and substantial differences between observers. Note also that even though the detection component of the model was defined as an additive effect between day and observer (i.e., $p(\text{Day}+\text{Obs})$, with a consistent difference between each observer each day), that was on the logit-scale and once converted to the probability scale, that effect is no longer consistent. For example, on day 3, the difference in detection probabilities for observer 1 and 3 is 0.14, but on day 5 that difference is 0.26. The inconsistent difference on the probability scale is a consequence of the non-linear transformation and occurs whenever

the logit-link function is used (e.g., logistic regression), and it is for this reason why interpreting effects in terms of the odds-ratio is recommended in general. For example, the odds-ratio for detection for the observer 3 compared to observer 1 can be calculated as:

$$\begin{aligned}\widehat{OR}_{Obs3} &= e^{\widehat{b7}} \\ &= e^{1.07} \\ &= 2.92\end{aligned}$$

which means that for every survey where weta are not detected at an occupied plot, observer 3 will detect weta in (approximately) 3 times as many surveys than observer 1. An approximate 95% confidence interval for the odd-ratio would be:

$$\begin{aligned}&= \left(e^{\widehat{b7}-2 \cdot SE(\widehat{b7})}, e^{\widehat{b7}+2 \cdot SE(\widehat{b7})} \right) \\ &= \left(e^{1.07-2 \cdot 0.46}, e^{1.07+2 \cdot 0.46} \right) \\ &= \left(e^{0.15}, e^{1.99} \right) \\ &= (1.16, 7.32)\end{aligned}$$

3.4 Example 3: Multiple and non-linear covariates, and producing species occurrence maps

This final single-season example for PRESENCE highlights a number of points; that PRESENCE is not restricted to fitting univariate models, covariates maybe continuous as well as categorical and that the results from PRESENCE can be used to produce species distribution maps. The associated spreadsheet for this example is also a useful reference for some of the steps necessary for preparing the data prior to entering it into PRESENCE. The data itself is fictitious but was simulated to replicate a real-world example.

3.4.1 Preparing the data

From the PRESENCE sample data folder, open the spreadsheet **Single-season example.xls** and go to the **Study Sites** sheet. Detection/nondetection data has been collected for 148 units, each surveyed four times. **Habitat Type** and **Elevation** for each unit has been recorded (these are site-specific covariates), and further to the right of the page, note that the time of the survey has also been recorded, which is going to be used as a sampling-occasion covariate (**Time of Day**). Also on the sheet are transformations of the covariates that have to be made prior to entering the data into PRESENCE.

Habitat Type is a categorical covariate with three levels which has been recorded as either type A, B or C. As noted earlier, this must be entered into PRESENCE as a series of indicator variables, in this case as the variables **HabA**, **HabB** and **HabC** which equal 1 if the unit was of the respective type, and 0 if the unit was one of the other habitat types. Calculation of the indicator variables was done using the `if` formula.

Elevation is a continuous covariate, and while not strictly necessary, it is often advisable to standardize continuous covariates particularly those whose values tend to be a long way from 0. Here, elevation has been standardized by subtracting 1000 (which is close to the mean elevation), then dividing the result by 100. That is, standardized elevation (Ele_i) has been calculated as:

$$Ele_i = \frac{E_i - 1000}{100} \quad (3.56)$$

where E_i is the actual elevation recorded at each unit.

Additional site-specific covariates have also been defined to enable an interaction between the **Habitat Type** and **Elevation** covariates. An interaction might be considered if it is thought that the effect of elevation on occupancy may be different in different habitat types. To represent this interaction term then a series of covariates have to be defined that are simply the product of the habitat indicator variables and the standardized elevation covariate, hence this results in three additional covariates being calculated (labeled **HabA.Ele**, **HabB.Ele** and **HabC.Ele**). These interaction covariates have a similar form to the indicator variables, although instead of having values that are either 0 or 1, the values are either 0 or the standardized elevation. Note that the interaction terms are not separately standardized as then the necessary relationships between the covariates would be lost. Take some time to examine spreadsheet and understand how the site-specific covariates to entered into PRESENCE have been calculated.

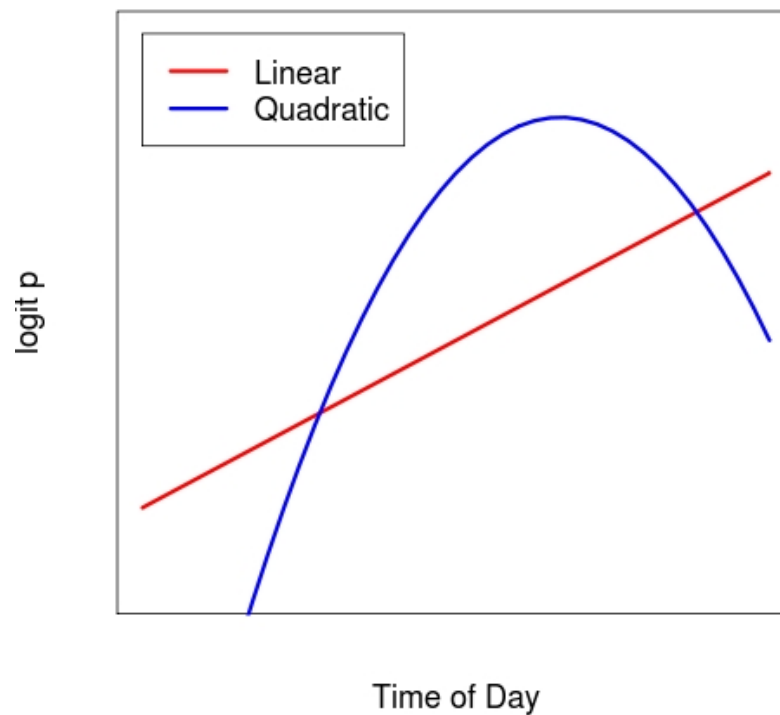


Figure 3.25: Linear and quadratic relationships between Time of Day and logit-detection.

Time of Day (ToD) for each survey was also recorded as it was expected that the detectability of the species would change during the day due to activity patterns. Furthermore, not only was it expected that detectability would change during the day, but that it would peak at a certain point when the species is most active and be lower at other times. Hence, using only ToD as a covariate would not be sufficient as that would assume a linear relationship between ToD and detection, i.e., detection as gradually increasing or decreasing throughout the day (Figure 3.25). One approach to allow the situation envisaged would be to use a quadratic relationship between detection and ToD day (Figure 3.25) which can be accomplished by defining a new covariate that equals ToD^2 . That is, the detection component that will be fit to the data will be of the form:

$$\text{logit}(p_{i,j}) = b_1 + b_2 \cdot ToD + b_3 \cdot ToD^2, \quad (3.57)$$

by doing so, we are incorporating a non-linear relationship between the probability and covariate of interest. Note that as the quadratic relationship has been drawn (concave down), b_3 would have to be a negative value. The quadratic relationship could also be concave up when b_3 is positive. However, by including the ToD^2 values as a covariate in a model, PRESENCE will estimate a value for b_3 from the data. Note that one property of a quadratic relationship is that it is symmetric about the maximum (or minimum). Also note that when converted to the probability scale, the quadratic relationship can appear more like a bell-curve (Figure 3.26). Other non-linear relationships with covariates, e.g., cubic, square-root or logarithmic, could be defined in a similar way.

As recorded (minutes since 6am), the ToD covariate has values that are large and tend to be a long way from 0, therefore the values have been standardized in the spreadsheet as hours since 8:30am using the formula:

$$ToD_{ij} = \frac{Time_{ij} - 150}{60} \quad (3.58)$$

where $Time_{ij}$ is the original time of the survey. 8:30am (i.e., 150 minutes after 6am) was used because it was close to the mean time of the surveys. Following the standardization, The standardized ToD has been squared; note that the square covariate has not been re-standardized as the desired relationship with ToD would be lost.

As can be seen from this example, a good deal of forethought needs to happen about the types of models that will be fit to the data in an analysis prior to even starting PRESENCE such that the appropriate covariates can be defined. I would argue this a good point as it forces the analyst to think careful about what it is they are trying to achieve with the data rather than following a more haphazard process. It should be noted, however, that is possible to add additional covariates to a PRESENCE data file even after a project has begun if required. Once the covariates have been added and the data file re-saved, close then restart PRESENCE and open the desired project so that PRESENCE recognizes the new covariates are available.

3.4.2 Entering the data in PRESENCE

Start a new project in PRESENCE and click the button in the lower-right corner to bring up the **Data Input Form** window. Then,

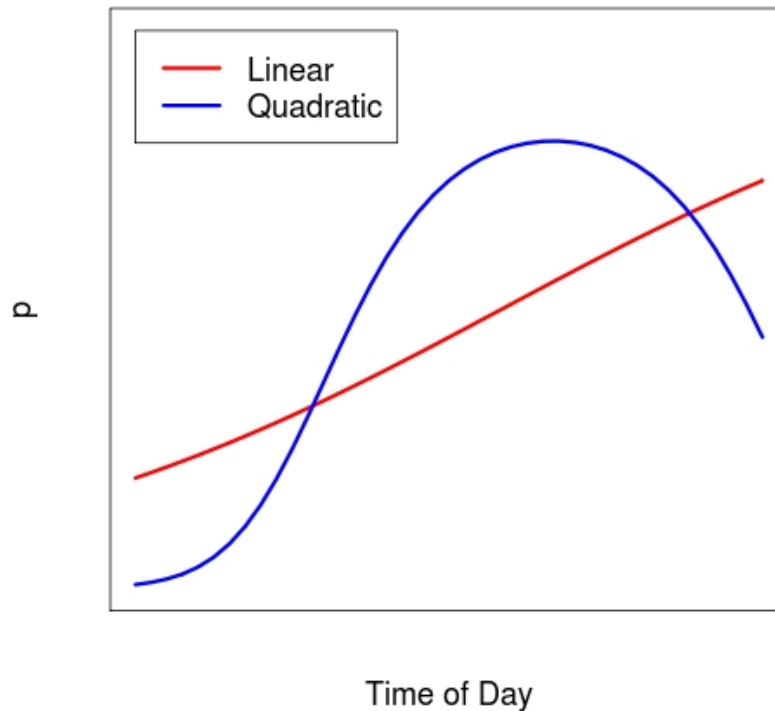


Figure 3.26: Linear and quadratic relationships between Time of Day and logit-detection when converted to the probability scale.

1. Copy and paste the detection data from the spreadsheet (recall that to paste the data, within the **Data Input window** select Edit>Paste>Paste Values). PRESENCE will automatically add the required number of rows and columns.
2. Change the number of site covariates to **7** and number of sampling covariates to **2**.
3. In the spreadsheet, select and copy the 7 site covariates that have been defined (i.e., 3 indicator variables, standardized elevation and 3 interaction terms), including the row of covariate names.
4. Select the **Site Covariate** tab in PRESENCE, make sure the top left-most cell is selected then paste in the covariates along with the covariate names (Edit>Paste>Paste w/covnames) to automatically rename the covariates.
5. Copy the standardized *ToD* sampling occasion covariate from the spreadsheet.
6. Select the **SampCov1** tab in PRESENCE, make sure the top left-most cell is selected and paste in the covariate values (Edit>Paste>Paste Values). Rename the covariate as *ToD* (Edit>Rename covariate).

7. Repeat the previous 2 steps for the *ToD*-squared covariate, renaming it appropriately (e.g., *ToD_sq*).
8. Save the data file with an appropriate name and title, selecting **No** when prompted about whether the last column contains frequency data.
9. Once the data file has been saved, close the **Data Input Form** window, review the information in the **Project Setup Window** for accuracy, and if satisfied, select **OK** to complete the creation of the project.

After completing these steps, an empty **Results Browser** table should be on the screen. If not, you have not yet created the project and should check that all of the above steps have been complete. You should not attempt to setup a model for analysis as it will not run without seeing the empty **Results Browser**.

3.4.3 Fitting a model

Here we are going to fit a single model to the data, and use those results to produce a species distribution map using third-party software, in this case, with the software R.

In this model, occupancy probability shall be function of habitat type and elevation, and detection probability a function of habitat type and time of day. More specifically, the model is to allow the effect of elevation on species occurrence to be different in different habitat types which is achieved by including the interaction terms between these two covariates. Without the interaction terms, we can only fit a model where the effect of elevation is the same in different habitats, which may often be a reasonable option, but is not a model of interest here. Habitat Type A will be used as a standard against which the effect of the other habitat types will be compared, therefore not all of the indicator variables and interaction terms will be required here. The equation representing the model that is to be fit is:

$$\text{logit}(\psi_i) = a1 + a2 \cdot \text{HabB}_i + a3 \cdot \text{HabC}_i + a4 \cdot \text{Ele}_i + a5 \cdot \text{HabB.Ele}_i + a6 \cdot \text{HabC.Ele}_i. \quad (3.59)$$

As with the previous examples, it is the terms associated with each of the regression coefficient that will get entered into the occupancy design matrix in PRESENCE (presuming a '1' for $a1$). For those with some experience with regression and linear-modeling, the derivation of this equation may be relatively straight forward, but for those with less experience it may be beneficial to work through the process. Consider the sketch of the model we want to fit to the data (Figure 3.27), noting that there are 3 lines with different intercepts and slopes, one for each habitat type.

The general formula for any straight line is $y = a + b \cdot x$, where a is the intercept (the value at $x = 0$, where the line crosses the y -axis) and b is the slope of the line (how quickly y changes as x changes). With this in mind, we can write an equation for each of the three lines in Figure 3.27. For habitat A (the red line), the intercept has been labeled as $a1$ and the slope of the line is $a4$, therefore the equation for the line is:

$$\text{logit}(\psi_i) = a1 + a4 \cdot \text{Ele}_i. \quad (3.60)$$

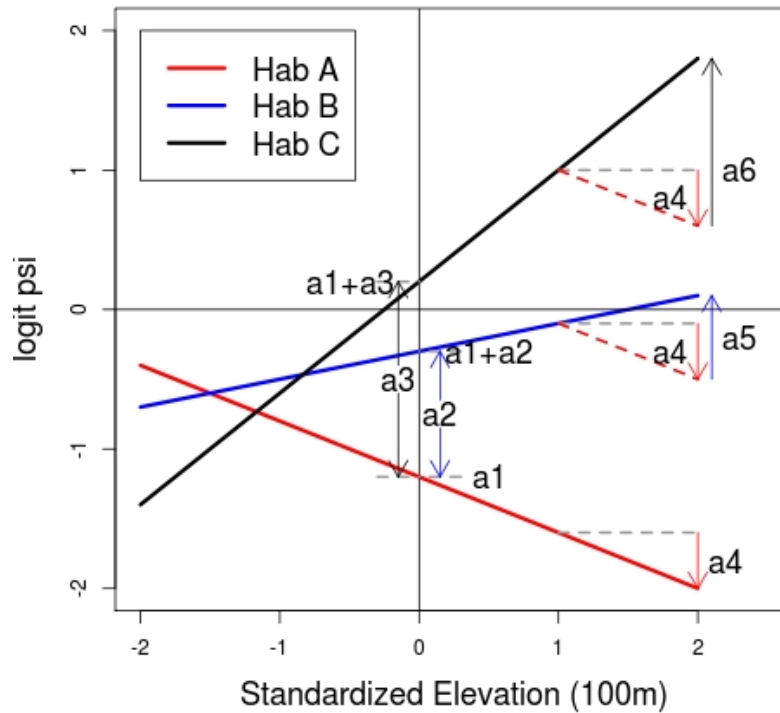


Figure 3.27: Graphical representation of the model for occupancy probability with an interaction between Elevation and Habitat Type.

For habitat B (the blue line), the intercept has been labeled as $a1 + a2$ with $a2$ indicating how different the intercept is for habitat B compared to habitat A. The slope of the line is $a4 + a5$ with $a5$ indicating how different the slope of the line is for habitat B compared to habitat A, i.e., how different is the effect of elevation on occupancy in habitat B compared to habitat A. Therefore, the equation for the blue line is:

$$\text{logit}(\psi_i) = (a1 + a2) + (a4 + a5) \cdot Ele_i. \quad (3.61)$$

Similarly for habitat C, the black line. The intercept has been labeled as $a1 + a3$ with $a3$ indicating how different the intercept is for habitat C is compared to habitat A. The slope of the line is $a4 + a6$ with $a6$ indicating how different the slope of the line is for habitat C compared to habitat A, i.e., how different is the effect of elevation on occupancy in habitat C compared to habitat A. Therefore, the equation for the black line is:

$$\text{logit}(\psi_i) = (a1 + a3) + (a4 + a6) \cdot Ele_i. \quad (3.62)$$

Note that if $a5$ and $a6$ were 0, that would imply that the effect of elevation on occupancy was the same in all three habitats. Also note that the lines as drawn in Figure 3.27 are completely arbitrary and are just to conceptualize the model that is to be fit to the data; the estimated parameters may be quite different from those indicated.

The next step is to combine the three equations into one through the use of the habitat indicator variables and interaction terms. With these terms the appropriate regression coefficients get included where required. Recall that the indicator variables = 1 where the unit is of that type and = 0 otherwise, therefore multiplying the indicator variable by a regression coefficient means that the regression coefficient gets included in the equation for a unit of that type, otherwise it does not. Similarly with the interaction terms, they equal the elevation of the unit (i.e., Ele_i) if the unit was of the respective habitat type, and 0 otherwise. Hence, multiplying the interaction term by a regression coefficient will equal the coefficient times the elevation of the unit if the unit is of that habitat type and 0 otherwise. For example, $a5 \cdot HabB.Ele_i$ will equal $a5 \cdot Ele_i$ if the unit is habitat type B and 0 otherwise. Using these variables, the three equations can be combined into a single equation as in Equation 3.59. To verify the form of the general equation is correct we can substitute in some covariate values to ensure the equations for the individual habitat types are obtained. For example, if a unit is habitat type C, Equation 3.59 becomes:

$$\begin{aligned} \text{logit}(\psi_i) &= a1 + a2 \cdot 0 + a3 \cdot 1 + a4 \cdot Ele_i + a5 \cdot 0 + a6 \cdot Ele_i \\ &= a1 + a3 + a4 \cdot Ele_i + a6 \cdot Ele_i \\ &= (a1 + a3) + (a4 + a6) \cdot Ele_i \end{aligned} \quad (3.63)$$

Based upon Equation 3.59, the required occupancy design matrix is given in Figure 3.28. Once again, note that the entries within the cells are simply the terms associated with the respective regression coefficients from the estimating equation.

	a1	a2	a3	a4	a5	a6
psi	1	HabB	HabC	Ele	HabB.Ele	HabC.Ele

Figure 3.28: Design matrix for occupancy probability

The detection component of the model is to have both a quadratic time of day effect, and a habitat effect. That is, the quadratic time of day effect is intended to represent the assumption that there is a optimal time of day where detection probability is highest, while the effect of habitat type on detection is that detection is consistently different in different habitat types throughout the day. Therefore, the peak in detection happens at the same time of day in all habitats. It will not be done here, but if it was believed that the highest detection probability of achieved at different times of day in different habitats, that would require an interaction between time of day and habitat type, which would have to have defined as separate covariates outside of PRESENCE, then included in the data file.

Figure 3.29 is a sketch of the intended relationship between detection, time of day and habitat type. Note that the resulting parabolic shape of the quadratic relationship with time of day is symmetric, and also that the optimum does not occur where $x = 0$ (i.e., at 8:30am given

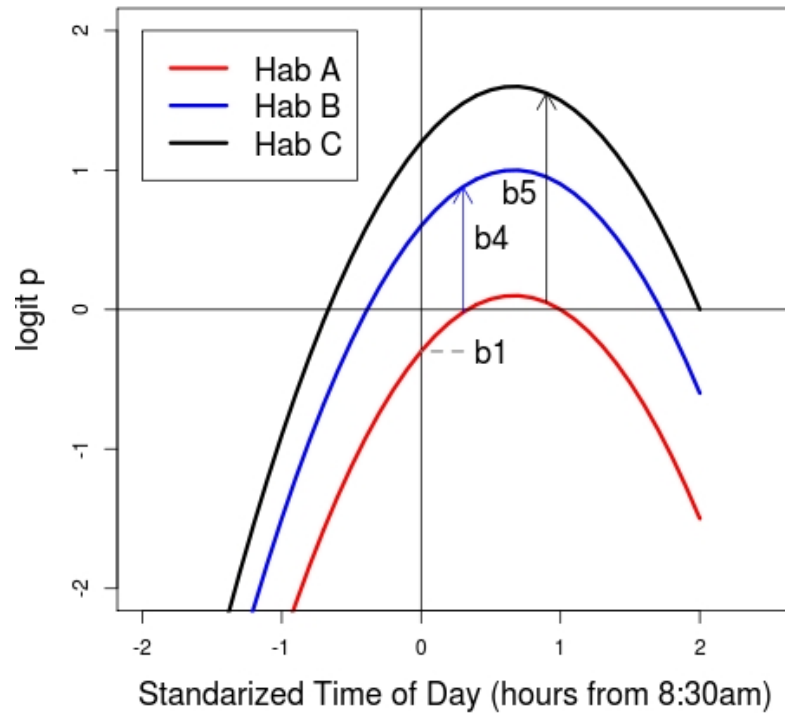


Figure 3.29: Graphical representation of the model for detection probability with a quadratic Time of Day effect and an additive effect of Habitat Type

the standardized covariate). In order to allow this then both ToD and ToD^2 terms are required. That is, the equation of the red line for habitat A could be expressed as:

$$\text{logit}(p_{i,j}) = b1 + b2 \cdot ToD_{i,j} + b3 \cdot ToD_{i,j}^2 \quad (3.64)$$

The equation for habitat B (the blue line) is similar, although it has been shifted up by an amount $b4$. This amount could be combined with the intercept term for the red line, but here will be left separate given the labeling that has been used, i.e.,

$$\text{logit}(p_{i,j}) = b1 + b2 \cdot ToD_{i,j} + b3 \cdot ToD_{i,j}^2 + b4 \quad (3.65)$$

Using the same reasoning, the equation for detection in habitat C would be:

$$\text{logit}(p_{i,j}) = b1 + b2 \cdot ToD_{i,j} + b3 \cdot ToD_{i,j}^2 + b5 \quad (3.66)$$

Using the indicator variables that have been defined for each habitat type, the following general equation for detection can be developed:

$$\text{logit}(p_{i,j}) = b1 + b2 \cdot ToD_{i,j} + b3 \cdot ToD_{i,j}^2 + b4 \cdot HabB_i + b5 \cdot HabC_i \quad (3.67)$$

-	b1	b2	b3	b4	b5
p1	1	ToD	ToD_sq	HabB	HabC
p2	1	ToD	ToD_sq	HabB	HabC
p3	1	ToD	ToD_sq	HabB	HabC
p4	1	ToD	ToD_sq	HabB	HabC

Figure 3.30: Design matrix for detection probability

As there is no additional among-survey variation in detection (beyond that caused by the time of the surveying), the same equation will apply to all four surveys and the four rows of the detection design matrix will be identical (Figure 3.30).

Once the design matrices have been constructed;

1. return to the **Setup Numerical Estimation Run** window
2. provide a meaningful name (**psi(Habitat*Ele),p(ToD_sq+Habitat)**, is suggested)
3. select Print V-C matrix from the list of options on the right-hand side
4. run the model

Here, the option to print the variance-covariance (V-C) matrix has been selected because when the probability of occurrence is predicted for unsurveyed places, a measure of the uncertainty is also desired. To obtain this, the variance-covariance matrix is required to account for any correlation between the estimated beta parameters.

After fitting this model to the data, the estimated beta parameters from the output are given in Table 3.2, which allows the completion of the logistic regression equations for both occupancy and detection probabilities.

$$\begin{aligned} \text{logit}(\psi_i) = & -1.14 + 0.73 \cdot \text{HabB}_i + 2.77 \cdot \text{HabC}_i \\ & + 0.48 \cdot \text{Ele}_i - 0.18 \cdot \text{HabB.Ele}_i + 0.45 \cdot \text{HabC.Ele}_i. \end{aligned} \quad (3.68)$$

From this model, it appears that there is both a habitat and elevational effect on occupancy. The elevational effect may be different in each habitat type (a_5 and a_6 are somewhat different from 0, but do have large standard errors relative to the size of the estimate), and the intercepts also appear to be quite different for each habitat type (consider the estimates for a_2 and a_3). Figure 3.31 presents the estimated relationships between occupancy, elevation and habitat type on both the logit and probability scale respectively. Again, it should be stressed that as the methods used to analyze the data have accounted for imperfect detection, the estimated

Table 3.2: Estimated beta parameters and standard errors (SE) for model $\text{psi}(\text{Habitat} \cdot \text{Ele}), p(\text{ToD_sq} + \text{Habitat})$

Beta	Estimate	SE
a1	-1.138	0.520
a2	0.727	0.593
a3	2.773	0.950
a4	0.483	0.233
a5	-0.179	0.267
a6	0.449	0.390
b1	0.513	0.457
b2	-0.453	0.121
b3	-0.509	0.101
b4	-0.013	0.507
b5	0.187	0.497

regression coefficients have been automatically corrected hence they can be used to make inferences about species occurrence without further regard for detection issues.

$$\text{logit}(p_{i,j}) = 0.51 - 0.45 \cdot \text{ToD}_{i,j} - 0.51 \cdot \text{ToD}_{i,j}^2 - 0.01 \cdot \text{HabB}_i + 0.19 \cdot \text{HabC}_i \quad (3.69)$$

For detection, there does appear to be an optimal time of day where detection is highest as the estimate for $b3$ is negative which gives a concave-down parabolic shape (a positive value for $b3$ would imply a concave-up parabolic shape suggesting detection would be lowest at some point during the day and higher at other times). Habitat-type does not appear to have much of an effect on detection given the small estimates for $b4$ and $b5$. Figure 3.32 present the estimated relationships between detection, time of day and habitat type on both the logit and probability scale respectively. These suggest that detection is highest at approximately 8am.

3.4.4 Producing a species distribution map in R

In this final section the software R shall be used to produce a species distribution map for a landscape based upon the model fit to the data in this example. The basic rationale is that regions with high predicted probabilities of occupancy would represent cores areas of the species distribution where the species is likely to be found more frequently, while regions with low occupancy probabilities would represent the edge of the species distribution. For the construction of this map, it is presumed that the resolution of the map is set such that the size of each pixel is at the same scale as the sampling units from which the data was collected. In fact, if a desired outcome of such an occupancy analysis is to produce a map, the required resolution of the map can be useful to determine an appropriate definition of a sampling unit. It is also important to recognize that in order to be able to make predictions at unsurveyed locations using the estimating equation, the equation can only contain covariates whose values are available for the unsurveyed locations. Therefore, covariates whose values can only be known from physically visiting a location may not be useful for this type of

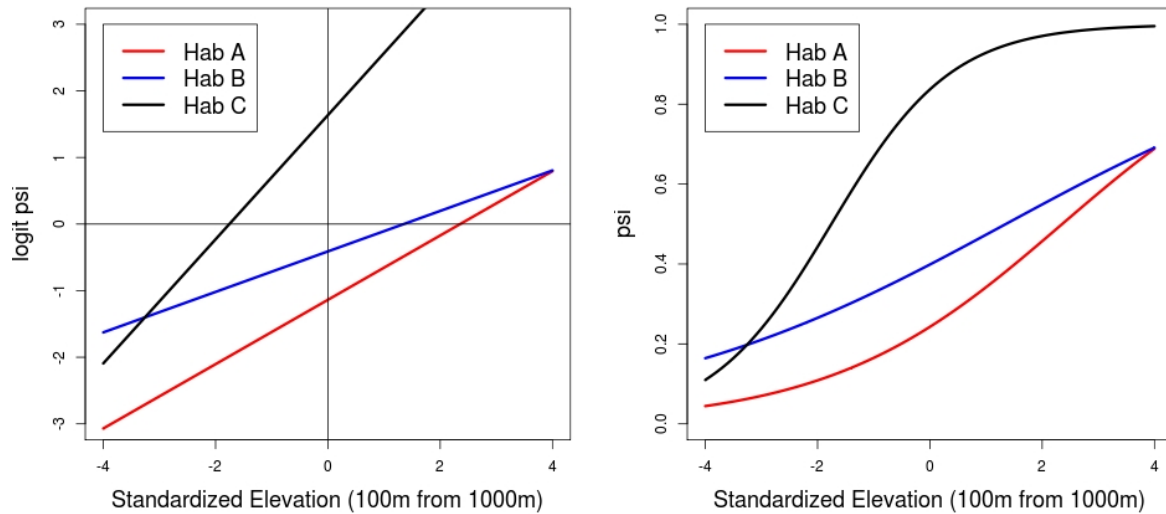


Figure 3.31: Estimated relationship between elevation and habitat type with logit-occupancy (left) and occupancy probability (right).

exercise. This holds for any technique used to develop a species distribution model and is not unique to these methods.

Detailed instructions are not given here as they require some knowledge of R, but the basic procedure is to use the regression equation for occupancy that results from the analysis in PRESENCE to predict the probability of occupancy at a set of locations in the area of interest based upon the available covariate values for each location. The portion of the V-C matrix in the output that relates to the beta parameters associated with occupancy (here, $a1-a6$) is used to calculate a standard error for the predicted occupancy probability, and also limits of a confidence interval; all of which can also be plotted.

In the PRESENCE sample data folder, there is a text file called **Ex3 landscape.txt**. This is a tab-delimited text file that contains the habitat type and elevational information for 1000 locations on a 50×20 grid. The habitat type and elevation variables have been converted to the same set of variables used in the analysis using the same indicator variables, standardization and interactions. The R code to perform the calculations and produce the graphs is provided in the PRESENCE sample data folder, **ex3Rcode.r**. The R code will need slightly modified to specify the location of the PRESENCE sample data folder, otherwise the code can be executed or copy and pasted into R. Once completed, a series of plots should be produced indicating the predicted distribution of the species and measures of uncertainty (Figures 3.33-3.36)

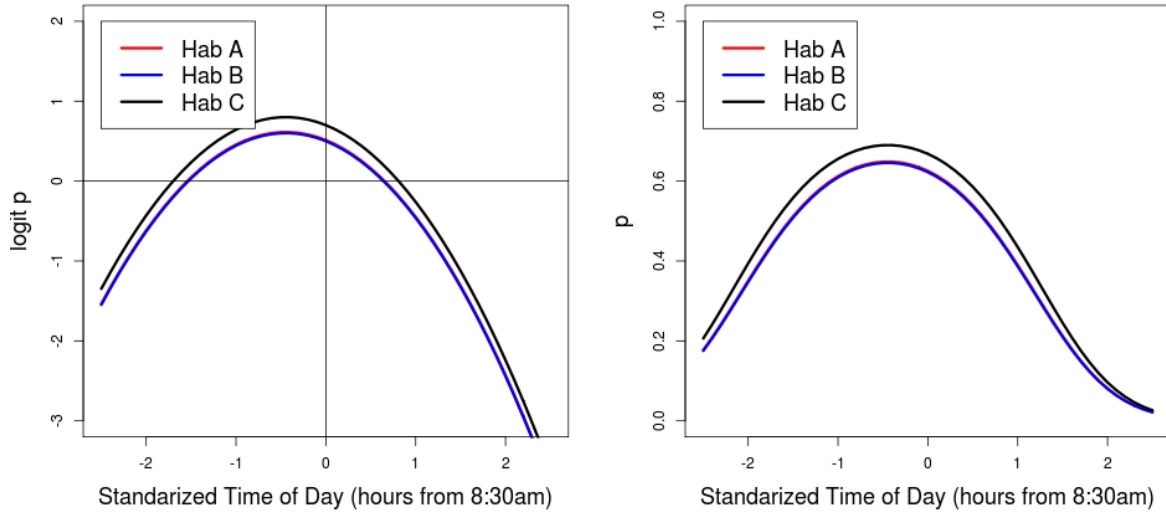


Figure 3.32: Estimated relationship between time of day and habitat type with logit-detection (left) and detection probability (right).

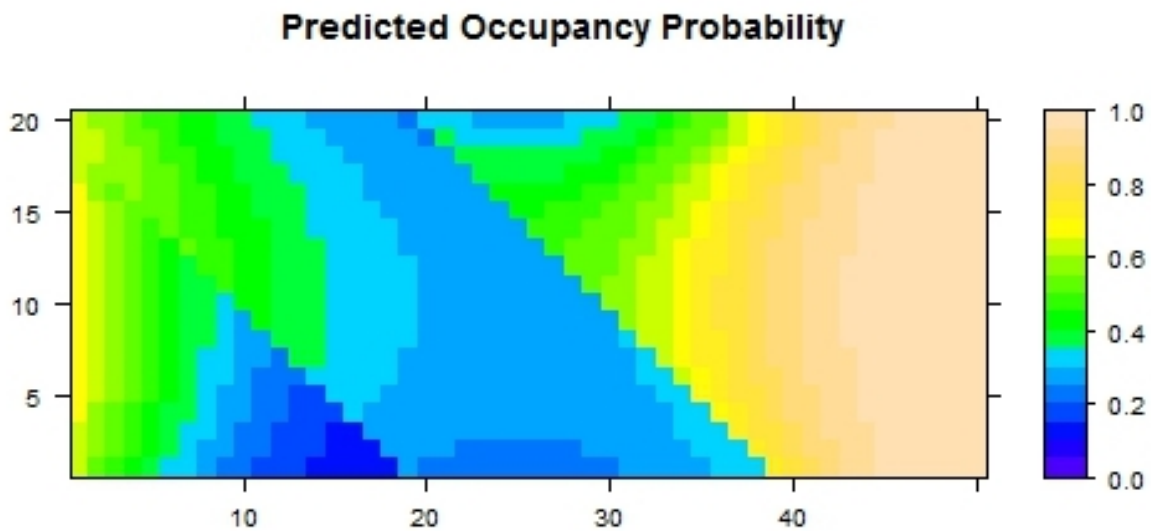


Figure 3.33: Predicted probability of occupancy across a fictitious landscape

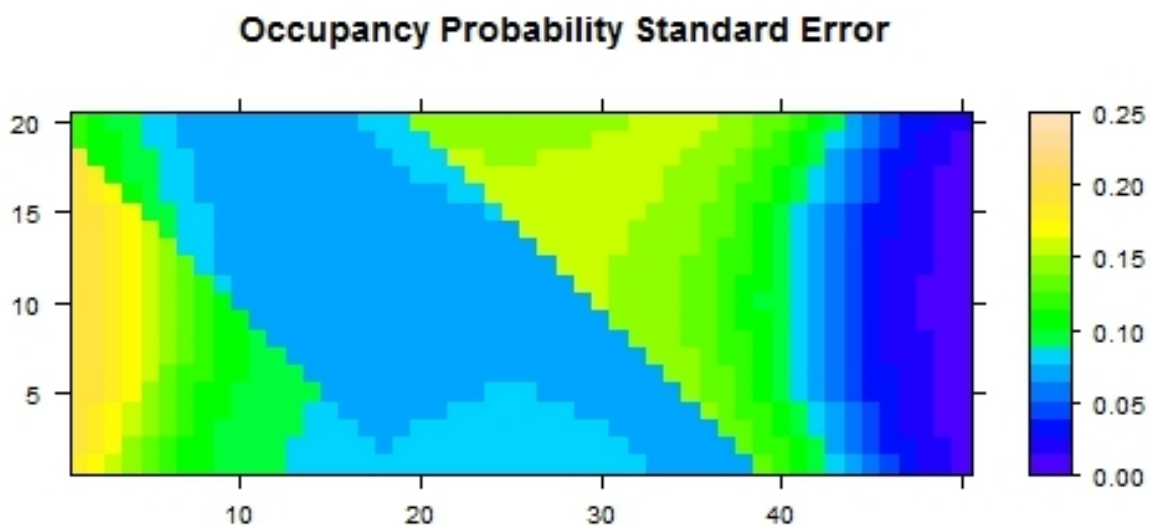


Figure 3.34: Standard error of predicted probability of occupancy across a fictitious landscape

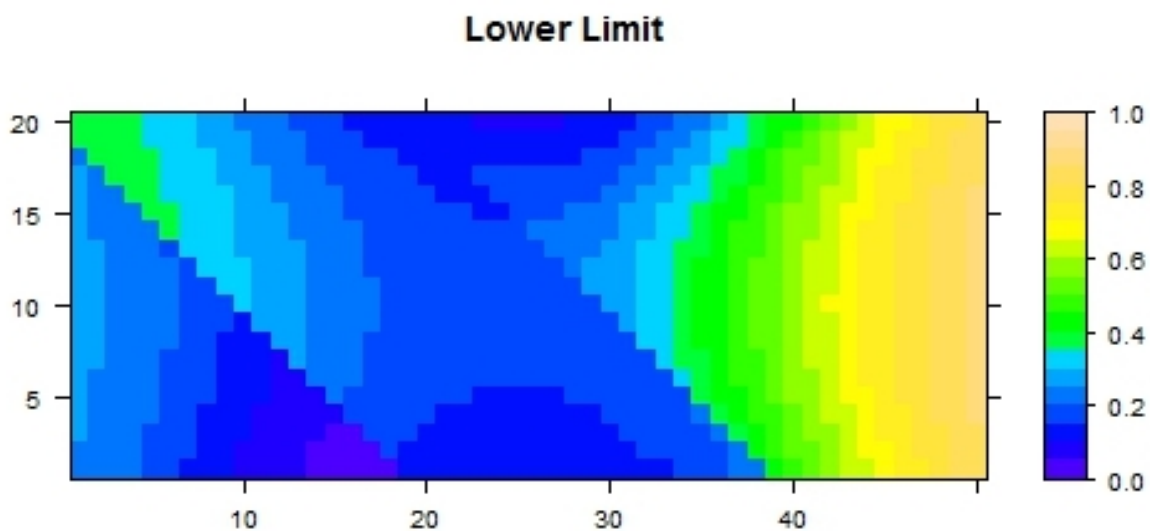


Figure 3.35: Lower limit of a 95% confidence interval on predicted probability of occupancy across a fictitious landscape

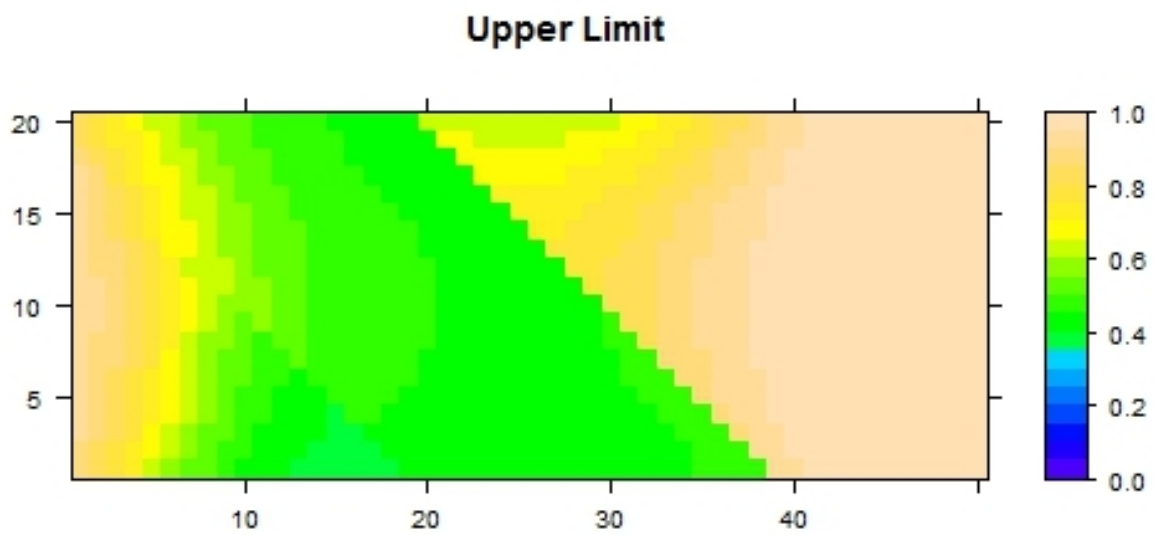


Figure 3.36: Upper limit of a 95% confidence interval on predicted probability of occupancy across a fictitious landscape

Chapter 4

Multi-season model

4.1 Basic modeling approaches

When data has been collected for more than one time period, or sampling season, then questions about changes in occupancy or species distributions naturally occur. There are two general approaches that could be taken here. The first is to simply regard the multi-season data as a series of single-season data sets and apply the single-season model to each season of data either individually or through a combined analysis. By doing so, one could address questions such as whether the pattern in occupancy each season is changing or examine the data for overall effects such as a trend in occupancy. Such an approach may be suitable for some types of problems. A second approach is to consider changes in occupancy through time at the scale of individual sampling units (Figure 4.1), i.e., whether the species is present or absent at a location at a particularly point in time and how that changes through time.

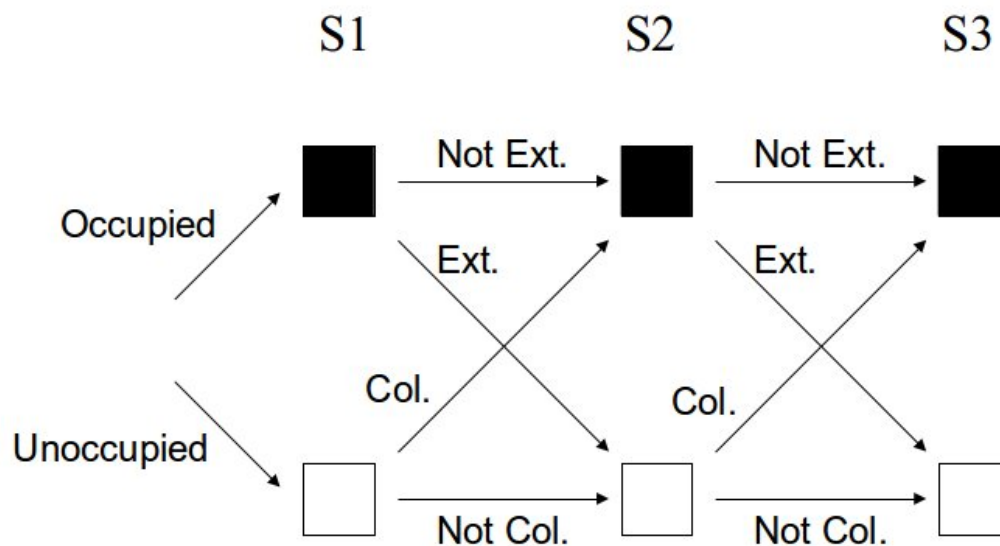


Figure 4.1: Schematic of unit-level changes in occupancy through time.

Using this alternative approach, rather than modeling changes in occupancy patterns, the underlying processes of change in occupancy are modeled which will likely yield much more interesting and useful insights. While there are a variety of terms that could be used, following MacKenzie et al. (2003) and MacKenzie et al. (2006) the processes shall be described in terms of local colonization and local extinction probabilities. These could be defined as:

γ_t : probability a unit is colonized by the species, the unit goes from unoccupied to occupied, between seasons t and $t + 1$.

ε_t : probability the species goes locally extinct from a unit, the unit goes from occupancy to unoccupied, between seasons t and $t + 1$.

Note that these processes are between season events, i.e., changes occur between the main survey periods; within a season, the species is presumed to be either present or absent at a unit for the duration of the surveying, i.e., units are *closed* to changes in occupancy. In addition to these dynamic occupancy probabilities, an initial occupancy probability must also be defined for the first season, e.g., ψ_1 . In combination, these occupancy-related parameters can be used to fully describe the changes in unit-level occupancy.

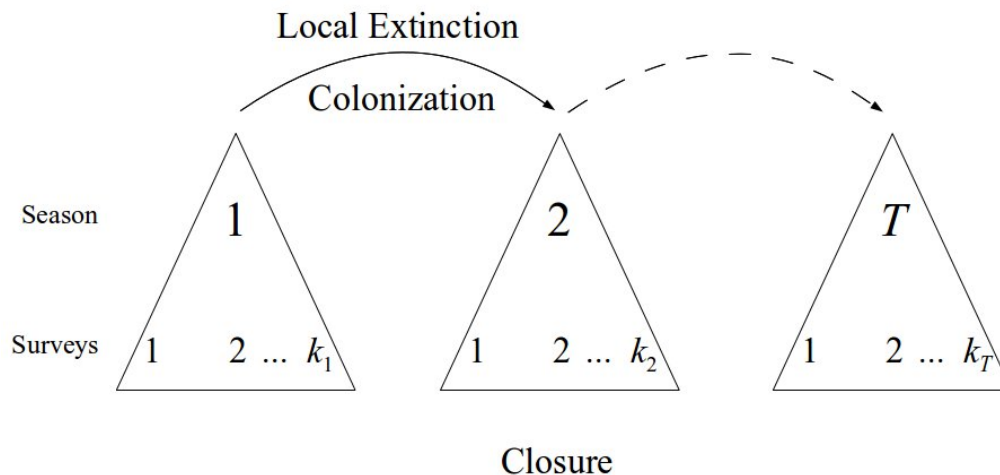


Figure 4.2: General sampling framework for multi-season occupancy model. Changes in occupancy happen between seasons, and repeat surveys are conducted within seasons when units are closed to changes in occupancy.

However, nondetection of the target species also creates problems in terms of colonization and extinction events. For example, if the species was found at a location in one year, but not found the next, has it really gone locally extinct from that place? Dealing with imperfect detection in a multi-season setting is similar to the single-season approach and requires repeat surveys within each season. The general sampling framework is given in Figure 4.2. In terms of the modeling, the basic approach is to again develop probability statements that include all possible options for what may have actually happened based upon the observed data. For example, consider the follow detection history:

$$h_i = 11\ 00\ 01,$$

which represents a location that was surveyed for three seasons, with two surveys per season. The species was detected in both surveys of the first season, not detected in either of the surveys in season 2 and only detected in the second survey of season 3. Because of imperfect detection, there is ambiguity about whether the species was present at the unit in season 2 or not. A verbal description for what may have occurred would be:

*The species was present at the unit in season 1 and detected in both surveys. Between seasons 1 and 2, the species either went locally extinct (so it was absent and couldn't be detected in season 2) then colonized the unit between seasons 2 and 3 (so it was present at the unit immediately before the first survey of the third season), **OR** the species did not go locally extinct between seasons 1 and 2 (so it was still present there), and was simply not detected in either survey of season 2, and did not locally extinct between seasons 2 and 3. In season 3, the species was not detected in the first survey, but was detected in the second survey.*

Note that with the verbal description, the species is presumed to be present in both seasons 1 and 3 as, from the observed data, the species was detected at least once in those seasons. The ambiguity is whether the species was actually present or absent in the second season (when it was never detected) and the verbal description is just working through the different possibilities for what may have happened in terms of colonization and extinction events for each of those cases. Translating the verbal description into a probability statement is just a matter of substituting in the respective parameters for the appropriate phrases (and noting that the probability of something not happening is $1 -$ the probability of it happening). Therefore, the probability statement for this observed detection history becomes:

$$\begin{aligned} \Pr(h_i) = & \psi_1 p_{1,1} p_{1,2} \\ & \times [\varepsilon_1 \gamma_2 + (1 - \varepsilon_1)(1 - p_{2,1})(1 - p_{2,2})(1 - \varepsilon_2)] \\ & \times (1 - p_{3,1}) p_{3,2} \end{aligned} \quad (4.1)$$

The term in the square brackets addresses the ambiguity with respect to the presence or absence of the species at the unit in the second season. As, from the observed data, it is not possible to discount either of the two possibilities, the respective probabilities are added together in the probability statement.

After constructing the probability statements for each surveyed unit, these are combined to form the likelihood equation which is then maximized, as in the single season model, to obtain the maximum likelihood estimates (MLE's) for the model parameters. For further discussion on the development of the model, see MacKenzie et al. (2003) and MacKenzie et al. (2006).

4.2 Reparameterizations

It is important to understand how the occupancy, colonization and extinction probabilities are associated with one another to gain a full appreciation of what is possible with the multi-season

framework, and also as this leads to possible reparametrizations of the model that are available in PRESENCE (but not explained in detail in the current User Manual). By reparameterizing the model, different types of questions can be addressed with the modeling, depending up on the questions of interest.

Note that from first year occupancy (ψ_1), colonization (γ_i) and extinction probabilities (ε_i), occupancy in future years (ψ_{t+1}) can be calculated with the recursive equation:

$$\psi_{t+1} = \psi_t(1 - \varepsilon_t) + (1 - \psi_t)\gamma_t \quad (4.2)$$

Therefore, given values for ψ_1 , ε_1 and γ_1 , the value for ψ_2 can be calculated; with ψ_2 , ε_2 and γ_2 , ψ_3 can be calculated, and so on. Hence, from the initial formulation of the model, overall occupancy in the subsequent seasons can be calculated. As future occupancy has not been estimated directly in the model, it is called a *derived parameter*, much like the conditional probability of occupancy was with the single season model. With this recursive equation, if any 3 of the 4 values values associated with it are supplied, the fourth can always be obtained with a bit of high-school algebra.

This means the recursive equation could be rearranged with either ε_t or γ_t on the left-hand side such that the parameters on the right-hand side, which are the ones to be directly estimated, include both ψ_t and ψ_{t+1} . Hence, rather than estimating first-season occupancy, colonization and extinction probabilities directly, it is possible to estimate seasonal occupancy directly along with either colonization or extinction probabilities. This could be useful when covariate effects on seasonal occupancy are of interest, or if there are questions about trends in occupancy over time.

Another reparameterization is that rather than talking about local extinction probabilities (i.e., the probability of the species becoming absent from a unit), persistence probabilities may be of interest. That is, the probability of the species remaining to be present at a unit between successive seasons. This could simply be defined as:

$$\phi_t = 1 - \varepsilon_t \quad (4.3)$$

It should be stressed that colonization and persistence probabilities are occupancy probabilities; they are just occupancy probabilities conditional upon the presence or absence of the species in the previous season. Cast in this light, one can make valid inferences about factors affecting occupancy based upon the colonization and persistence probabilities, the population of interest is just being subsetted into those places that were unoccupied and occupied in the preceding season.

4.3 Example: Grand Skinks

The grand skink (*Oligosoma grande*) is an endangered giant skink species endemic to New Zealand and presently only found in a few locations within the Otago province. It grows to approximately 300mm in length and gives birth to 2-3 live, fully-formed young each year. They are omnivorous, diurnal and do not hibernate during the winter. They are mostly found

in upland grassland areas on large rocky outcrops. The dataset that will be used here was collected by the New Zealand Department of Conservation near the Macreas township in eastern Otago over a five-year period in the 1990's.



Figure 4.3: Grand skink, *Oligosoma grande*. Photo credit: Catherine Roughton, University of Otago.

During the five years, data was collected from 338 rocky-outcrops. Each outcrop is considered here as a sampling unit, hence the presence and absence of grand skinks on the outcrops will be modeled. Each outcrop was surrounded by either a native grassland, tussock, or a modified habitat where the native grassland had been converted to pasture. The surrounding habitat type is a site-specific covariate and will be used in the analysis. In this example, we shall work thorough the steps of getting multi-season data into PRESENCE, then fit a range of models with and without covariates and examine the output.

4.3.1 Getting the data into PRESENCE

Open the file **Grand Skinks.xls**, located in the PRESENCE sample data folder. The data is arranged with one row for each outcrop and 15 columns; there are 5 years of data with up to 3 surveys of each outcrop per year, although note there are a number of missing values and some outcrops did not get surveyed at all in some years. The first survey of all outcrops each year are aligned in the same column, so the 15 columns are in 5 blocks of 3. In this example, the columns just represent the first, second and third survey of each outcrop and the surveys do not align chronologically. That is, some outcrops may have already been surveyed three times during a year before others had been surveyed for the first time. As such, fitting a model with a survey-specific detection probability would not make biological sense as the first survey of different outcrops may have occurred at quite different times of the year. If there was a desire to allow to allow detection to vary within each year that could easily be achieved by defining suitable sampling-occasion covariates based up survey date, but that is not done

here. There is also a pair of site-specific covariates called Tussock and Pasture which =1 if the surrounding habitat was of that type, and 0 otherwise.

Perform the following steps to enter the data into PRESENCE. If you already have been using PRESENCE for a previous analysis it is always advisable to start a new instance of PRESENCE before starting a new, or reopening a, project.

1. Start a new project in PRESENCE and click on **Input Data Form** to open a blank data form
2. Go to the spreadsheet, select and copy the detection data
3. Return to the data form and paste in the detection data (Paste>Paste Values)
4. Change the number of site-covariates to 2
5. On the spreadsheet, select and copy the name and values of the two covariates
6. In the data input form, select the **Site Covars** tab, ensure the top-leftmost cell is selected, then paste in the covariate names and values (Paste>Paste w/covnames)
7. Change the number of surveys per season (**No. Occ/season**) from 15 to 3
8. Save the data, clicking on **No**, do not save the last column as frequency data and providing a suitable title when prompted
9. Once saved, close the **Data Input Form**, review the data summary for accuracy (especially ensure the number of occasions per season is correct), then click **OK** to complete the project setup
10. When completed, an empty **Results Browser** window should be visible

In this example, each season had the same number of surveys, hence the number of occasions per season could be specified as a single number. When the number of surveys differs, then the number of occasions should be specified as a list of numbers separated by a comma. For example, '3,3,3,2,4' could be entered if there were a maximum of 3 surveys conducted in the first three seasons, then 2 and 4 surveys in the fourth and fifth seasons respectively. The list is just telling PRESENCE how the 15 columns of data should be broken into the different seasons.

4.3.2 Fitting multi-season models without covariates

To begin a multi-season analysis select Run>Analysis: multi-season (approximately half-way down the list of models), which will bring up the window in Figure 4.4. Note that it is a similar format to the single-season analysis, but with a different selection of models in the model box. There are no predefined models, and all models have to be fit using the design matrix. In the model box there are four different parameterizations available; only the first parameterization will be used here. Looking at the design matrix, there are four tabs

in the design matrix window, one for each of the parameter types; occupancy, colonization, extinction and detection. Exploring the design matrix window, noting the following points:

- on the occupancy tab there is only one row for the one real occupancy parameter, ψ_{11} , for first-season occupancy
- on the colonization and extinction tabs, there are four real parameters for colonization and extinction. As these are between season events, and there is 5 years of data, there are 4 between-year periods. The parameter names have been abbreviated to **gam** and **eps** respectively
- on the detection tab there are 15 real parameters, one for each survey occasion. The real parameters are named using the convention $p[t-j]$ where t indicates season and j survey within season

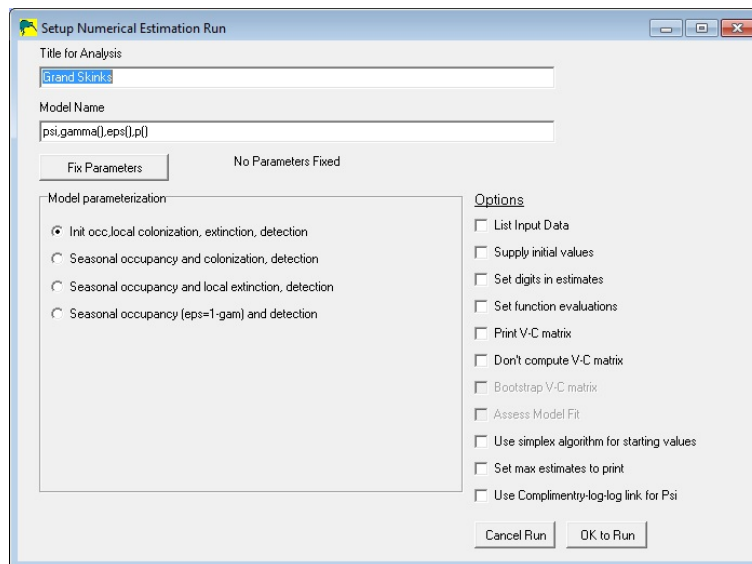


Figure 4.4: Multi-season Setup Numerical Estimation Run window.

The first model to fit is **psi1(.),gamma(.),eps(.),p(.)**, i.e., each of the parameters are constant across time and space. This is the default model when the analysis window is opened, hence no changes to the design matrices are required. Simply rename the model, then press **Ok to Run**. Confirm the model then open the output.

The beginning of the output is very similar to the single-season output, with the main exception being that there are now two additional design matrices, one for colonization (**gamma**) and one for extinction (**eps**) probabilities. These design matrices are treated in exactly the same way as before; the single column of 1's is indicating that these parameters all have the same value in each year. Focusing on the estimation part of the output, first the beta parameters are reported which could be used to write out the corresponding logistic regression equations. This is followed by the real parameter estimates, i.e., the probabilities. As there are no covariates in this particular model, only the value for the first outcrop (site) is reported as the value is assumed to be the same for all outcrops. From this model we would therefore conclude that;

1. the probability of occupancy in the first year was 0.39
2. between all seasons, the probability that skinks colonize a previously unoccupied rocky outcrop is 0.07
3. between all seasons, the probability that skinks go locally extinct from an occupied rocky outcrop is 0.10
4. given an outcrop is occupied by skinks in a year, the probability of detecting skinks in a single survey is 0.69

Below the real parameter estimates, some derived estimates are reported. Occupancy estimates for years 2-5 are calculated using the recursive equation defined previously. Here, the occupancy estimates are slightly increasing. Below, the occupancy estimates are some parameters called **lam**, short for *lambda* (λ). These values are simply the ratio of successive occupancy estimates and have a similar interpretation to a finite population growth rate, e.g.,

$$\lambda_t = \frac{\psi_{t+1}}{\psi_t} \quad (4.4)$$

Looking at the derived parameter estimates, some may question how can overall occupancy increase when the local-extinction probability is greater than the colonization probability? This happens because these dynamic processes only affect occupied and unoccupied units respectively. As initial occupancy is estimated to be <0.5 , there are more unoccupied units that could be colonized compared to the number of occupied units from where the skinks could go locally extinct. Further, the overall fraction of units that go locally extinct between years 1 and 2 can be calculated as $\psi_1 \varepsilon_1 \approx 0.4$, while the overall fraction that get colonized for the same time period is $(1 - \psi_1)\gamma_1 \approx 0.42$, which is slightly larger, leading to the small increase in overall occupancy.

Now, fit the model **psi1(.),gamma(year),eps(year),p(year)**, that is, colonization, extinction and detection probabilities are allowed to be different in each year. After opening the multi-season analysis window, set up the following design matrices:

1. leave the occupancy design matrix as the default
2. select the colonization tab, then select from the design matrix window menu `Init>Full Identity`
3. repeat the above for extinction probability
4. select the detection tab, then select from the design matrix window menu `Init>Seasonal Effects`

Screen shots of the four design matrices required are given in Figure 4.5. As with the single-season models, these design matrices are representing a series of logistic regression

Figure 4.5: Design matrices for the model $\psi(\cdot)$, $\gamma(\text{year})$, $\epsilon(\text{year})$, $p(\text{year})$.

equations. For example, for colonization probability:

$$\text{logit}(\gamma_{i,1}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{0} = b1 \quad (4.5)$$

$$\text{logit}(\gamma_{i,2}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{1} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{0} = b2 \quad (4.6)$$

$$\text{logit}(\gamma_{i,3}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{1} + b4 \cdot \mathbf{0} = b3 \quad (4.7)$$

$$\text{logit}(\gamma_{i,4}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{1} = b4 \quad (4.8)$$

That is, the result of each equation will be a different amount, therefore the probability in each year is allowed to be different. The design matrix that has been defined for detection probability allows detection probability to be different between years, but forces it to be the same within years. Recall that the surveys within years are not aligned in chronological order so different outcrops may have been surveyed for the first time in a year on quite different dates, hence a model that allows detection probability at the first, second and third surveys each year to be different from one another (but the same across all outcrops each survey) makes little biological sense. Had a model been required where each of the 15 surveys were to be allowed a different detection probability then that could have been achieved by selecting `Init>Full Identity`. Note the `Seasonal effects` option only becomes enabled when the detection tab is selected for multi-season models.

Once the design matrices have been set up, provide a model name then run the model. After confirming the results, open the output and locate the real parameter estimates. These should indicate that the estimated probability of;

1. occupancy in the first year is 0.38
2. colonization is 0.12, 0.01, 0.07 and 0.10 for each of the between year periods respectively
3. extinction is 0.07, 0.07, 0.14 and 0.17 for each of the between year periods respectively
4. detecting the skinks in a single survey is 0.70, 0.65, 0.69, 0.84 and 0.66 in each of the 5 years respectively

That the model $\mathbf{psil}(\cdot), \mathbf{gamma}(\mathbf{year}), \mathbf{eps}(\mathbf{year}), \mathbf{p}(\mathbf{year})$ is ranked above the simpler model suggests that there is some important annual variation in at least one of those three parameters that is much better explained by the current model. Which of the parameter types is reasonable for this result could be identified by fitting further models that include annual variation in some of those parameters, but not others, then comparing the results. In actual fact, if identification of which parameters exhibited substantial annual variation and which were relatively constant across the 5-year period was an intended outcome of the analysis then the best approach would be to define a set of candidate models representing different combinations of the questions of interest for each parameter type.

4.3.3 Fitting multi-season models with covariates

Defining the models

Incorporating potential covariates into a multi-season analysis proceeds exactly as for a single-season analysis. The most important step is to ensure that the relevant variables of interest have been defined appropriately and included in the PRESENCE data file. For example, if the question of interest was whether occupancy in the first year was different for those outcrops surrounded by pasture vs tussock (i.e., the modified vs natural habitat), then the model that represents the following equation should be fit to the data:

$$\text{logit}(\psi_{i,1}) = a1 + a2 \cdot \text{Pasture}_i \quad (4.9)$$

where Pasture_i is the indicator covariate defined in the data file which = 1 if the outcrop was surrounded by pasture or = 0 if it was surrounded by tussock. Substituting in these values it should be apparent that for an outcrop surrounded by pasture Equation 4.9 becomes:

$$\text{logit}(\psi_{i,1}) = a1 + a2, \quad (4.10)$$

and for an outcrop surrounded by tussock:

$$\text{logit}(\psi_{i,1}) = a1. \quad (4.11)$$

Therefore, $a2$ indicates how different (on the logit scale) occupancy is for outcrops surrounded by pasture compared to tussock; a negative value would mean occupancy is lower while a positive value would indicate it is higher. Modifying Equation 4.9 slightly highlights the values that need to be included in the design matrix, i.e.,

$$\text{logit}(\psi_{i,1}) = a1 \cdot \mathbf{1} + a2 \cdot \mathbf{Pasture}_i \quad (4.12)$$

so inserting the values that are associated with the regression coefficients (the beta parameters $a1$ and $a2$), the design matrix should appear as in Figure 4.6.

The approach is generalized for the other parameter types that involve more real parameters (rows in the design matrices). For example, suppose that a pasture effect was to be included for the colonization probabilities (i.e., probability of skinks colonizing an outcrop was different if the outcrop was surrounded by pasture compared to tussock), in addition to

	a1	a2
-		
psi1	1	Pasture

Figure 4.6: Design matrix for including a pasture effect on first-year occupancy.

annual variation. At this point, it will be assumed that the effect of being surrounded by pasture is the same each year. In this case, the effect of the variables year and pasture are *additive* and this component could be notated as $\mathbf{gamma}(\mathbf{year}+\mathbf{pasture})$. Building on the set of equations given previously for a year effect on colonization probabilities, the set for this model would be;

$$\text{logit}(\gamma_{i,1}) = b1 + b5 \cdot \text{Pasture}_i \quad (4.13)$$

$$\text{logit}(\gamma_{i,2}) = b2 + b5 \cdot \text{Pasture}_i \quad (4.14)$$

$$\text{logit}(\gamma_{i,3}) = b3 + b5 \cdot \text{Pasture}_i \quad (4.15)$$

$$\text{logit}(\gamma_{i,4}) = b4 + b5 \cdot \text{Pasture}_i \quad (4.16)$$

so $b1 - b4$ allow the annual variation, and $b5$ is the effect pasture has on colonization each year. Expanding this set of equations out to identify the design matrix gives;

$$\text{logit}(\gamma_{i,1}) = b1 \cdot \mathbf{1} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{0} + b5 \cdot \mathbf{Pasture}_i \quad (4.17)$$

$$\text{logit}(\gamma_{i,2}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{1} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{0} + b5 \cdot \mathbf{Pasture}_i \quad (4.18)$$

$$\text{logit}(\gamma_{i,3}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{1} + b4 \cdot \mathbf{0} + b5 \cdot \mathbf{Pasture}_i \quad (4.19)$$

$$\text{logit}(\gamma_{i,4}) = b1 \cdot \mathbf{0} + b2 \cdot \mathbf{0} + b3 \cdot \mathbf{0} + b4 \cdot \mathbf{1} + b5 \cdot \mathbf{Pasture}_i \quad (4.20)$$

Note that each equation contains all beta parameters, although many are multiplied by 0 to indicate if they are not required for that particular equation. This may seem long-winded, but ultimately provides a great deal of flexibility for the types of models that can be fit to the data. The design matrix for the colonization component of this multi-season model is in Figure 4.7.

What if the effect of the variable of interest was thought to be different each year, e.g., sometimes, under certain environmental conditions, skinks on the outcrops surrounded by pasture did better than those surrounded by tussock, but in other years that pattern was reversed (or at least the magnitude of any effect was different)? The effect of the variables year and pasture would now be called *multiplicative*, and there is an interaction between these two variables. To fit a model with such an interaction between year and pasture on extinction probabilities, a design matrix is needed that represents the following equations;

$$\text{logit}(\epsilon_{i,1}) = c1 + c5 \cdot \text{Pasture}_i \quad (4.21)$$

$$\text{logit}(\epsilon_{i,2}) = c2 + c6 \cdot \text{Pasture}_i \quad (4.22)$$

$$\text{logit}(\epsilon_{i,3}) = c3 + c7 \cdot \text{Pasture}_i \quad (4.23)$$

$$\text{logit}(\epsilon_{i,4}) = c4 + c8 \cdot \text{Pasture}_i \quad (4.24)$$

	b1	b2	b3	b4	b5	
-						
qam1	1	0	0	0		Pasture
qam2	0	1	0	0		Pasture
qam3	0	0	1	0		Pasture
qam4	0	0	0	1		Pasture

Figure 4.7: Design matrix for including a consistent pasture effect on colonization probability with additive annual variation.

Here, $c1-c4$ provides for an overall year effect and $c5-c8$ indicates how different extinction probabilities are between outcrops surrounded pasture vs tussock each between year period. After expanding the equations out, the required design matrix is given in Figure 4.8.

	c1	c2	c3	c4	c5	c6	c7	c8
-								
eps1	1	0	0	0	Pasture	0	0	0
eps2	0	1	0	0	0	Pasture	0	0
eps3	0	0	1	0	0	0	Pasture	0
eps4	0	0	0	1	0	0	0	Pasture

Figure 4.8: Design matrix for including an interaction between year and pasture on extinction probability.

Notating the extinction component of the model, it could be called **eps(Year*Pasture)** with the '*' indicating the multiplicative effect (as opposed to the '+' used when the variables were additive).

While covariates can also be added for detection (either site-specific or sampling-occasion covariates), none shall be used here. However the mechanics of doing so is the same as that used above, and also for the single-season models where covariates were applied to detection probabilities. Here, use a year effect of detection as was done for the previous model fit to the data. Given the structure specified for each component, this model could be called **psi1(Pasture),gamma(Year+Pasture),eps(Year*Pasture),p(Year)**. Rename the model accordingly then hit **Ok to Run**.

Table 4.1: Beta parameter estimates, standard errors (*SE*) and odds ratios (*OR*) for the effect of pasture on extinction probabilities

Beta Parameter	Estimate	<i>SE</i>	<i>OR</i>
<i>c5</i>	1.62	0.96	5.05
<i>c6</i>	-1.25	1.39	0.29
<i>c7</i>	1.40	0.69	4.04
<i>c8</i>	0.56	0.74	1.75

Interpretation of the results

From this model, $\widehat{a2} = -1.12$ ($SE = 0.029$) suggesting that skink occupancy in the first year was lower for outcrops surrounded by pasture compared to pasture. The respective occupancy probabilities are also given in the output, although as noted previously, interpreting the effects in terms of an odds ratio can be useful, particularly for more complicated models. Therefore, the estimated odds ratio for the effect of pasture on occupancy would be:

$$\begin{aligned}\widehat{OR}_{Pasture} &= e^{-1.12} \\ &= 0.33\end{aligned}$$

This implies that for every unoccupied outcrop in the first year, the number of occupied outcrops surrounded by pasture will be approximately one-third of the number of occupied outcrops surrounded by tussock.

For colonization, the estimated effect of pasture is $\widehat{b4} = -0.87$ ($SE = 0.37$) indicating that colonization probabilities are lower for those outcrops surrounded by pasture. Interpreting this in terms of an odds ratio, it could be concluded that for every unoccupied outcrop that does not become colonized, the number of outcrops surrounded by pasture that become colonized is 0.42 ($= e^{-0.87}$) times the number of outcrops surrounded by tussock that become colonized. Note that this effect is assumed to be consistent across all years so is unaffected by what the overall level of colonization might be in any particular time period.

The effect of pasture on extinction probabilities was allowed to be different over time in this model, and the beta parameter estimates are given in Table 4.1. Note that the estimated effect of pasture is variable, and the size of the standard error is relatively large compared to the estimated effect size in some years suggesting the effect may not be estimated very precisely. Taken at face value, however, the results would suggest outcrops surrounded by pasture have a higher probability of skinks going locally extinct than outcrops surrounded by tussock in most years, although there was one year where extinction probability may have been lower when the outcrop was surrounded by pasture.

4.3.4 Other points for consideration

There are a number of short points to be made before leaving this example.

Alternative design matrices

The first is that often it will be possible to define the same biological model with different design matrices. Estimated probabilities, AIC and log-likelihood values *should* come out to be the same values, but what will be different will be the estimated beta parameters and they will have different interpretations. The other values *should* be the same in theory, but in practice different results can be obtained as the algorithms used by PRESENCE (and other software) to maximize the likelihood and obtain the MLE's are not perfect and they can converge to an incorrect result depending on the particularly design matrices that have been defined and the set of data being used. The 'true' maximum will hold for the different constructions of the model, the algorithms may just sometimes need a bit of assistance in finding that 'true' maximum as they are getting stuck on another local maximum. The other option, of course, is the design matrices have not been set up correctly and that the user is actually attempting to fit models with different biological interpretations.

As an example of using a different design matrix, consider again the extinction component of the previous model with an interaction between the variables year and pasture. The design matrix could have been defined as in Figure 4.9.

	c1	c2	c3	c4	c5	c6	c7	c8
eps1	1	0	0	0	Pasture	0	0	0
eps2	1	1	0	0	Pasture	Pasture	0	0
eps3	1	0	1	0	Pasture	0	Pasture	0
eps4	1	0	0	1	Pasture	0	0	Pasture

Figure 4.9: Design matrix for including an interaction between year and pasture on extinction probability.

Rather than having 1's and the pasture variable name on just the diagonal, there is now a column of 1's in the first column and a column of 'Pasture' in the fifth column. Most importantly is that this alternative construction of the model has the exact same biological mechanism; the effect of pasture on extinction probability is allowed to vary in each year. What is different is the interpretation of the beta parameters or regression coefficients. This design matrix represents the series of equations:

$$\text{logit}(\varepsilon_{i,1}) = c1 + c5 \cdot \text{Pasture}_i \quad (4.25)$$

$$\text{logit}(\varepsilon_{i,2}) = c1 + c2 + c5 \cdot \text{Pasture}_i + c6 \cdot \text{Pasture}_i \quad (4.26)$$

$$\text{logit}(\varepsilon_{i,3}) = c1 + c3 + c5 \cdot \text{Pasture}_i + c7 \cdot \text{Pasture}_i \quad (4.27)$$

$$\text{logit}(\varepsilon_{i,4}) = c1 + c4 + c5 \cdot \text{Pasture}_i + c8 \cdot \text{Pasture}_i \quad (4.28)$$

which could be rewritten as:

$$\text{logit}(\epsilon_{i,1}) = c1 + c5 \cdot \text{Pasture}_i \quad (4.29)$$

$$\text{logit}(\epsilon_{i,2}) = (c1 + c2) + (c5 + c6) \cdot \text{Pasture}_i \quad (4.30)$$

$$\text{logit}(\epsilon_{i,3}) = (c1 + c3) + (c5 + c7) \cdot \text{Pasture}_i \quad (4.31)$$

$$\text{logit}(\epsilon_{i,4}) = (c1 + c4) + (c5 + c8) \cdot \text{Pasture}_i \quad (4.32)$$

Compare this to the form of equations used previously, keeping in mind that some of the beta parameters have different interpretations so are not directly comparable (to help reduce confusion, the previous beta parameters will be denoted with an accent, e.g., $\acute{c}2$). The first equation is exactly the same in each, so $\acute{c}1$ and $c1$, $\acute{c}5$ and $c5$ have the same interpretation both times. However, $\acute{c}2$ is now replaced by the term $c1 + c2$ and, theoretically, once estimated the two terms should have the same value, logit-extinction probability for outcrops surrounded by tussock between years 2 and 3. What is different is that by including the column of 1's the first extinction probability is now being treated as a benchmark against which the other extinction are compared. That is, while $\acute{c}2$ was the absolute value of the logit-extinction probability, $c2$ indicates how *different* the second extinction probability was compared to the first. Similarly, $c3$ and $c4$ indicate how different the third and fourth extinction probabilities were compared to the first. The same change in interpretation occurs with respect to the regression coefficients for the pasture variable. In both cases $\acute{c}5$ and $c5$ is the effect pasture has on the first extinction probability, however while $\acute{c}6$ is the absolute effect pasture has on the second extinction probability, $c6$ is how *different* the effect of pasture is on the second extinction probability compared to the first, and similarly for $c7$ and $c8$.

In some cases, there may be a natural choice in terms of which type of design matrix to use. For example, if a prime question is how different one parameter is compared to another, then using a parameterization similar to the latter one may be preferable. It should also be noted that the column of 1's, or the variable name, does not have to go in the first position and they could have been placed elsewhere which simply would have changed which extinction probability is being treated as the 'standard' against which the other probabilities are being compared. For example, had the column of 1's been placed in the third column (and 1's on the diagonal otherwise) then $c1$, $c2$ and $c4$ would indicate how different extinction was in the respective period compared to the third probability.

Defining the model set

The second point for consideration is that with these multi-state models, the number of potential models you could fit to the data grows exponentially with the number of variables available for each parameter type. Without much effort, the number of possible models can quickly become thousands even with only 3 or 4 factors of interest for each parameter type. The key is to have a good, pragmatic strategy for dealing with such a large set of models. There may often be multiple options for doing so, and one suggestion is to focus on only one parameter type at a time. While doing so, however, it is also suggested that a relatively general model (as opposed to the very simple model, e.g., a constant model) be maintained for the other parameter types to provide them with some flexibility should there actually be some variation in

those other parameters. If there is some variation in those other parameter types, and a very simple model is enforced upon them, then that variation may manifest itself into the parameter currently being focus on leading to misleading inferences.

Maps

Another point is that based upon the PRESENCE output from a multi-season model, it is possible to create maps of occupancy, colonization and extinction probabilities in much the same way as was demonstrated in the final single-season model example. Such maps could be used to illustrate areas of special interest. Maps could also be produced based upon reparameterizations or derived parameters, dependent upon what type of information is ultimately of interest.

Prediction

Finally, a natural application of the multi-season occupancy model is to predict the occurrence or distribution of the species into the future. Once parameter values have been estimated from real data (or failing that, parameter values have been assumed), the general framework of the multi-season model could be used to predict what level of occupancy would be expected 5, 10 or 50 years into the future. Depending upon the available information, these predictions could be in the form of simple numerical summaries (e.g., by using the previously defined recursive equation), or visual summaries (e.g., maps).

Chapter 5

Closing thoughts

Hopefully, this User Manual has been a useful introduction to how to use PRESENCE. I am aware that many topics have been left untouched and, over time, some of those blanks will be filled in. However, as noted at the beginning of this manual, the intent of this User Manual is to be a beginners 'How to' reference and to be complementary to other sources of information that is already available for PRESENCE and occupancy modeling in general. It is highly unlikely that this manual will ever evolve into a font of all knowledge about occupancy modeling; the topic is simply too broad and scope of possible applications is even broader. Further reading will always be required.

Bibliography

- Bailey, L. L., Reid, J. A., Forsman, E. D., & Nichols, J. D. (2009). Modeling co-occurrence of northern spotted and barred owls: Accounting for detection probability differences. *Biological Conservation*, 142(12):2983–2989.
- Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multi-Model Inference*, (2nd ed.). Springer.
- Hines, J. E., Nichols, J. D., Royle, J. A., MacKenzie, D. I., Gopalaswamy, A. M., Kumar, N. S., & Karanth, K. U. (2010). Tigers on trails: occupancy modeling for cluster sampling. *Ecological Applications*, 20(5):1456–1466.
- MacKenzie, D. I., Bailey, L. L., Hines, J. E., & Nichols, J. D. (2011). An integrated model of habitat and species occurrence dynamics. *Methods in Ecology and Evolution*, 2(6):612–622.
- MacKenzie, D. I., Bailey, L. L., & Nichols, J. D. (2004). Investigating species co-occurrence patterns when species are detected imperfectly. *Journal of Animal Ecology*, 73(3):546–555.
- MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., & Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology*, 84(8):2200–2207.
- MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., & Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255.
- MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., & Hines, J. E. (2006). *Occupancy estimation and modeling : inferring patterns and dynamics of species occurrence*. Elsevier.
- MacKenzie, D. I., Nichols, J. D., Seamans, M. E., & Gutiérrez, R. J. (2009). Modeling species occurrence dynamics with multiple states and imperfect detection. *Ecology*, 90(3):823–835.
- MacKenzie, D. I. & Royle, J. A. (2005). Designing occupancy studies: general advice and allocating survey effort. *Journal of Applied Ecology*, 42(6):1105–1114.
- Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., & Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology*, 92(7):1422–1428.

- Nichols, J. D., Bailey, L. L., O'Connell Jr., A. F., Talancy, N. W., Grant, E. H. C., Gilbert, A. T., Annand, E. M., Husband, T. P., & Hines, J. E. (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology*, 45(5):1321–1329.
- Nichols, J. D., Hines, J. E., Mackenzie, D. I., Seamans, M. E., & Gutiérrez, R. J. (2007). Occupancy estimation and modeling with multiple states and state uncertainty. *Ecology*, 88(6):1395–1400.
- Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.
- Royle, J. A. & Nichols, J. D. (2003). Estimating abundance from repeated presence-absence data or point counts. *Ecology*, 84:777–790.
- Stauffer, H. B., Ralph, C. J., & Miller, S. L. (2004). Ranking habitat for marbled murrelets: new conservation approach for species with uncertain detection. *Ecological Applications*, 14(5):1374–1383.
- Tyre, A. J., Tenhumberg, B., Field, S. A., Niejalke, D., Parris, K., & Possingham, H. P. (2003). Improving precision and reducing bias in biological surveys: estimating false-negative error rates. *Ecological Applications*, 13(6):1790–1801.
- Wintle, B. A., McCarthy, M. A., Parris, K. M., & Burgman, M. A. (2004). Precision and bias of methods for estimating point survey detection probabilities. *Ecological Applications*, 14:703–712.