CHAPTER 22

Occupancy models – single-species

Brian D. Gerber, Daniel Martin, Larissa Bailey, *Colorado State University*
Thierry Chambert, *Penn State University & USGS*
Brittany Mosher, *University of Vermont*

As ecologists, we are often interested in how species and communities respond to changes in available resources over space and time. Previous chapters discussed robust methods for evaluating this relationship when we are able to obtain mark-recapture information for individuals or groups. When we are unable to mark animals or when the primary interest is in patterns of species occurrence or proportion of a study area that is occupied or used by a target species, *occupancy models* provide a flexible alternative for elucidating associations between species occurrence and the environment. Our sample unit is thus no longer an individual, but rather a 'site', which is defined based on a study's objective (e.g., 1km^2 grid cell, habitat patch, camera-trap site, transect segment, point-count station).

Occupancy models enable us to estimate the probability of occurrence of a species among sampled sites, while exploring hypotheses about factors (e.g., habitat, environmental conditions, etc.) thought to influence the species' occurrence. The basic sampling design involves randomly selecting a set of independent sites and surveying each site multiple times (i.e., sample 'surveys') during a time period when the state of a site (occupied or not), does not change (see section 22.3, below). The resulting site-specific encounter histories enable us to estimate occupancy – the probability that a site is occupied – while accounting for imperfect species detection. The occupancy approach also enables us to model variation in occupancy and detection simultaneously, relative to site-specific covariates. Additional survey-specific factors can be incorporated into the detection process. The assumptions required by occupancy models share similarities with closed mark-recapture, and are discussed in detail later in this chapter. Historically, occurrence modeling used logistic or probit regression, which assumes a species was absent at sites where it was not detected, and thus inference is known to be biased when the probability of detecting a species at a site is less than one (MacKenzie *et al.* 2017).

Occupancy models have been used to: assess habitat use by single species and communities (e.g., Martin *et al.* 2007, Ruiz-Gutiérrez *et al.* 2010), estimate co-occurrence of competing species (e.g., Bailey *et al.* 2009), estimate prevalence of pathogens (e.g., Lachish *et al.* 2012), quantify effects of habitat fragmentation (e.g., Gerber *et al.* 2012), assess differences in occurrence at multiple scales and to compare different detection methods (Nichols *et al.* 2008), and much more. New approaches are regularly being developed that provide additional flexibility to handle complex study designs, relax model assumptions, and evaluate complex relationships (e.g., potential species interactions) or improve inference by accounting for sources of bias (e.g., 'false-positive' error associated with species misidentification). Here, we begin with the basics: a static, single-season occupancy model.

22.1. The static (single-season) occupancy model

The basic static (single season, single species) occupancy model has two stochastic processes that determine whether the target species is detected at a site. First, the site i may be occupied by the target species with probability ψ_i, or unoccupied, $(1 - \psi_i)$. Assumptions are covered in detail later on (in section 22.3), but it is important to note that in this static model we assume that a site's occupancy status (occupied or unoccupied) does not change between surveys. If site (i) is occupied, the second stochastic process is that there is some probability (p_{ij}) of detecting the species at the site during each survey (j). Conversely, the probability of not detecting the species at an occupied site on survey (j) is $(1 - p_{ij})$. If the site is unoccupied then we cannot detect the species. Of course, this makes the implicit assumption that a species cannot be erroneously detected at an unoccupied site (i.e., no false positive detections; see section (22.3.4) for more details on this assumption and possible ways to relax it). Notice that for each site, we consider whether it is occupied or not occupied and thus are assuming that the state of a site cannot change over surveys $j = 1$ to J, which is the defined season. In the occupancy literature, the term 'season' is similar to a primary period in a robust design (Chapter 16). The definition of a 'site' and a 'season' should be based on the objective of the study (Gerber *et al.* 2014, MacKenzie *et al.* 2017).

Similar to mark-recapture scenarios, now that we have identified our stochastic processes (in this case, ψ and p), we can link our parameters to our detection/non-detection data through probability statements. Let's imagine the simplest scenario where site (i) is surveyed on two occasions $(J = 2)$ and record the detection history, $h_i =$ '01'. We know this site is occupied (because of the detection on $j = 2$). The species was not detected on survey 1, but was detected on survey 2. We can translate this detection history into a probability statement, such that,

$$\Pr(h_i = \text{'01'}) = \psi(1 - p_1)p_2.$$

Notice how our verbal and mathematical statement agree; the site is occupied (ψ) and was not detected on the first survey $(1 - p_1)$, but was detected on the second survey (p_2).

Now, let's consider the slightly more difficult situation when there were no detections of the species, such that $h_i =$ '00'. In this case, we don't know if the site is occupied or not; if the site was occupied we know that the species was not detected on survey 1 or 2. We allow for both possibilities, that the site was occupied or unoccupied, by adding the two probabilities together,

$$\Pr(h_i = \text{'00'}) = \psi(1 - p_1)(1 - p_2) + (1 - \psi).$$

Thus, we are explicitly stating that if the site was occupied, we did not detect the species in survey 1 or 2 (the expression to the left of addition sign), while if the site wasn't occupied (the expression to the right of addition sign) then there was no probability of detecting the species. The addition sign between the two probability statements can be read as logical (Boolean) 'OR'.

To complete this exercise, here are the probability statements associated with the 4 possible detection histories that can be observed for a study with 2 surveys:

history, h_i	*probability expression*
11	$\psi\left[p_1 p_2\right]$
10	$\psi\left[p_1\left(1 - p_2\right)\right]$
01	$\psi\left[\left(1 - p_1\right) p_2\right]$
00	$\psi\left[\left(1 - p_1\right)\left(1 - p_2\right)\right] + \left(1 - \psi\right)$

But, what if we didn't actually survey all sites equally? Do we ignore this information? Absolutely not. Not being able to sample all sites equally is one of the most common circumstances in any field study. Thus, to specify the correct model, we need to differentiate a survey with no detection from that of not conducting a survey, such that there was no possibility of a detection. To do this, we acknowledge that the probability of detecting the target species when we don't survey a site is zero. For example, suppose we survey site 1 three times and obtained the detection history, h_1 = '`000`', while we only surveyed site 2 twice, having missed the second survey because our vehicle broke down. Thus, we obtained the detection history, h_2 = '`0.0`' (where '.' indicates no survey – use of the 'dot' notation is discussed in more detail in Chapter 2 and Chapter 4). We can write our respective probability statements as,

$$\Pr(h_1 = \text{'000'}) = \psi(1-p_1)(1-p_2)(1-p_3) + (1-\psi),$$
$$\Pr(h_2 = \text{'0.0'}) = \psi(1-p_1)(1-p_3) + (1-\psi).$$

Notice that the only difference between these probability statements is that site 2 doesn't include a probability associated with survey 2. In other words, the probability of detecting the species during survey 2 is zero for site 2, because the site was not surveyed.

Now that we can define all probability statements associated with any observed detection history, including when we are missing data, we can link all our data together in a single model likelihood,

$$\mathcal{L}(\psi, p_1, p_2, p_3 \mid h_1, h_2, \ldots, h_N) = \prod_{i=1}^{N} \Pr(h_i),$$

or as the log likelihood (which is more convenient to work with),

$$\log\big(\mathcal{L}(\psi, p_1, p_2, p_3 \mid h_1, h_2, \ldots, h_N)\big) = \sum_{i=1}^{N} \log(\Pr(h_i)).$$

Now, let's consider a more tangible example; imagine a study where we surveyed 100 sites four times each ($N = 100, J = 4$). For simplicity, we will assume constant detection probability, such that $p_1 = p_2 = p_3 = p_4 = p$ and thus when a detection occurred is irrelevant (i.e., probability statements for histories with the same number of detections are equivalent). Our detection histories are given in the following table.

observed frequencies of each equivalent detection history	*detections per site*	*equivalent detection histories (h_i)*	*probability*
10	4	`1111`	ψp^4
20	3	`1110, 1101, 1011, 0111`	$\psi p^3(1-p)$
30	2	`1100, 1001, 1010, 0011, 0101`	$\psi p^2(1-p)^2$
20	1	`1000, 0100, 0010, 0001`	$\psi p(1-p)^3$
20	0	`0000`	$\psi(1-p)^4 + (1-\psi)$

The likelihood for our data and model, assuming constant detection probability (p) can be specified using the frequency of each detection history (data: $y_1 = 10, y_2 = 20, y_3 = 30, y_4 = 20, y_5 = 20$):

$$\begin{aligned}\mathcal{L}(\psi, p \mid \text{data}) &= (\psi p^4)^{y_1} \times (\psi p^3(1-p))^{y_2} \times (\psi p^2(1-p)^2)^{y_3} \times (\psi p(1-p)^3)^{y_4} \times (\psi(1-p)^4) + (1-\psi))^{y_5} \\ &= (\psi p^4)^{10} \times (\psi p^3(1-p))^{20} \times (\psi p^2(1-p)^2)^{30} \times (\psi p(1-p)^3)^{20} \times (\psi(1-p)^4) + (1-\psi))^{20}.\end{aligned}$$

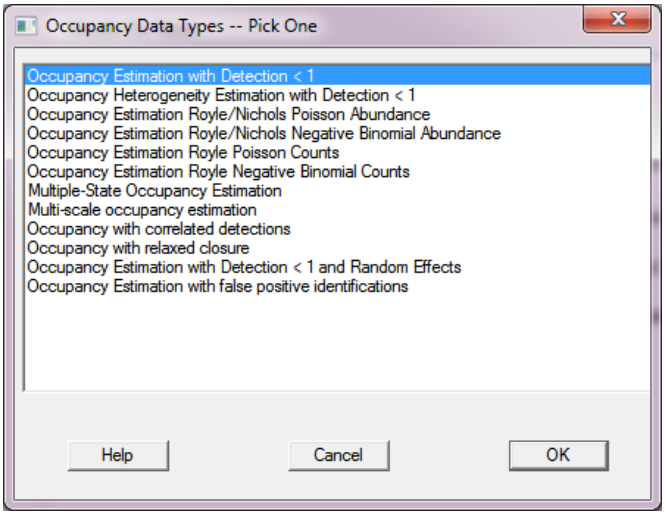
To specify our likelihood more generally, we can rewrite it as:

$$\mathcal{L}(\psi, \mathbf{p} \mid \text{data}) = \left[\psi^{N_d} \prod_{j=1}^{J} p_j^{s_j} (1 - p_j)^{N_d - s_j}\right] \left[\psi \prod_{j=1}^{J} (1 - p_j) + (1 - \psi)\right]^{N - N_d},$$

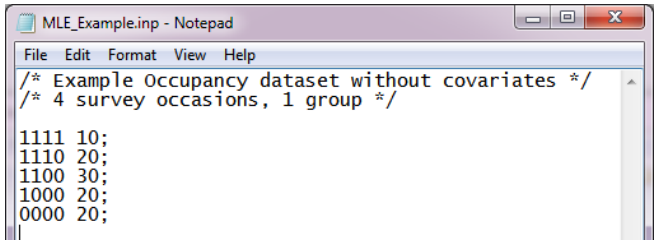
where N_d is the number of sites where the target species was detected at least once and s_j is the number of sites where the species was detected during survey (j). Let's now go through a simple example in **MARK** where we have a single season without covariates.

22.1.1. Single-season occupancy model – example without covariates

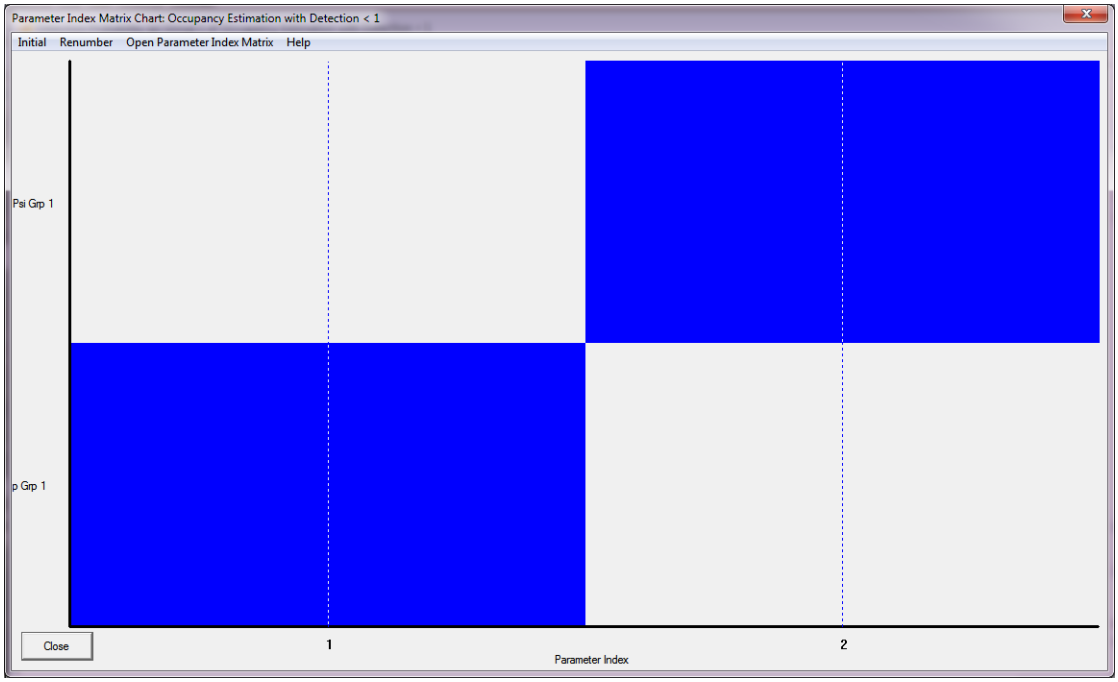
Start a new **MARK** analysis and select the '**`Occupancy Estimation Data Type | Occupancy Estimation with Detection < 1`**'; this refers to a static, single-season occupancy model.



Select the '**`MLE_Example.inp`**' file and view it. You'll notice there are 4 encounter occasions (surveys), no groups, and no individual covariates:



This is the same dataset given in the table above. After entering the appropriate information in the '**`Specification`**' boxes (title, number of sampling sessions, and so on...), click '**`OK`**' and let's build a model. Using the PIM chart (shown at the top of the next page), we can build a simple model assuming constant probability of occupancy and detection, $\{\psi_\cdot\, p_\cdot\}$.



If we run this simple model and view the real parameter estimates from this model, we see:

MLE example - single season occupancy

Real Function Parameters of {Psi(.)p(.)}

Parameter	Estimate	Standard Error	95% Confidence Interval Lower	Upper
1:p	0.5365503	0.0309055	0.4757344	0.5962980
2:Psi	0.8386912	0.0433079	0.7351859	0.9068655

22.1.2. Single-season occupancy model – incorporating covariates

In occupancy analysis we are often interested in how the probability of occupancy varies among sites or whether the probability of detection varies among sites or survey occasions. In these cases, we might have categorical or continuous covariates that represent hypotheses about ecological or observational processes. To parameterize a model is to link our parameters (which are bounded between 0 and 1) to a linear model on the real number line. We commonly use the logit-link function (see Chapter 6),

$$\text{logit}(\psi_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_M x_{i,M},$$

where $x_{i,1}$ through $x_{i,M}$ are the covariate values for covariates 1 through M for site (i). The effects, the βs, are estimated for each covariate.

Similarly, we can consider variation in the detection process across sites and survey occasions,

$$\text{logit}(p_i) = \alpha_0 + \alpha_1 w_{ij,1} + \alpha_2 w_{ij,2} + \cdots + \alpha_{iJP} w_{i,JP},$$

where $w_{ij,1}$ through $w_{i,JP}$ are the covariate values for site (i) and survey (j) for covariates 1 through J

and P; there are $(J \times P)$ coefficient effects (α). We will illustrate models with these types of relationships in the examples below.

22.1.3. An example with covariates

Let's consider a dataset involving northern spotted owls (*Strix occidentalis caurina*) and barred owls (*Strix varia*) in the Pacific Northwest. Barred owls have recently expanded their range and now are present over the entire range of northern spotted owls. Managers may want to know if northern spotted owls are being displaced from territories by barred owls. They may also wonder whether barred owls influence the vocalization behavior (and subsequently, detection probabilities) of northern spotted owls. These ideas were explored in Bailey *et al*. (2009) using a more complex two-species occupancy model, which accounts for the fact that barred owl detection, which we use as a covariate, is imperfect. While the two-species model is more appropriate for this dataset, we use known barred owl detection as a simple site covariate here for demonstrative purposes only.

Day and night surveys were conducted at 159 sites that were roughly the size of a northern spotted owl territory. Owls were detected using vocal imitation or playbacks of spotted owl calls. Multiple surveys were conducted at each site, and observers recorded whether or not spotted owls were detected on each day or night survey. They also recorded whether or not barred owls were ever detected or observed at the site during the season; this information was used as a static site covariate.

As in the preceding example, open **MARK** and select the '**Occupancy Estimation | Occupancy Estimation with Detection <1**' data type (we'll mention some of the other options listed later in this chapter). After giving your analysis a title and selecting the input file '**NSO_SSoccupancy.inp**', you'll need to enter some information about the dataset. Click '**view file**' to open the input file:

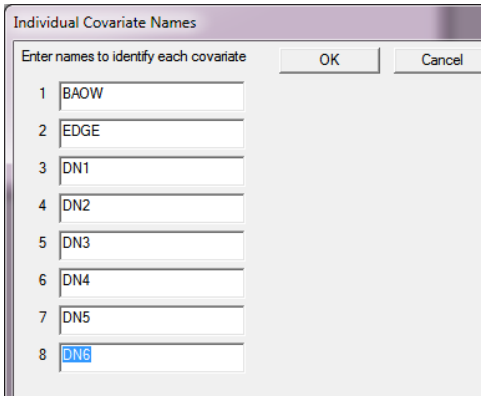
```
/*Subset of Detection-non-detection data from N=159 sites in Oregon. */
/*6 survey occasions, where '.' denotes when sites were not surveyed - ie. missing observations*/
/*8 covariates: 1st covariate (site-specific) = known Barred owl occurence (1) or not (0)*/
/*2nd covariate (site-specific) = proportion edge (0 to 1.0) */
/*Covariates 3-8 denote whether the survey was conducted during the day (0) or night (1) */
/*0.5 was used as missing covariate associated with missing observations - never enters into the likelihood */
/*Subset of data used in Bailey et al. 2009 Biological Conservation */

/*1  */   0000..   1   0  0.090  1  1  1  1  0.5  0.5;
/*2  */   011111   1   0  0.180  0  0  0  0  0  0;
/*3  */   0000..   1   0  0.030  0  1  1  1  0.5  0.5;
/*4  */   100000   1   0  0.170  0  0  0  1  1  1;
/*5  */   0000..   1   1  0.051  1  1  0  1  0.5  0.5;
/*6  */   00010.   1   0  0.030  0  1  1  1  1  0.5;
/*7  */   101111   1   0  0.590  0  0  0  0  0  1;
/*8  */   000000   1   0  0.690  0  0  1  1  1  1;
/*9  */   001000   1   0  0.030  0  1  1  1  1  1;
/*10 */   111111   1   0  0.920  0  0  0  0  0  0;
/*11 */   011101   1   0  0.430  0  1  0  0  0  1;
/*12 */   011111   1   0  0.670  0  1  0  0  1  1;
/*13 */   0000..   1   0  0.970  0  1  1  1  0.5  0.5;
/*14 */   111001   1   0  0.040  0  0  1  0  0  1;
/*15 */   00101.   1   0  0.250  0  1  1  1  1  0.5;
```

The input file contains the site number, followed by the encounter history at a particular site. We can see that up to six surveys were conducted at each site (encounter occasions = 6). Notice that some encounter histories (sites 1, 3, and 5, for example) contain dots ('.') indicating missing observations. Recall that missing observations do not enter into the likelihood and therefore do not influence estimates of detection probabilities.

The remaining columns in the input file are a frequency column, followed by 8 individual covariates. The first covariate denotes whether barred owls are known to exist at a given site, the second column gives the proportion of edge habitat at each site, and the remaining 6 columns denote whether a day (0) or night (1) survey was conducted for each of the 6 surveys.

You'll notice the value 0.5 also occurs in the day/night covariate columns. Unlike missing survey information, **MARK** does not tolerate missing covariate values. We 'trick' **MARK** into thinking the covariate exists by replacing missing covariate values with a placeholder (in our case 0.5). As long as those missing covariate values are paired with missing survey information, neither piece of information is included in the likelihood and thus won't affect estimates and inferences. It's best to name the individual covariates. We called ours **`BAOW`** (for barred owl), **`EDGE`** (for proportion of edge habitat), and **`DN1-DN6`** to indicate whether each survey was conducted during the day or night (Day = `0`, Night = `1`).



Open the PIM chart and investigate the parameters that we will be dealing with. You'll see there are just two parameter types listed and the default model structure includes time-varying detection probability and a constant occupancy probability. Remember that because occupancy is assumed to be closed to changes within a season, occupancy probabilities cannot vary with time (i.e., across surveys). Run the default model, which we might call $\{\psi_{.}\ p_{t}\}$. This model should have 7 parameters (one detection probability for each of 6 surveys, and one occupancy probability) and AIC_c = 840.58. If we look at the real parameter estimates, we see that the occupancy probability of northern spotted owls across sites is 0.59 with detection probabilities ranging from 0.50 to 0.68 depending on the survey:

NSO - single season

Real Function Parameters of {psi(.)p(t)}

Parameter	Estimate	Standard Error	95% Confidence Interval Lower	Upper
1:p	0.4999434	0.0521911	0.3990550	0.6008365
2:p	0.6169510	0.0511098	0.5131806	0.7110562
3:p	0.6754290	0.0500061	0.5709577	0.7649348
4:p	0.6498967	0.0529087	0.5406150	0.7454235
5:p	0.6708014	0.0587894	0.5473608	0.7744499
6:p	0.6728113	0.0650780	0.5353163	0.7858947

In occupancy modeling, we are generally interested in how occupancy and detection probabilities are influenced by environmental or survey covariates of interest. Here, we are interested in investigating:

1. how known barred owl occurrence may influence spotted owl occupancy,
2. how the proportion of edge habitat influences spotted owl occupancy,
3. if spotted owl detection probability varies among day and night surveys, or among sites with and without known barred owls.

To evaluate hypotheses related to individual covariates we have to use the design matrix. First, let's run a model where spotted owl detection depends on known (naïve) barred owl occurrence. Retrieve

your $\{\psi_. \, p_t\}$ model and open a (reduced) design matrix with 3 columns; the first six rows in this matrix labeled **1:p** through **6:p** represent survey-specific detection probabilities and the seventh row represents occupancy (**psi**). It is worth noting that both parameter types (ψ and p) share a single design matrix in **MARK**, as opposed to other software (notably, program **PRESENCE**). This makes it possible for these two parameter types to share covariates, but remember, just because something is possible doesn't mean you should do it. You may want to refer to Chapters 6 and 11 for a refresher on constructing linear models in **MARK** using individual covariates.

Suppose we want to model spotted owl detection probability as a function (on the logit scale) of known barred owl occurrence at site (i), and model spotted owl occupancy as constant among sites:

$$\text{logit}(p_i) = \beta_1 + \beta_2(\texttt{BAOW}_i),$$
$$\text{logit}(\psi) = \beta_3.$$

B1: p Int	Parm	B2 p-BAOW	B3: Psi Int
1	1:p	BAOW	0
1	2:p	BAOW	0
1	3:p	BAOW	0
1	4:p	BAOW	0
1	5:p	BAOW	0
1	6:p	BAOW	0
0	7:Psi	0	1

It's (very) helpful to label your columns, as shown, by clicking the '**Appearance**' tab and then '**Label Column**' while the design matrix is open. Run this model. The model should have 3 parameters and an AIC_c of 798.95.

Let's run a few more models using the design matrix to round out our model set. Fit the model $\{\psi_{\texttt{BAOW}} \, p_t\}$ by opening a reduced design matrix with 8 columns. The detection probabilities should have a common intercept that is interpreted as the detection probability at a reference survey (time 1 in the example design matrix below), and an additional parameter for each subsequent survey. We wish to model occupancy as a function of known barred owl occurrence. This is translated to a logit-linear model where detection varies by survey j and occupancy varies by site i:

$$\text{logit}(p_j) = \beta_1 + \beta_2(\texttt{time2}) + \beta_3(\texttt{time3}) + \beta_4(\texttt{time4}) + \beta_5(\texttt{time5}) + \beta_6(\texttt{time6}),$$
$$\text{logit}(\psi_i) = \beta_7 + \beta_8(\texttt{BAOW}_i).$$

B1 p-t1	B2 p-t2	B3 p-t3	B4 p-t4	Parm	B5 p-t5	B6 p-t6	B7 psi-int	B8 psi-BAOW
1	0	0	0	1:p	0	0	0	0
1	1	0	0	2:p	0	0	0	0
1	0	1	0	3:p	0	0	0	0
1	0	0	1	4:p	0	0	0	0
1	0	0	0	5:p	1	0	0	0
1	0	0	0	6:p	0	1	0	0
0	0	0	0	7:Psi	0	0	1	BAOW

Run this model. It should have 8 parameters and $\text{AIC}_c = 842.76$.

Let's also run $\{\psi_{\texttt{BAOW}}\, p_{\texttt{Night/Day}}\}$ to explore the effect of the survey time (day or night) on detection probability. This model should have 4 parameters:

$$\text{logit}(p_j) = \beta_1 + \beta_2(\texttt{DN}_j),$$
$$\text{logit}(\psi_i) = \beta_3 + \beta_4(\texttt{BAOW}_i).$$

B1 p-int	B2 p-DayNight	Parm	B3 psi-int	B4 psi-BAOW
1	DN1	1:p	0	0
1	DN2	2:p	0	0
1	DN3	3:p	0	0
1	DN4	4:p	0	0
1	DN5	5:p	0	0
1	DN6	6:p	0	0
0	0	7:Psi	1	BAOW

We could fit $\{\psi_{\texttt{BAOW}}\, p_{\texttt{BAOW}}\}$ to explore the hypothesis that barred owls affect both the occupancy and detection probabilities of spotted owls. This model should have 4 parameters:

$$\text{logit}(p_i) = \beta_1 + \beta_2(\texttt{BAOW}_i),$$
$$\text{logit}(\psi_i) = \beta_3 + \beta_4(\texttt{BAOW}_i).$$

Finally, let's run a model where occupancy of spotted owls varies with the proportion of edge habitat at a site and detection varies with barred owl occurrence. We'll call this model $\{\psi_{\texttt{EDGE}}\, p_{\texttt{BAOW}}\}$.

$$\text{logit}(p_i) = \beta_1 + \beta_2(\texttt{BAOW}_i),$$
$$\text{logit}(\psi_i) = \beta_3 + \beta_4(\texttt{EDGE}_i).$$

Now, take a look at the model selection table we have created. It should contain 6 models, three of which have some support. All three supported models suggest that known barred owl occurrence influences spotted owl detection probability, but there is uncertainty in whether occupancy varies with known barred owl occurrence (model 2), proportion of edge habitat (model 3), or is constant (model 1). Both the second and third models differ from the first by a single parameter.

Model	AICc	Delta AICc	AICc Weight	Model Likelihood	No. Par.	Deviance	-2Log(L)
{psi(.)p(BAOW)}	798.9511	0.0000	0.55514	1.0000	3	792.7963	792.7963
{psi(BAOW)p(BAOW)}	800.5382	1.5871	0.25105	0.4522	4	792.2784	792.2784
{psi(EDGE)p(BAOW)}	801.0558	2.1047	0.19381	0.3491	4	792.7960	792.7960
{psi(BAOW)p(DayNight)}	830.2793	31.3282	0.00000	0.0000	4	822.0195	822.0195
{psi(.)p(t)}	840.5764	41.6253	0.00000	0.0000	7	825.8347	825.8347
{psi(BAOW)p(t)}	842.7599	43.8088	0.00000	0.0000	8	825.7999	825.7999

If we look closely at the AIC_c of these three models, we notice that the $\{\psi_{\texttt{EDGE}}\, p_{\texttt{BAOW}}\}$ is almost exactly 2 AIC_c units different from the top model, and has a nearly identical deviance. This indicates that the 2-unit penalty of adding an additional parameter was not compensated with an increase in model fit. The '**EDGE**' covariate appears to be a 'pretending' or 'uninformative' variable – it rides the coattails of the best model's structure, and does not add anything to our biological understanding of this system (see the `-sidebar-` on pp. 61-62 in Chapter 4).

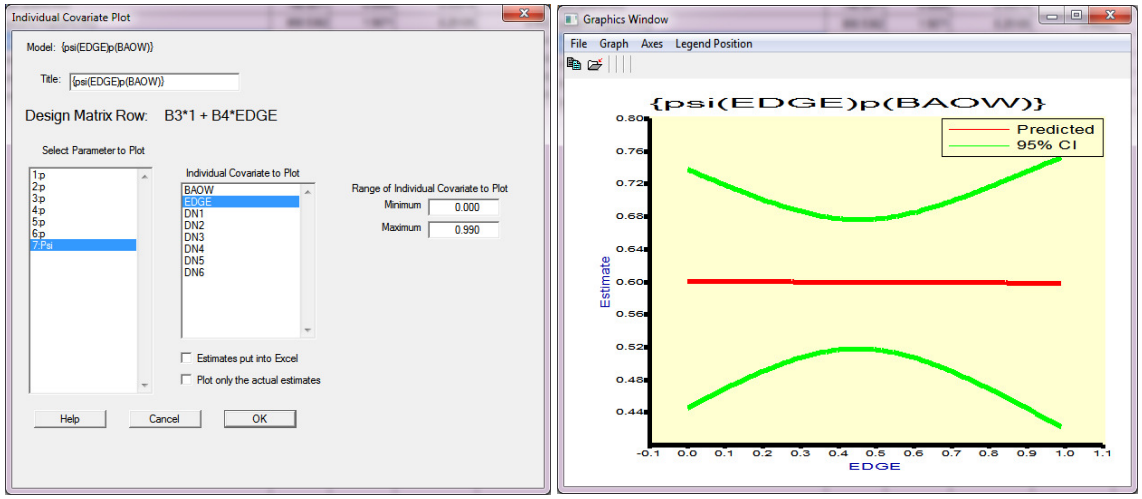
We can confirm that EDGE is an 'uninformative' variable in a couple of other ways. First, let's look at the estimated coefficient describing the effect of **EDGE** (β_4):

NSO - single season

LOGIT Link Function Parameters of {psi(EDGE)p(BAOW)}

Parameter	Beta	Standard Error	95% Confidence Interval Lower	Upper
1:p-int	0.9035150	0.1206998	0.6669434	1.1400867
2:p-EDGE	-1.4716362	0.2342371	-1.9307409	-1.0125314
3:psi-int	0.4079630	0.3205795	-0.2203728	1.0362988
4:psi-EDGE	-0.0088269	0.6005373	-1.1858799	1.1682262

As expected, the effect of edge is estimated very close to zero, with a confidence interval that includes both negative and positive values. Another way to visualize that edge has little effect on occupancy is to plot occupancy over the range of observed edge values using the '**Individual Covariate Plot**' function (first discussed in section 11.5 in Chapter 11). Highlight the $\{\psi_{\text{EDGE}}\, p_{\text{BAOW}}\}$ model in the browser, and click the blue '**Individual Covariate Plot**' icon. Select the parameter (ψ) and covariate (EDGE) that we want to plot and press '**OK**':



We see that occupancy is virtually constant across values of the proportion edge covariate, confirming that edge is a pretending variable. As we continue to work through the rest of this exercise, you might consider whether the effect of barred owl occurrence on occupancy probability is also a pretending variable.

All supported models in the candidate set suggest that known barred owl occurrence influence spotted owl detection probability. The best-supported model received 0.56 of the weight and suggested occupancy was constant among sites with detection probability depending on the occurrence of barred owls. If we look at the estimated coefficient that describes the relationship between detection probability and known barred owl occurrence in the top model (top of the next page), we see that the effect of barred owl occurrence on detection probability is negative and is estimated to be –1.47 on the logit (or log-odds) scale.

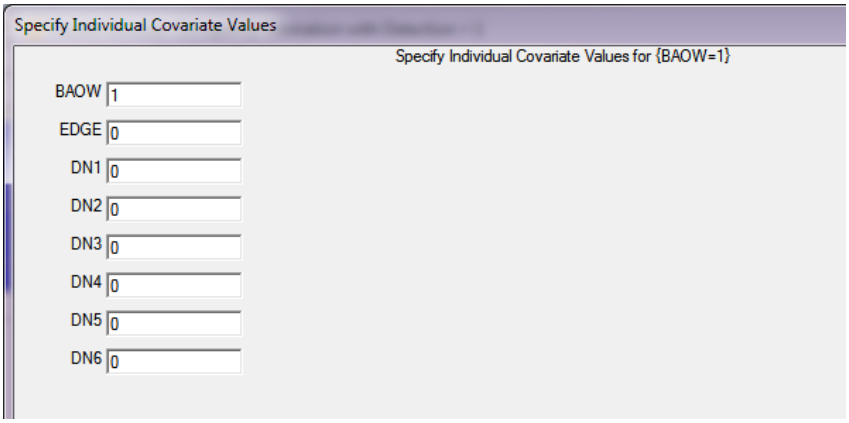
NSO - single season

LOGIT Link Function Parameters of {psi(.)p(BAOW)}

Parameter	Beta	Standard Error	95% Confidence Interval Lower	Upper
1:p Int	0.9035177	0.1206996	0.6669466	1.1400889
2:p-BAOW	-1.4716912	0.2342194	-1.9307612	-1.0126211
3:Psi Int	0.4039630	0.1693957	0.0719474	0.7359787

But just how different is the detection probability of northern spotted owls when barred owls are and are not detected at a site? If we simply look at the real parameter estimates, what we see is the detection probability at the average value of the BAOW covariate (0.23). This is not informative for a binary variable like BAOW. To see what the detection probability is when 'BAOW=1' or 'BAOW=0', we can either re-run the model and check '**User-specified covariate values**' in the '**Run**' dialogue box, or we can use the '**ReGenerate Real and Derived Estimates**' function in the browser '**Run**' menu.

We'll illustrate the '**User-specified covariate**' option here, and the '**Re-generate**' option in the model averaging section below. Retrieve the top model and click the '**Run**' icon. In the bottom right corner of the dialogue box select '**User-specified covariate values**'. This tells **MARK** that we'd like real parameter estimates for a certain covariate value, rather than for the mean value. Change the model name to include 'BAOW=1' and press '**OK to Run**'. When prompted, input a '1' for the **BAOW** covariate value:



You'll notice that the AIC_c of this model is identical to the first model you ran with this structure – that's because it's the same model. However, when we examine the real parameter estimates, you'll see that the detection probabilities given are for the particular case where BAOW=1. Repeat these steps but set BAOW=0. You'll find that the detection probability of northern spotted owls is 0.71 when barred owls are not found at a site, and is only 0.36 when barred owls are known to be present. An odds ratio for this difference can be obtained by exponentiating the coefficient of the **BAOW** effect (β_2):

$$\begin{aligned} e^{\beta_2} &= e^{(-1.47)} \\ &= 0.229. \end{aligned}$$

Thus, the detection of northern spotted owls is 0.23 times lower when barred owls are known to be present! Perhaps more intuitively, northern spotted owl detection is 4.35 times higher when barred owls are not detected ($1/e^{\beta_2} = 4.35$). For a refresher on odds ratios, see section 6.13.1 in Chapter 6.

22.2. Model averaging

How can we obtain our best estimate of northern spotted owl occupancy, given the slight model selection uncertainty we have? Model averaging (introduced in Chapter 4) gives us a way to honestly represent the uncertainty of our estimates, which include two components: parameter uncertainty and model uncertainty. When we use model selection criteria, there is virtually always uncertainty regarding which model is the best. Therefore, the measures of uncertainty that we report (e.g., SE, 95% CI) should include this model selection uncertainty in addition to the parameter uncertainty present in any given model.

Before we model average, let's remove the $\{\psi_{\texttt{EDGE}}\, p_{\texttt{BAOW}}\}$ model because we have confirmed (above) that this model doesn't contain any information not already present in other models. Let's also remove the $\{\psi_{\cdot}\, p_{\texttt{BAOW=1}}\}$ and $\{\psi_{\cdot}\, p_{\texttt{BAOW=0}}\}$ models because they are identical to $\{\psi_{\cdot}\, p_{\texttt{BAOW}}\}$ and we don't want to artificially inflate that model's support. We are left with a set of 5 models, with only two models receiving any AIC_c weight:

Model	AICc	Delta AICc	AICc Weight	Model Likelihood	No. Par.	Deviance	-2Log(L)
{psi(.)p(BAOW)}	798.9511	0.0000	0.68859	1.0000	3	792.7963	792.7963
{psi(BAOW)p(BAOW)}	800.5382	1.5871	0.31141	0.4522	4	792.2784	792.2784
{psi(BAOW)p(DayNight)}	830.2793	31.3282	0.00000	0.0000	4	822.0195	822.0195
{psi(.)p(t)}	840.5764	41.6253	0.00000	0.0000	7	825.8347	825.8347
{psi(BAOW)p(t)}	842.7599	43.8088	0.00000	0.0000	8	825.7999	825.7999

When we model average, we take the real parameter estimates from each model and weight them according to that model's support. Therefore, before we model average we need to consider which real parameters are of interest and which covariate values are necessary to create those real parameter values. Suppose we wanted estimates of spotted owl occupancy and detection when barred owls are present. Before model averaging, we'd want to update our models to reflect '`BAOW = 1`', so that real parameters that are averaged are comparable. We could retrieve and rerun each model and use the '**User-specified Individual Covariates**' function that we discussed above, but that could get tedious, especially for large model sets or several covariate combinations of interest.

An alternative is to use the '**Regenerate Real and Derived Estimates**' function in the '**Run**' menu which does not rerun models, but does calculate new real parameter estimates for particular covariate values. Select this option in '**Run**', then select all models and press '**OK**'. Now you can input covariate values of interest. In our case, we want to set '`BAOW = 1`'. **EDGE** no longer occurs in our model set, so it can be ignored. The day or night survey indicator does exist in the model set, but these models containing the survey time covariate didn't receive any weight so that covariate can also be ignored. After pressing '**OK**', **MARK** will regenerate real and derived parameter estimates for each model. Take a look at the real estimates from the top model to be sure it worked. Sure enough, '`BAOW = 1`' is now listed as the covariate value of interest, and the corresponding detection probability (0.36) is reported.

Now that all models are reporting real parameter estimates for a consistent scenario (`BAOW = 1`), we can model average the detection probability estimates at sites where barred owl were detected during the season. The '**Model Averaging**' option is in the '**Output**' menu, and we will select '**Real**' parameter estimates. Next, specify the parameters of interest. Because the models with time variation in detection receive no AIC_c weight, we can get away with selecting any of the 6 detection probabilities. We will also select parameter 7, which is occupancy. Once we press '**OK**' our two estimates of interest are generated (shown at the top of the next page).

At sites where barred owls are detected during the season our best estimates of spotted owl detection and occupancy probability are $\hat{p} = 0.36$ and $\hat{\psi} = 0.62$. We can use the '**Regenerate Real and Derived Estimates**' again for the scenario where barred owls are not detected (`BAOW=0`). If we do this, and model

```
                              NSO - single season
      Estimates only for data type Occupancy Estimation with Detection < 1

                Detection Probability (p) Group 1 Parameter 1
Model                                      Weight    Estimate      Standard Error
------------------------------------------ -------   ------------- -------------
{psi(.)p(BAOW)}                           0.68859   0.3616584     0.0464368
{psi(BAOW)p(BAOW)}                        0.31141   0.3514682     0.0493544
------------------------------------------ -------   ------------- -------------
Weighted Average                                     0.3584851     0.0473454
Unconditional SE                                                   0.0475991
95% CI for Wgt. Ave. Est. (logit trans.) is 0.2713826 to 0.4560459
Percent of Variation Attributable to Model Variation is 1.06%


                      Occupancy (Psi) Group 1 Parameter 7
Model                                      Weight    Estimate      Standard Error
------------------------------------------ -------   ------------- -------------
{psi(.)p(BAOW)}                           0.68859   0.5996394     0.0406672
{psi(BAOW)p(BAOW)}                        0.31141   0.6661388     0.1019703
------------------------------------------ -------   ------------- -------------
Weighted Average                                     0.6203478     0.0597574
Unconditional SE                                                   0.0729730
95% CI for Wgt. Ave. Est. (logit trans.) is 0.4709660 to 0.7499444
Percent of Variation Attributable to Model Variation is 32.94%
```

average once again, we find that our best estimates of spotted owl detection and occupancy are $\hat{p} = 0.71$ and $\hat{\psi} = 0.60$. By comparing the real estimates and viewing the model selection results, we can infer, the impact of known barred owl presence on spotted owls seems to be largely related to detection and not to occurrence.

begin sidebar

conditional site occupancy

Ecologists are often interested in the probability of occupancy at a site, conditional upon the species not being detected. We know that if a species was detected at a site that it is occupied, but what about the sites where there were no detections? We can derive an occupancy estimate, conditional on the survey effort (detection history), using the following expression,

$$\hat{\psi}_{i,condl} = \frac{\hat{\psi}_i(1-\hat{p}_i)^J}{(1-\hat{\psi}_i) + \hat{\psi}_i(1-\hat{p}_i)^J},$$

using estimated probabilities of occupancy and detection (p) and the number of surveys (J). Notice what this expression represents: the numerator is the probability that the species occurred, but was not detected during J surveys. The denominator represents the probability that the species was not detected, either because it was absent $(1 - \psi_i)$ or because it was present but not detected. So the expression represents the probability that the site was occupied, given that the species was not detected. Ideally, this probability would be very low for barred owls to justify using known barred owl detection as a covariate in the example above.

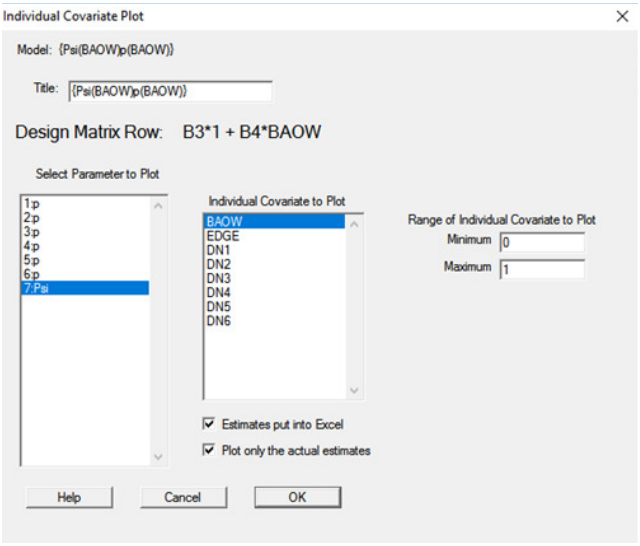
end sidebar

We may want to compare our occupancy estimate to a naïve occupancy estimate, which is simply the proportion of sites where the species was detected at least once. We refer to this estimate as 'naïve' because it ignores the possible influences of imperfect detection (thus underestimating occupancy) and typically lacks a measure of precision. In our example, we detected spotted owls at 92 of the 159 sites, so our naïve estimate of occupancy is (92/159) or 0.58. To compare this value to an occupancy estimate, we typically report the estimate of occupancy from a model that accounts for variation in detection probability, but assumes that occupancy is constant among sites.

For example, the model $\{\psi\,.\,p_{\text{BAOW}}\}$ accounts for the important variation in detection probability among

sites with and without known barred owls; the occupancy estimate from this model is 0.60 ($\widehat{SE}(\hat{\psi}) = 0.04$). With this example, there is little difference between the estimated and naïve occupancy values because the probability of detecting spotted owls at least once on sites without known barred owls is 1 ($p^* = 1 - (1 - p_{(\texttt{BAOW=0})})^6 = 1 - (1 - 0.71)^6 = 1.0$). This probability is also very high on sites with known barred owls ($p^* = 1 - (1 - p_{(\texttt{BAOW=1})})^6 = 1 - (1 - 0.36)^6 = 0.93$); however, on sites with missing surveys this cumulative detection probability is obviously lower.

Sometimes biologists are interested in estimating the proportion of occupied sites in the sample, in addition to using the sample to make inferences about occupancy for the entire population of interest. In these cases, the occupancy estimates from a model with constant occupancy e.g., $\{\psi\,.\,p_{\texttt{BAOW}}\}$) provide a good approximation for the overall proportion of sites occupied in the sample. Another estimator that is often used, if you want to incorporate covariate effects, is to calculate the mean of the site-specific occupancy estimates for each site in the sample. For example, let's suppose that our second model $\{\psi_{\texttt{BAOW}}\, p_{\texttt{BAOW}}\}$ was much better supported than the constant occupancy model. In this case, we might want to know the estimated proportion of sites occupied using this model because it accounts for variation in occupancy probabilities among sites. An easy way to calculate the mean of the site-specific occupancy estimates is to use features found in the '**`Individual Covariate Plot`**'. To demonstrate, highlight the $\{\psi_{BAOW}\, p_{BAOW}\}$ model in the browser, and click the blue '**`Individual Covariate Plot`**' icon. Select the parameter ψ (**`Psi`**) and covariate (**`BAOW`**) and click the two boxes at the bottom to export the actual estimates into Excel, then click '**`OK`**'.



A graph will appear with only two estimates, one for `BAOW = 0` ($\psi = 0.59$) and another for '`BAOW = 1`' (this estimate may be hidden behind the '**`Predicted`**' box, but it is approximately 0.67). Additionally, an Excel spreadsheet with the 159 site-specific occupancy estimates should appear. Notice the site-specific values of each covariate in the model are reported for each site and the associated site-specific occupancy estimate - there are only two different estimates in our case because there are only two values for the **`BAOW`** covariate. We can easily calculate the mean of these site-specific occupancy estimates (0.605) and the associated standard error (0.03). A final estimator for the proportion of sites occupied in the sample would be to calculate the site-specific conditional occupancy estimates (see above side-bar) which would account for variation in survey effort across sites, and then sum the conditional occupancy estimates.

Importantly, the mean of the site-specific occupancy estimates and the sum of the conditional occupancy estimates are estimators for the proportion of sites occupied. How relevant these estimates

are to the proportion of sites occupied in the entire area of interest depends on how representative the sample is when compared to the entire population of sites. Random selection of sites will help insure that the surveyed sites are representative of the entire population of sites in the area of interest; however, if a stratified random design is employed there are occupancy estimators that weight strata-specific occupancy estimates by appropriate inclusion probabilities (e.g., Gould *et al*. 2012).

22.3. Model assumptions

As with any statistical model, we make assumptions which if correct, allow for appropriate inference from the data, based on model parameters. However, just because we violate a model assumption, does not necessarily mean that our inferences are wrong; in some cases, inferences may be robust to violations of an assumption. In these cases, we can say that a model is robust to the assumption. Generally, we want to try and meet all model assumptions through study design or model specification.

The five basic occupancy model assumptions are:

1. The occupancy status at each site does not change during the survey period. In other words, sites are closed to changes in occupancy status. For example, in the case where we are studying the occurrence of snapping turtles at a sample of wetlands, if a wetland is occupied, we assume that the species is available for detection during each survey during the sampling season.
2. The probability of occupancy is constant across all sites, or if heterogeneous, it is appropriately modeled using site-level covariates.
3. The probability of detection is constant across all occupied sites, or if heterogeneous, it is appropriately modeled using site or survey level covariates.
4. The detection of species and thus the detection histories at each site are independent.
5. The species is not misidentified, there are no false positives, meaning it is impossible for a detection to occur at an unoccupied site: false detections from species misidentification do not occur. Thus, detections can only occur at occupied sites.

These assumptions are discussed in detail in MacKenzie *et al*. (2017); the authors describe potential violations of these assumptions and study design considerations to meet these assumptions. In this section, we briefly summarize these assumptions and describe recent work to relax or test the assumptions.

22.3.1. Closure

The model described above applies to a single season. The occupancy state is assumed to be constant across the entire season, meaning that the occupancy state does not change between surveys. A site that is occupied (or unoccupied) remains so for all surveys of that season. This assumption also extends to the dynamic (multi-season) occupancy model (discussed later in this chapter). The dynamic model considers several seasons between which the occupancy status might change, and is similar to the robust design presented in Chapter 16. In the occupancy literature, primary periods are often referred to as 'seasons' and secondary sessions as 'surveys'. Kendall & White (2009) reviewed the robustness of the single-season occupancy model to violations of the closure assumption. They confirmed that if a species is randomly available for detection at a site, as might be the case for a wide-ranging, territorial species that moves in and out of the sample unit, the resulting occupancy estimates are unbiased but should be interpreted as probability of the unit being 'used' by the species.

In another example, suppose we are interested in the occurrence of a neotropical bird species within the breeding range and a randomly placed fixed-radius point count was located near the territory of a single male. The species may not always be within the sample unit (fixed-radius point count area) during every survey conducted at the site, but the species movement in and out of the unit likely resembles a random process.

In these cases, detection probability is really a product of two probabilities: (1) the probability the species is available for detection within the site during a given survey, and (2) the probability of detection given the species is available. These two probabilities cannot be separated in the single-season occupancy model, but sub-sampling at a finer temporal scale can allow separation of these two processes – see Nichols *et al*. (2008) and Rota *et al*. (2009) for examples.

In situations where availability is non-random, bias in occupancy estimates is expected (Kendall & White 2009, Kendall *et al*. 2013, MacKenzie *et al*. 2017). There are various ways to reduce or eliminate bias caused by non-random species availability, or non-random movement in and out of sites, depending on the temporal pattern of species availability. When species movement consists of ingress only, or egress only, surveys can be pooled to eliminate bias (see Kendall & White 2009, and Kendall 1999 for details). Weir *et al*. (2005) dealt with seasonal changes in detection for different species of anurans by modeling detection probabilities as a function of date, and this approach is effective at eliminating bias in occupancy (Kendall *et al*. 2013).

Several studies have employed additional surveys at a finer temporal scale to separate the processes of availability and detection (Rota *et al*. 2009, Nichols *et al*. 2008). For example, Otto *et al*. (2013) suspected that seasonal availability of a terrestrial salamander species might change during the summer, a time period over which occupancy is usually assumed to be static. Additional temporal sub-sampling allowed these authors to differentiate seasonal availability and potential behavioral responses of the species due to habitat disruption during surveys. Temporal sub-sampling enables investigators to use dynamic occupancy models to test for closure by fitting models where extinction and colonization probabilities are fixed to zero and comparing them to models where these probabilities are estimated.

Finally, Kendall and colleagues have developed an occupancy model akin to the open robust design models presented in Chapter 16. The model explicitly estimates entry and exit parameters, enabling investigators to estimate residence or stopover time (phenology) among sites during a given season (Kendall *et al*. 2013, Chambert *et al*. 2015).

22.3.2. Unmodeled heterogeneity in occupancy or detection probability

The second and third occupancy assumptions imply that variation in occupancy and detection probability is appropriately modeled with covariates. The impact of unmodeled variation in occupancy probability among sites has not been well studied relative to the other model assumptions. Still, this situation likely occurs, especially when analyzing historic data where relevant covariates were simply not collected. For example, the distribution of most carnivores is likely a function of local prey abundance or density, but these covariates are often not available. Coarse-scale habitat metrics may serve as a proxy for such data, but in some cases this information may not be available or is a poor representation of the abundance distribution of primary prey species. In these cases, we suspect that the resulting occupancy estimates represent the average occupancy for the sampled sites and the reported variances are likely conservative (MacKenzie *et al*. 2017).

Unmodeled heterogeneity in detection probability will often result in negatively biased occupancy estimates (e.g., Royle & Nichols 2003, MacKenzie & Bailey 2004, Royle 2005, MacKenzie *et al*. 2017). Detection probabilities may vary among sites, for example among habitat types, or among surveys due to environmental factors, observers, or seasonal behavior. If this variation is not modeled via covariates,

the resulting estimates can be biased. The degree of bias relates to the magnitude of the variation and the number of surveys and sites. One unique form of heterogeneity that cannot be easily modeled with covariates is variation in detection due to the size of the local population at each site (i.e., a species local abundance).

Indeed, if a site is occupied by many individuals (of the focal species), it should have a higher detection than sites that are occupied by only a few individuals. Ways of dealing with this type of heterogeneity in detection is the focus of the next section.

22.3.3. Lack of independence

Non-independence among detection histories can arise if sites are located too close to one another, allowing an individual animal to be detected at multiple sites simultaneously. For example, if sites are located nearby, a spotted owl hooting at one site could also be recorded during the same survey at the other nearby site. Likewise, if remote-cameras are located near each other, a single individual may be detected at multiple cameras (sites) during a given week (survey). In these instances, the number of sites surveyed (or the number of detection histories) is not a good representation of the number of independent units in the study. This is a form of overdispersion (Chapter 5), and in these cases, the estimates of occupancy are often unbiased, but estimates of precision are too small because the true number of independent sites is smaller than the number of sites surveyed (MacKenzie & Bailey 2004, MacKenzie *et al*. 2017). Employing sampling designs where sites are randomly selected with appropriate spacing to assure independence is the best way to meet this assumption; however, goodness-of-fit assessments have some power to detect and adjust for this type of overdispersion (MacKenzie & Bailey 2004, MacKenzie *et al*. 2017 p. 155 '*Assessing Model Fit*').

Another type of detection dependence may occur in monitoring programs where an observer is assigned a subset of sites, and surveys those sites multiple times. In these cases, once the observer detects the species, they know where to look for the species on subsequent surveys, increasing the detection probability after first detection. This is particularly relevant if multiple surveys are conducted on the same visit. For example, suppose that we are again studying the occurrence of a neotropical bird species and during a single visit to our randomly chosen fixed-radius points, a single observer conducts three 5-minute surveys, perhaps with little time between surveys. In this case, once the observer sees or hears a given species, the observer is alerted to its location and will be more likely to see or hear it in the subsequent surveys. Obviously, the detection of the species is no longer independent and the process resembles a behavioral response that is common in many closed-population models, where the initial capture probability is lower than the recapture probability (i.e., a trap-happy response, Chapter 14). Accordingly, we can fit models that account for this potential effect. How do we do this without the equivalent of a 'recapture probability' for detection? Well, one way is to develop survey-specific covariates that indicate a permanent change in the detection probability after first detection at a site.

For example, consider the detection histories from a couple of the spotted owl sites in the previous example, and let's pretend we had no other covariates. Here are the detection histories for sites 1, 2, 4 and 6, followed by the frequency column:

```
/* 1 */   0000..    1
/* 2 */   011111    1
/* 4 */   100000    1
/* 6 */   00010.    1
```

To develop the six survey-specific covariates to represent this behavioral effect (call these `beh_t1`, `beh_t2`, `beh_t3`, `beh_t4`, `beh_t5`, `beh_t6`), we simply denote whether the species had been previously detected (`1`) or not (`0`) for each survey.

So, the six covariates would look like:

```
/* 1 */   0000..   1  0 0 0 0 0 0;
/* 2 */   011111   1  0 0 1 1 1 1;
/* 4 */   100000   1  0 1 1 1 1 1;
/* 6 */   00010.   1  0 0 0 0 1 0.5;
```

Notice for site 2 the covariate values are `0` (`beh_t1=t2=0`) until after the species is detected, when the values switch to `1` (`beh_t3=t4=t5=t6=1`). Likewise, the covariate remains `0` for site 6 until after the 4th survey, when it changes to `1` for survey 5, and there is a missing value for survey 6 (denoted with 0.5).

To fit this 'behavioral' detection structure with constant occupancy, $\{\psi . p_{behav}\}$ you need to use the design matrix and it would look like this:

B1 p - intercept	Parm	B2 p - BehEffect	B3 psi - intercept
1	1:p	beh_t1	0
1	2:p	beh_t2	0
1	3:p	beh_t3	0
1	4:p	beh_t4	0
1	5:p	beh_t5	0
1	6:p	beh_t6	0
0	7:Psi	0	1

Hand-coding these survey-specific covariates for each site can get a little tedious, so **MARK** has a function that calculates the appropriate value for each site. '**PriorCapL**' indicates whether a species was previously captured or observed between any two surveys. For example, `priorcapl(i,j)` will return the value of '`0`' if the species was not previously captured on surveys $(i, i+1, i+2, ..., j)$, and '1' if the animal was captured during this set of surveys. `Priorcapl(1,1)` is valid – again returning '`0`' if the species was not detected on survey 1, and '1' if the species was detected. The function calculates the appropriate value (0 or 1) for each site in the sample, and uses it as a survey- and site-specific covariate.

An appropriate design matrix for the model $\{\psi . p_{\text{behav}}\}$ would look like this:

B1: p Int	Parm	B2: p-BAOW	B3: Psi Int
1	1:p	0	0
1	2:p	PriorCapL(1,1)	0
1	3:p	PriorCapL(1,2)	0
1	4:p	PriorCapL(1,3)	0
1	5:p	PriorCapL(1,4)	0
1	6:p	PriorCapL(1,5)	0
0	7:Psi	0	1

Real and derived parameter estimates (shown at the top of the next page) are based on the first encounter history, and this encounter history is printed with the real parameter estimates in the output file. In our case, the first detection history consists of all zeros, so only the initial detection probability is reported ($\hat{p}_{\texttt{initial}} = 0.46$).

NSO - single season

Real Function Parameters of {psi(.)p(behav)}
Estimates based on the following encounter history:
0000..

Parameter	Estimate	Standard Error	95% Confidence Interval Lower	Upper
1:p	0.4580948	0.0550142	0.3538001	0.5661959
2:p	0.4580948	0.0550142	0.3538001	0.5661959
3:p	0.4580948	0.0550142	0.3538001	0.5661959
4:p	0.4580948	0.0550142	0.3538001	0.5661959
5:p	0.4580948	0.0550142	0.3538001	0.5661959
6:p	0.4580948	0.0550142	0.3538001	0.5661959
7:Psi	0.6286818	0.0477388	0.5313948	0.7165449

However, the estimated effect is positive,

NSO - single season

LOGIT Link Function Parameters of {psi(.)p(behav)}

Parameter	Beta	Standard Error	95% Confidence Interval Lower	Upper
1:p Int	-0.1680150	0.2216133	-0.6023770	0.2663471
2:p-BAOW	0.9412031	0.2536470	0.4440549	1.4383512
3:Psi Int	0.5265659	0.2045007	0.1257445	0.9273873

Thus, the estimate of detection probability does *increase* following initial detection:

$$p_{\text{after}} = \frac{e^{(-0.168017+0.9412005)}}{1+e^{(-0.168017+0.9412005)}}$$
$$= 0.68,$$

but this model is not well supported relative to the detection structures that include the effect of barred owl ($\Delta\text{AIC}_c = 25.4$).

22.3.4. False positives

The fifth assumption implies that detections can only occur at occupied sites. If a detection is made at a given site, it is assumed to be true detection, which confirms (without error) the occupied status of that site. However, occupancy monitoring can be vulnerable to false positives, i.e., detections reported at sites that are not occupied by the species of interest. False positives can be the result of different types of processes. First, they can result from the misidentification of a similar species. This is relatively common in acoustic surveys (e.g., McClintock *et al*. 2010, Miller *et al*. 2012), especially when many individuals are calling (full choruses) or in the presence of background noises (e.g., wind) . Misidentification can also be a pervasive issue in visual surveys, especially when sister species live in sympatry (e.g., passerines, some lizards and freshwater mussels) or when monitoring larval or juvenile life stages that are hard to identify (e.g., tadpoles).

In some cases, false positives can originate from unreliable or untrustworthy observers. This is especially true when data are obtained from public surveys, where 'observers' (people that are interviewed) might have a personal interest in providing a desired answer (e.g., the presence of a rare species). A good example is provided by Pillay *et al*. (2014), where observations of large mammals in India where obtained by interviewing local people. For species involved in human conflicts (e.g., tigers), local informants were more likely to provide 'false positive' observations.

When not accounted for, false positives will systematically induce a positive bias in occupancy estimators. McClintock *et al*. (2010) and Miller *et al*. (2012) found that even small levels of false positive detections (as little as 1% of all detections) could cause severe overestimation of site occupancy probability. These studies, also found that the occurrence of false detections was more common than expected. For instance, an experiment (frog survey) by Miller *et al*. (2012) found that about 8% of all recorded detections were false positives, a rate that would induce severe biases. This issue should thus not be overlooked!

When designing an occupancy study, think about what sources could cause false detections and try to mitigate these. Mitigating false positives through your study design and the methods of detection you employ (e.g., visual vs. acoustic detection; animal signs vs. direct observations) is important. If false positives are still likely to occur in your dataset despite all your efforts, don't worry. Occupancy models have been extended to deal with this issue and correct the bias – see details in section (22.7).

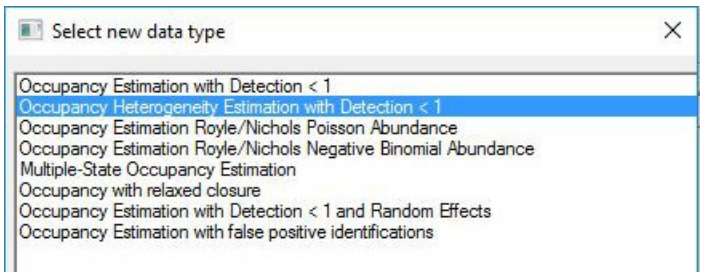
22.4. Unobserved detection heterogeneity

In this section we discuss methods that can be used to model heterogeneity in species detection probability without collected covariates.

22.4.1. Finite mixtures

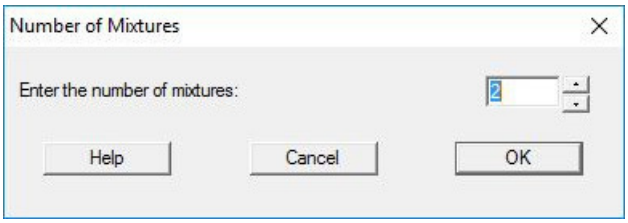
One of the banes of mark-recapture and occupancy modeling is individual or site-level heterogeneity in the detection process (see Chapter 15). In many cases, we have explicit hypotheses regarding why the probability of detection may vary across sites, which we model via individual covariates. However, there are also many times that site-level heterogeneity is unexplainable. We don't want to ignore this potential variation as it may lead to biased estimates of occupancy. There are several ways we can handle general heterogeneity in the single-season occupancy model when working in **MARK**. It should be noted that comparing models with different types of heterogeneity or between models with and without heterogeneity using AIC should be done with some caution. AIC depends on regularity conditions for properties of consistency ('*AIC will do the right thing*') which are not generally met with mixture models (Pledger & Phillpot 2008). There is also some question as to whether the same is true for comparing random effect models' integrated likelihood with model likelihoods without a random effect using AIC (G. C. White, *pers. comm.*). Whether AIC performs appropriately may depend on many factors, and is an issue which merits further study.

The first option to deal with unmodeled heterogeneity is the discrete- or finite-mixture approach (introduced in Chapter 15), which is the second data type listed when you select the Occupancy Estimation data type (labeled '**`Occupancy Heterogeneity with Detection <1`**').



After choosing this model type, the first thing we will see is a dialog box asking us how many *a priori*

mixtures we want to choose:



Generally, most data sets will only be able to accommodate 2 to 3 mixtures. The basic idea is that the number of mixtures corresponds to the number of ways the data can be split with different corresponding detection probabilities.

For two mixtures, the detection probability for site i (p_i) is,

$$p_i = \begin{cases} p_{i,A} & \text{with } \Pr(\pi) \\ p_{i,B} & \text{with } \Pr(1-\pi), \end{cases}$$

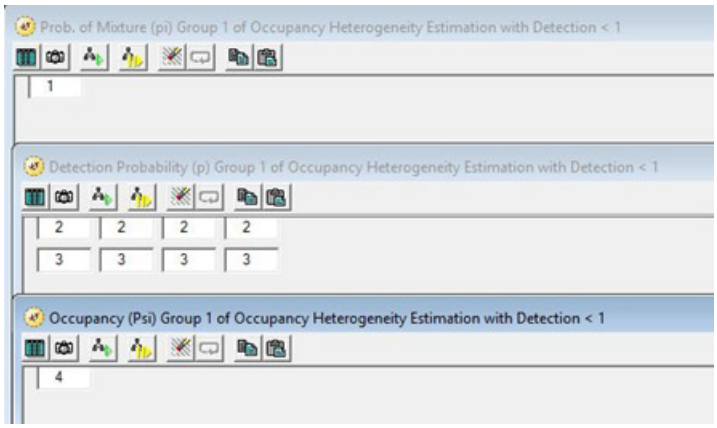
where A and B represent the two mixtures. Thus, for two mixtures, we need to estimate 3 parameters, $p_{i,A}$, $p_{i,B}$, and π. The parameter π represents the proportion of sites with detection probability $p_{i,A}$, while $1-\pi$ is the proportion of sites with detection probability $p_{i,B}$. What π really means, depends strongly on the underlying detection process, which may not be discrete at all; thus, not interpreting this parameter to mean anything is probably safest (see Chapter 15). For > 2 mixtures, there are additional π parameters that need to be estimated and constrained to sum to 1.

For the encounter history, $h_i =$ '1001', we would write our probability statement indicating 2 mixtures and no temporal variation as:

$$\Pr(h_i) = \text{'1001'} = \psi\left[\pi p_A^2\left(1-p_A^2\right) + (1-\pi)p_B^2\left(1-p_B^2\right)\right].$$

We know the site is occupied (ψ), but we don't know which detection probability is most appropriate (p_A, p_B), so we allow for both possibilities. Note that generally, we need 5 or more sampling occasions to fit heterogeneity models.

In **MARK**, we have three PIM's: ψ, π , and the p's.



The first row of the detection probability PIM corresponds to the p_A parameters, while the second row corresponds to the p_B parameters. Here, we specifying the model $\{\psi.\ \pi.\ p_{het-2mixtures}\}$, which corresponds to the same model definition used for the previous probability statement.

22.4.2. Random effects

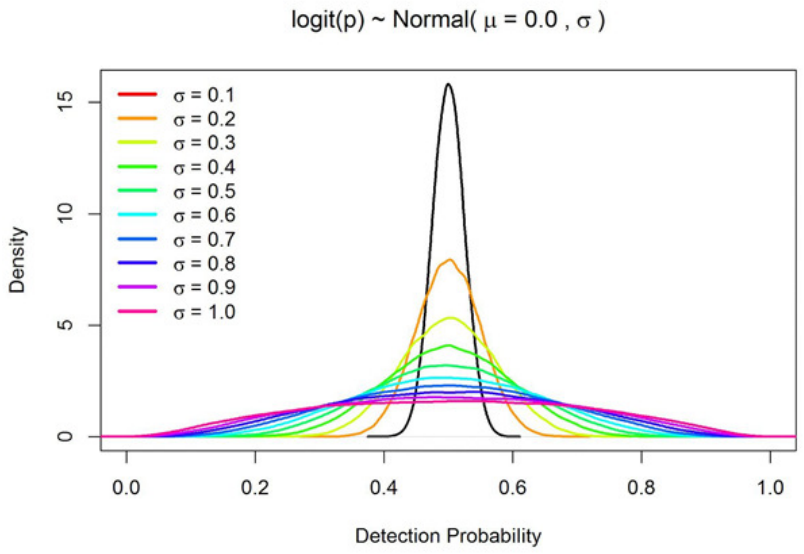
A different way to accommodate variation in the detection process is through a *random effect*. Instead of considering a number of different finite detection probabilities (mixtures), we might want to consider a distribution of detection probabilities with a mean μ and standard deviation σ. **MARK** now allows for random effects in many types of models, including the occupancy models (for an introduction to the theory, and a collation of which data types allow for random effects, see Chapter 15, and the addendum to that chapter). The simplest random effect model includes a single mean process (μ) and variation around this mean (σ). This model would look like,

$$\text{logit}(p_{ij}) \sim \mathcal{N}(\mu, \sigma).$$

A more complicated model might consider the mean process to vary by sampling occasions, such as,

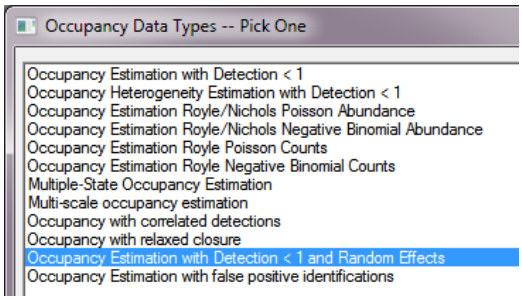
$$\text{logit}(p_{ij}) \sim \mathcal{N}(\mu_j, \sigma),$$

where we are estimating J means (μ), but only one σ . A figure might help in understanding how the standard deviation on the logit-scale influences the variation in the detection process:

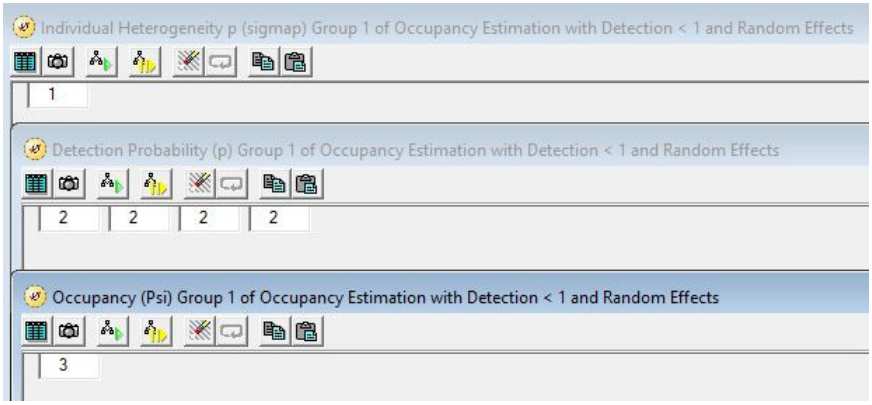


To fit random effects using likelihood theory, **MARK** integrates outs the random effect using *Gauss-Hermite numerical quadrature* (McClintock & White 2009, Gimenez and Choquet 2010, White & Cooch 2017 – also, see the introduction to this topic in Chapter 15), which provides an integrated likelihood. This likelihood can be maximized as normal to find our MLEs and variance-covariance matrix.

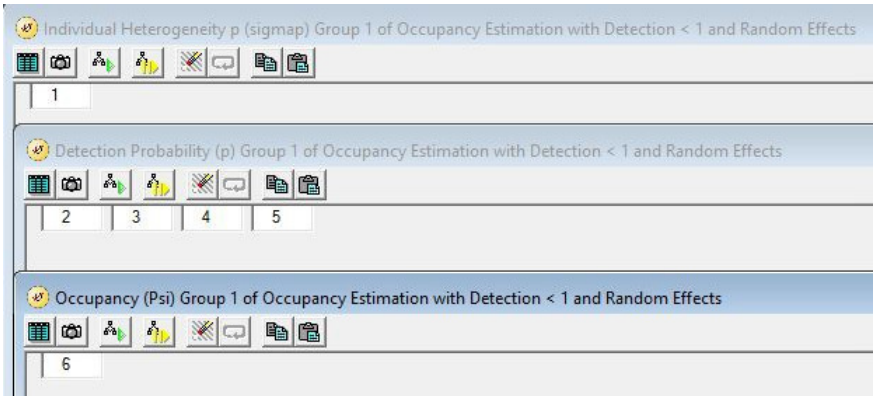
To fit this random effect model in **MARK**, first select the appropriate data type, '**`Occupancy Estimation with Detection <1 and Random Effects`**'.



Similar to the finite-mixture modeling approach (above), we have three PIM's: ψ, the mean detection probability (μ_p) for each sampling occasion j, and σ. In this PIM, we are specifying a single mean detection process (μ_p) with a measure of variation around this mean (σ_p):



If we want to consider a model where the mean changed for each sampling occasions, the PIM would look like:



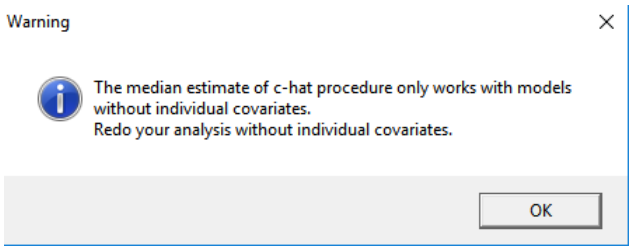
22.5. Goodness of fit

As with other models discussed in this book, occupancy models assume that at least one model in the candidate set provides adequate fit to the data (see Chapter 5). Few goodness of fit tests exist

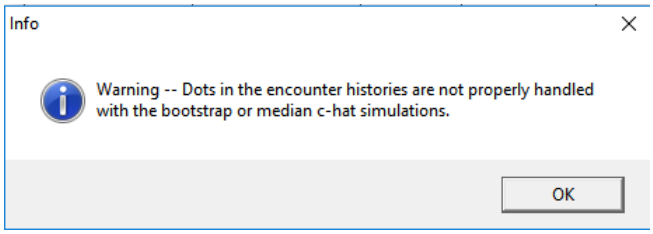
for occupancy models (see MacKenzie & Bailey 2004, MacKenzie *et al*. 2017) and the only method incorporated into program **MARK** is the median-$\hat{c}$ (section 5.7, White 2002). The conceptual motivation for this approach is covered in Chapter 5 and we refer readers to that chapter for more details. The median-$\hat{c}$ approach is well suited for detecting, and adjusting for, overdispersion that may result when the independence assumption is violated.

The typical approach, as discussed in Chapter 5, is to use the most general model structure to estimate the variance inflation factor, $\hat{c}$. Unfortunately, this is the most general structure without covariates, as there is no straightforward way to simulate data using covariates in **MARK**. Usually, this means that the most general structure is $\{\psi_{\cdot}\, p_t\}$ or if there are groups of sites the general structure may be $\{\psi_g\, p_{g*t}\}$.

Note, if we retrieve the model $\{\psi_{\cdot}\, p_t\}$ from our spotted owl analysis and attempt to perform a median-$\hat{c}$ analysis, we get the following message:



Unfortunately, if we reformat our data set and omit the covariates and attempt the process again, the following message appears:



MARK will still run the median approach but only using the detection histories without missing values. In the spotted owl case the method fails entirely because the observed deviance is negative (−179.27380), creating a negative observed $\hat{c}$.

In summary, few goodness-of-fit tests exist for occupancy models and we refer interested readers to MacKenzie *et al*. (2017) for a complete discussion of the existing methods.

22.6. The dynamic occupancy model

A natural extension of the basic occupancy model (single species, single season) is to link several 'seasons' together to investigate site-level dynamics. This allows researchers to focus on the processes that govern occurrence patterns, such as how and why occupied sites become extirpated (or conversely, persist) or how and why unoccupied sites become colonized. Examining these processes can help predict future patterns of species occurrence, often better than simply understanding occupancy-environment relationships (Yackulic *et al*. 2015). Linking data in this way is similar to Pollock's robust design used in mark-recapture studies (Pollock 1982), where seasons and surveys represent primary and secondary

periods, respectively (see Chapter 16). The assumptions (e.g., 'closure' during primary sampling periods – section 16.5) and complications (e.g., unequal intervals between primary sampling periods – section 16.11) are also largely equivalent.

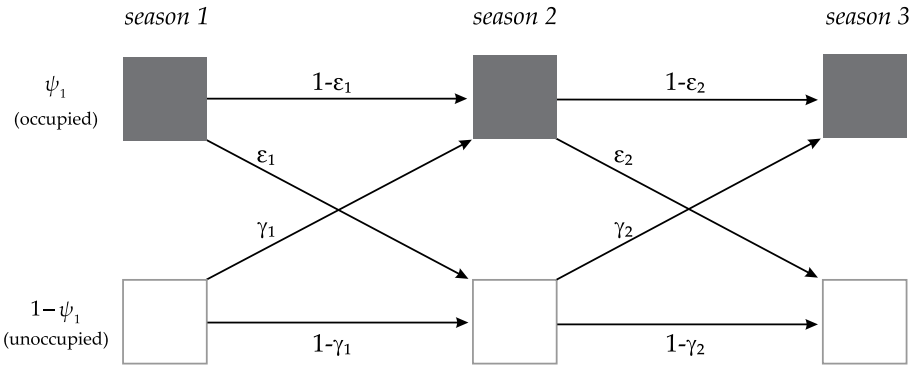
To model the dynamic changes in species occurrence species distribution from one season to the next, it is mathematically convenient and often biologically reasonable to do so based on a first-order Markov process. Simply, that the probability a site is occupied in season (t) depends on the occupancy state of the site in the previous season, $(t-1)$. Thinking about how this assumption matches the life history of the species of interest and the goal of the study may help in defining the season.

A first-order Markov dynamic occupancy model includes two additional parameters beyond that of occurrence ($\psi_{(t,i)}$) and detection ($p_{(t,i,j)}$),

γ_t = the probability that an unoccupied site in season (t) is occupied in season $(t+1) \Rightarrow$ *colonization*,

ϵ_t = the probability that an occupied site in season (t) is unoccupied in season $(t+1) \Rightarrow$ *extirpation*.

There are several alternative parameterizations of the dynamic occupancy model (MacKenzie *et al*. 2002), which are implemented in program **MARK**. Here, we follow the parameterization based on initial (season 1) occupancy (ψ_1) and time-specific extirpation and colonization. This model is illustrated in the following, where all possible site-level changes across three seasons are indicated:



A site is either occupied ($\psi_{(t,i)}$) or not ($1-\psi_{(t,i)}$) in each season (rows in the diagram). Changes in occupancy between seasons (columns) are based on the dynamic parameters of colonization (γ_t) and extirpation probability (ϵ_t).

While the parameterization we present here focuses on initial occupancy probability, occupancy in subsequent seasons can be derived using the recursive equation:

$$\psi_{t+1} = \psi_t(1-\epsilon_t) + (1-\psi_t)\gamma_t.$$

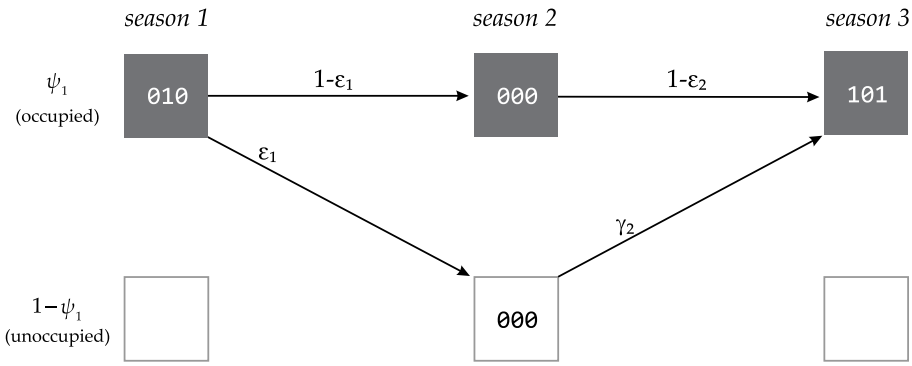
This equation estimates the probability of occupancy in subsequent seasons by starting with the initial occupancy estimate (ψ_1) and projecting that estimate forward using estimates of colonization and extirpation. Luckily, **MARK** computes this derived quantity of interest for us (shown in the example in the next section), complete with standard errors and confidence intervals.

The data for the dynamic occupancy model are the same as the basic model; a target species is either detected (`1`) or not detected (`0`) at site i on survey j. These detections and non-detections are replicated

at the same sites for seasons $t = 1, 2, \ldots, T$. Seasons do not need to have the same number of surveys and missing surveys can be easily accommodated by using a dot ('.'). An example detection history (h) for a study conducted for three seasons, with three surveys in each season is,

$$h_i = \text{'010 000 101'}$$

[To make the discrete seasons more obvious, we've added a space in the encounter history. In the **MARK** `.INP` file, however, there are no spaces in the history, which is a contiguous string of 1's and 0's.] We know the species occurs at this site in season 1 because it was detected on survey 2, but was not detected on surveys 1 or 3. During season 2, we don't know whether the species was extirpated and thus could not be detected, or whether the species persisted at the site and we simply did not detect the species on all three surveys. During season 3, we know the site was occupied because the species was detected on surveys 1 and 3, but went undetected on survey 2. Because we don't know whether the site was occupied during season 2, we also don't know whether the site remained occupied and thus persisted from season 2 to 3 or was unoccupied and thus was recolonized from season 2 to 3. We can diagram the possible pathways of site-level changes that could have happened as,



As with our single season models, we can link our parameters to our detection/non-detection data through probability statements. The probability of the observed detection history can be written as,

$$\begin{aligned}\Pr(h_i = \text{'010 000 101'}) = {} & \psi_1(1-p_{1,1})p_{1,2}(1-p_{1,3}) \\ & \times \left[(1-\epsilon_1)\prod_{j=1}^{3}(1-p_{2,j})(1-\epsilon_2) + \epsilon_1\gamma_2\right] \\ & \times p_{3,1}(1-p_{3,2})p_{3,3}.\end{aligned}$$

Starting on the first line of the equation right of the equal sign, we know the site was occupied in season 1 (ψ_1), because of the detection on survey 2 ($p_{1,2}$). Thus, we also know that we did not detect the species on surveys 1 and 3 ($(1 - p_{1,1})$ and $(1 - p_{1,3})$, respectively).

For season 2, we don't know whether the site was (i) occupied (i.e., no detections) or (ii) not occupied, and thus we need to represent each possible outcome in terms of both the dynamic and detection parameters; these probability outcomes are written on line 2 of the above equation. The first possibility is that the species persisted ($1 - \epsilon_1$) but was undetected. The second possibility is that the species was locally extirpated (ϵ_1). If the first possibility occurred, the species must have persisted ($1 - \epsilon_2$) from season 2 to 3, as we detected it in season 3. However, if the second possibility occurred, the site must have been recolonized after extirpation ($\epsilon_1\gamma_2$).

Finally, regardless of the pathway, we know the site was occupied in season 3 because of the detections on surveys 1 and 3 ($p_{3,1}$, $p_{3,3}$, respectively), though we did not detect the species on survey 2 ($1 - p_{3,2}$).

Similar to the basic occupancy model, after stating our data in terms of probability statements, we can link the statements together into a single likelihood, thus allowing us to estimate the unknown parameters:

$$\mathcal{L}(\psi_1, \epsilon_1, \epsilon_2, \gamma_1, \gamma_2, p_{1,1}, p_{1,2}, p_{1,3}, p_{2,1}, p_{2,2}, p_{2,3}, p_{3,1}, p_{3,2}, p_{3,3} \mid h_1, h_2, \ldots, h_n, N) = \prod_i^N \Pr(h_i).$$

22.6.1. Dynamic (multi-season) occupancy – an example

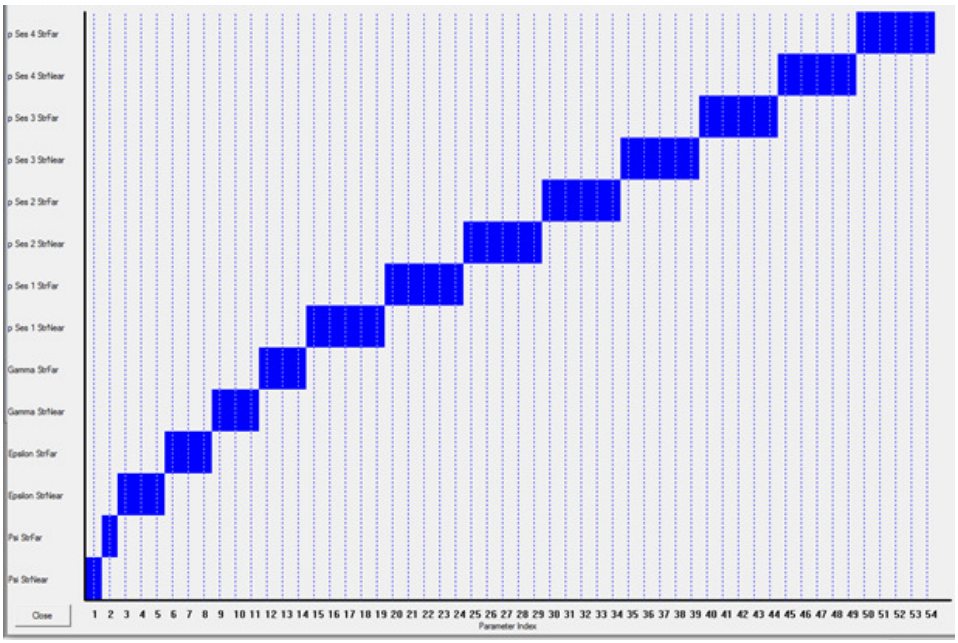
Let's explore the dynamic occupancy model using a 4-year dataset on Blue-ridge salamander (*Eurycea wilderae*) occurrence in the Appalachian Mountains of the eastern United States (the survey data are found in **blueridge_slamander.inp**). This species undergoes a seasonal migration from streams to upland terrestrial areas during summer months, and can often be found some distance from running water. Because *E. wilderae* occurs at low densities in terrestrial habitats, occupancy has been proposed as a state variable of interest for large-scale monitoring programs.

Thirty-nine sites were sampled from 1998-2001 (4 years). Sites were located near trails, with ≈250m between sites. Natural cover and cover boards were surveyed at each site. In the first year of the study, sites were sampled once per month from June-August for a total of three surveys. In subsequent years (1999-2001), every site was surveyed once every two weeks from April through late June for a total of five surveys in each of these seasons. Covariates measured at each site include elevation (a continuous covariate) and stream proximity (a binary value). Stream proximity values were categorized as two '**Groups**': `group 1` sites had a stream located within 50m of the site, and sites in `group 2` were not near streams. The researchers note that rainfall was variable from 1998-2001, and declined over the last 3 years of the study.

Create a new **MARK** project, and select '**Robust Design Occupancy**' from the list of available data types. You'll see that many variations of this model are implemented in **MARK**. The first three entries correspond to the standard dynamic occupancy model, and differ only in their parameterization. Different parameterizations may be selected depending on the study objectives, but we will use the third entry, '**Robust Design Occupancy with psi(1), gamma, epsilon**'. We typically use this parameterization because we find it to be the most mechanistic parameterization and because it is less prone to the numerical convergence issues that sometimes arise when using the other parameterizations.

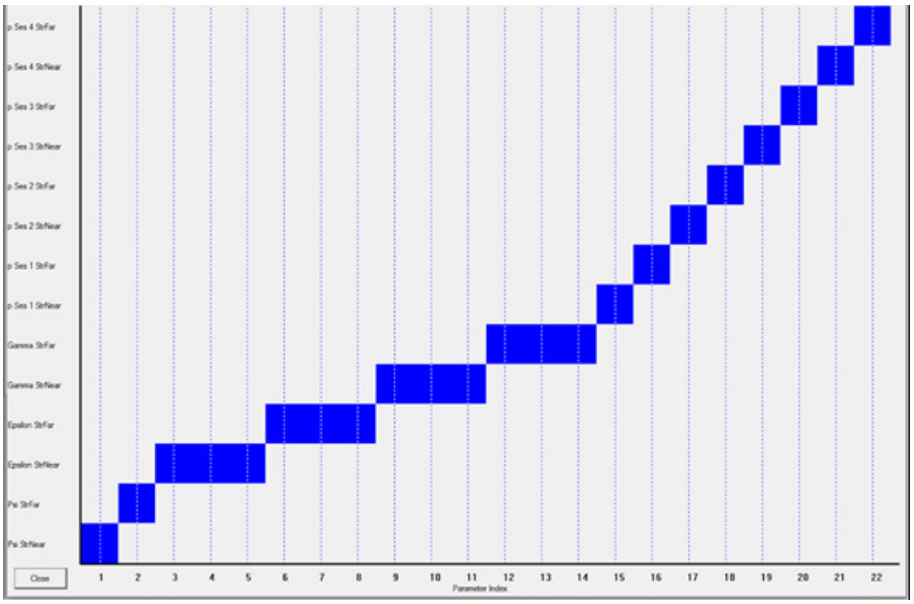
Select the **blueridge_salamander.inp** input file. There are 20 encounter occasions in this dataset (4 seasons × 5 surveys per season). In season 1 (1998), only 3 surveys were conducted so surveys 4 and 5 are coded as missing observations ('.'), but we could have included 3 entries in this season too; either approach is fine. Use the '**Easy Robust Design Times**' button, enter the number of seasons (termed primary occasions) and verify that the number of surveys in each season, or primary occasion, is 5. The dataset has two groups related to stream proximity, as described previously; we have labeled these groups '`StrmNear`' and '`StrmFar`'. Finally, we have one individual covariate '`elev`'. Label these groups and covariates using the buttons by each entry.

Open the PIM chart (shown at the top of the next page). The parameters listed on the y-axis include occupancy in the first year of the study (1998), local extirpation (`Epsilon`), local colonization (`Gamma`), and detection probability for each season (`p`). The default model shown has an initial occupancy for each group (`StrmNear`, `StrmFar`), colonization and extirpation varying by season (i.e., year) and group, and p varying by both season and survey in addition to group. We might name this model: '`Psi(g) Epsilon(g*t) Gamma(g*t) p(g*t)`'.



Run this model so that we have this PIM structure saved, but note that this model might be difficult to interpret biologically.

Let's explore a small set of biologically relevant hypotheses. To keep this example simple, we'll assume that detection probability is constant for all surveys in all seasons, but we will model initial occurrence, colonization, and extirpation as functions of covariates of interest. Let's run a 'null' model first, where all parameters are assumed to be constant. We'll do this in the design matrix, though a constant model could also be fit using the PIMs. Before moving to the design matrix, we'll simplify the PIM structure for detection. We will assume that detection probability is the same for all surveys within a year (primary occasion) for each group, so we can right-click on each detection block and select '**`Constant`**'.



Now let's open a reduced design matrix with 4 parameters (betas) by clicking '**Design | Reduced**', and entering 4. In this model, we will constrain groups to be equivalent for each type of parameter (initial occupancy, extirpation, colonization and detection probability).

B1:	B2:	Parm	B3:	B4:
1	0	1:Psi	0	0
1	0	2:Psi	0	0
0	1	3:Epsilon	0	0
0	1	4:Epsilon	0	0
0	1	5:Epsilon	0	0
0	1	6:Epsilon	0	0
0	1	7:Epsilon	0	0
0	1	8:Epsilon	0	0
0	0	9:Gamma	1	0
0	0	10:Gamma	1	0
0	0	11:Gamma	1	0
0	0	12:Gamma	1	0
0	0	13:Gamma	1	0
0	0	14:Gamma	1	0
0	0	15:p Session 1	0	1
0	0	16:p Session 1	0	1
0	0	17:p Session 2	0	1
0	0	18:p Session 2	0	1
0	0	19:p Session 3	0	1
0	0	20:p Session 3	0	1
0	0	21:p Session 4	0	1
0	0	22:p Session 4	0	1

This model has 4 parameters (`B1` through `B4`) and assumes that initial occupancy, colonization, extirpation, and detection are all constant across sites: (`Psi(1998) Epsilon(.) Gamma(.) p(.)`). We label occupancy as being for 1998 to remind ourselves that we are using the initial occupancy parameterization and that occupancy estimated is for the first year of the study (1998 in this case).

Suppose biologists are interested in whether initial occupancy, colonization, and extirpation differ between sites that are near or far from streams. Again, we could build this model using PIMs, or in the design matrix, because it does not use individual covariates. We will build it in the design matrix (which is, ultimately, more flexible than using PIMs):

B1:	B2:	B3:	Parm	B4:	B5:	B6:	B7:
1	0	0	1:Psi	0	0	0	0
1	1	0	2:Psi	0	0	0	0
0	0	1	3:Epsilon	0	0	0	0
0	0	1	4:Epsilon	0	0	0	0
0	0	1	5:Epsilon	0	0	0	0
0	0	1	6:Epsilon	1	0	0	0
0	0	1	7:Epsilon	1	0	0	0
0	0	1	8:Epsilon	1	0	0	0
0	0	0	9:Gamma	0	1	0	0
0	0	0	10:Gamma	0	1	0	0
0	0	0	11:Gamma	0	1	0	0
0	0	0	12:Gamma	0	1	1	0
0	0	0	13:Gamma	0	1	1	0
0	0	0	14:Gamma	0	1	1	0
0	0	0	15:p Session 1	0	0	0	1
0	0	0	16:p Session 1	0	0	0	1
0	0	0	17:p Session 2	0	0	0	1
0	0	0	18:p Session 2	0	0	0	1
0	0	0	19:p Session 3	0	0	0	1
0	0	0	20:p Session 3	0	0	0	1
0	0	0	21:p Session 4	0	0	0	1
0	0	0	22:p Session 4	0	0	0	1

We might call this model: '`Psi(strm) Epsilon(strm) Gamma(strm) p(.)`'.

The researchers also wonder if the decreasing rainfall over the course of the study might influence colonization and extirpation parameters. Though we don't have rainfall as an annual covariate, we can model colonization and extirpation as a function of a monotonic or linear trend (designated '`T`' in model names) using the design matrix: '`Psi(1998) Eps(T) Gam(T) p(.)`'.

B1: Psi - 1998	B2: Eps - int	B3: Eps - T	Parm	B4: Gam - int	B5: Gam - T	B6: p
1	0	0	1:Psi	0	0	0
1	0	0	2:Psi	0	0	0
0	1	1	3:Epsilon	0	0	0
0	1	2	4:Epsilon	0	0	0
0	1	3	5:Epsilon	0	0	0
0	1	1	6:Epsilon	0	0	0
0	1	2	7:Epsilon	0	0	0
0	1	3	8:Epsilon	0	0	0
0	0	0	9:Gamma	1	1	0
0	0	0	10:Gamma	1	2	0
0	0	0	11:Gamma	1	3	0
0	0	0	12:Gamma	1	1	0
0	0	0	13:Gamma	1	2	0
0	0	0	14:Gamma	1	3	0
0	0	0	15:p Session 1	0	0	1
0	0	0	16:p Session 1	0	0	1
0	0	0	17:p Session 2	0	0	1
0	0	0	18:p Session 2	0	0	1
0	0	0	19:p Session 3	0	0	1
0	0	0	20:p Session 3	0	0	1
0	0	0	21:p Session 4	0	0	1
0	0	0	22:p Session 4	0	0	1

Finally, let's run a model where initial occupancy, colonization, and extirpation probabilities vary with both `Group` (i.e., stream distance) and `elevation`, in an additive fashion – we'll call this model '`Psi(strm+elev) Eps(strm+elev) Gam(strm+elev)p(.)`'.

B1: Psi - Near	B2: Psi - Far	B3: Psi - elev	B4: Eps - Near	B5: Eps - Far	Parm	B6: Eps - elev	B7: Gam - Near	B8: Gam - Far	B9: Gam - elev	B10: p
1	0	elev	0	0	1:Psi	0	0	0	0	0
1	1	elev	0	0	2:Psi	0	0	0	0	0
0	0	0	1	0	3:Epsilon	elev	0	0	0	0
0	0	0	1	0	4:Epsilon	elev	0	0	0	0
0	0	0	1	0	5:Epsilon	elev	0	0	0	0
0	0	0	1	1	6:Epsilon	elev	0	0	0	0
0	0	0	1	1	7:Epsilon	elev	0	0	0	0
0	0	0	1	1	8:Epsilon	elev	0	0	0	0
0	0	0	0	0	9:Gamma	0	1	0	elev	0
0	0	0	0	0	10:Gamma	0	1	0	elev	0
0	0	0	0	0	11:Gamma	0	1	0	elev	0
0	0	0	0	0	12:Gamma	0	1	1	elev	0
0	0	0	0	0	13:Gamma	0	1	1	elev	0
0	0	0	0	0	14:Gamma	0	1	1	elev	0
0	0	0	0	0	15:p Session 1	0	0	0	0	1
0	0	0	0	0	16:p Session 1	0	0	0	0	1
0	0	0	0	0	17:p Session 2	0	0	0	0	1
0	0	0	0	0	18:p Session 2	0	0	0	0	1
0	0	0	0	0	19:p Session 3	0	0	0	0	1
0	0	0	0	0	20:p Session 3	0	0	0	0	1
0	0	0	0	0	21:p Session 4	0	0	0	0	1
0	0	0	0	0	22:p Session 4	0	0	0	0	1

Take a look at the results browser with the 4 models we have run. The model where initial occupancy, colonization, and extirpation vary among sites with and without a stream nearby (among our two groups) is best-supported (AIC_c weight = 0.91).

Model	AICc	Delta AICc	AICc Weight	Model Likelihood	No. Par.	Deviance	-2Log(L)
{psi1(strm)epsilon(strm)gamma(strm)p(.)}	690.9792	0.0000	0.91126	1.0000	7	676.2174	676.2174
{psi1(strm+elev)epsilon(strm+elev)gamma(strm+elev)p(.)}	695.7028	4.7236	0.08589	0.0943	10	674.1750	674.1750
{psi1(1998)epsilon(.)gamma(.)p(.)}	702.7813	11.8021	0.00249	0.0027	4	694.5146	694.5146
{psi1(1998)epsilon(T)gamma(T)p(.)}	706.6522	15.6730	0.00036	0.0004	6	694.0846	694.0846

If we examine the real parameter estimates from that model, we see that initial occupancy is 1.0 for sites near streams, but occupancy a bit lower at sites further from streams ($\psi = 0.92$). Differences in extirpation and colonization probabilities are even more pronounced for the two types of sites. Sites near streams are less likely to go extinct and are more likely to be colonized than sites far from streams.

Blue-ridge Salamander

Real Function Parameters of {psi1(strm)epsilon(strm)gamma(strm)p(.)}

Parameter	Estimate	Standard Error	95% Confidence Interval Lower	Upper
1:Psi	1.0000000	0.0000000	1.0000000	1.0000000
2:Psi	0.9181377	0.1170845	0.3461629	0.9958088
3:Epsilon	0.0912805	0.0673260	0.0200562	0.3302082
4:Epsilon	0.0912805	0.0673260	0.0200562	0.3302082
5:Epsilon	0.0912805	0.0673260	0.0200562	0.3302082
6:Epsilon	0.5089388	0.0952613	0.3293053	0.6862933
7:Epsilon	0.5089388	0.0952613	0.3293053	0.6862933
8:Epsilon	0.5089388	0.0952613	0.3293053	0.6862933
9:Gamma	0.7462327	0.4587374	0.0248587	0.9970606
10:Gamma	0.7462327	0.4587374	0.0248587	0.9970606
11:Gamma	0.7462327	0.4587374	0.0248587	0.9970606
12:Gamma	0.1810863	0.0818624	0.0697213	0.3948355
13:Gamma	0.1810863	0.0818624	0.0697213	0.3948355
14:Gamma	0.1810863	0.0818624	0.0697213	0.3948355
15:p Session 1	0.3857662	0.0277264	0.3330348	0.4413222
16:p Session 1	0.3857662	0.0277264	0.3330348	0.4413222
17:p Session 2	0.3857662	0.0277264	0.3330348	0.4413222
18:p Session 2	0.3857662	0.0277264	0.3330348	0.4413222
19:p Session 3	0.3857662	0.0277264	0.3330348	0.4413222
20:p Session 3	0.3857662	0.0277264	0.3330348	0.4413222
21:p Session 4	0.3857662	0.0277264	0.3330348	0.4413222
22:p Session 4	0.3857662	0.0277264	0.3330348	0.4413222

In addition to knowing the initial occupancy, colonization, and extirpation probabilities researchers and managers will often be interested in seasonal estimates of occupancy. **MARK** derives the probability that a site is occupied in each season using estimates of initial occupancy along with the extirpation and colonization probabilities using the recursive equation described earlier in this section. These estimates are accessed using the '**Derived Parameters**' tab of the results browser.

Blue-ridge Salamander

Estimates of Derived Parameters
Psi Estimates of {psi1(strm)epsilon(strm)gamma(strm)p(.)}

Grp.	Occ.	Psi-hat	Standard Error	95% Confidence Interval Lower	Upper
1	1	1.0000000	0.0000000	1.0000000	1.0000000
1	2	0.9087195	0.0673260	0.7767606	1.0406784
1	3	0.8938876	0.0737375	0.7493621	1.0384131
1	4	0.8914776	0.0790150	0.7366082	1.0463470
2	1	0.9181377	0.1170853	0.6886505	1.1476248
2	2	0.4656860	0.0898560	0.2895683	0.6418037
2	3	0.3254373	0.0750071	0.1784233	0.4724512
2	4	0.2819637	0.0769687	0.1311051	0.4328223

Lambda Estimates of {psi1(strm)epsilon(strm)gamma(strm)p(.)}

Grp.	Occ.	Lambda-hat	Standard Error	95% Confidence Interval Lower	Upper
1	1	0.9087195	0.0673260	0.7767606	1.0406784
1	2	0.9836783	0.0448917	0.8956905	1.0716660
1	3	0.9973039	0.0153566	0.9672049	1.0274030
2	1	0.5072071	0.0997791	0.3116401	0.7027741
2	2	0.6988342	0.0939774	0.5146384	0.8830300
2	3	0.8664149	0.1033640	0.6638214	1.0690084

Lambda' Estimates of {psi1(strm)epsilon(strm)gamma(strm)p(.)}

Grp.	Occ.	Lambda'-hat	Standard Error	95% Confidence Interval Lower	Upper
1	1	0.1147600E-006	0.0000000	0.1147600E-006	0.1147600E-006
1	2	0.8461845	0.3867605	0.0881338	1.6042351
1	3	0.9751565	0.1339221	0.7126692	1.2376439
2	1	0.0777093	0.1157558	-0.1491721	0.3045907
2	2	0.5535392	0.1195507	0.3192198	0.7878586
2	3	0.8139577	0.1340890	0.5511432	1.0767722

The first estimates listed are the annual occupancy estimates for 1998-2001 for each group of sites. As you can see, seasonal occupancy is declining regardless of stream proximity, but the decline is more pronounced for sites that are far from streams (Group 2). One note of caution – when colonization and extirpation probabilities are constant, as in this model, a stationary Markov process is assumed where occupancy will trend toward an equilibrium value. This value (sometimes noted ψ_{EQ}) is simply $\gamma/(\gamma+\epsilon)$ and may be of interest to researchers and biologists as it can be compared to current occupancy estimates (as in Farris *et al*. 2017, Figure 3). ψ_{EQ} can be thought of as the projected future equilibrium state, assuming no change in colonization or extirpation. The validity of this assumption depends on the biology of the system and of the investigator's expectations. Using our output from the best model, the equilibrium occupancy estimate for sites near streams is 0.89 (= $0.75/(0.75+0.09)$) and for sites far from streams is 0.26 (= $0.18/(0.18+0.51)$).

The next set of derived parameter estimates are for λ, or the rate of change in occupancy. This derived estimate is conceptually comparable to the familiar growth rate (λ) associated with population models. λ is the ratio of annual occupancy probabilities ($\psi_{(t+1)}/\psi_t$); values of $\lambda > 1$ indicate an increase in occupancy probability while values < 1 indicate a decrease. This expression is another case of 'odds' – where we compare one type of proportion to another. Comparisons of λ can be difficult because 5% growth for an initial occupancy of 0.9 is a fundamentally different absolute change than the same growth when initial occupancy is 0.2. For this reason, we also may be interested in an odds ratio, which **MARK** calls λ' ('lambda prime').This metric is a ratio of the odds of occupancy at time (t) versus at time ($t+1$) : $\lambda' = (\psi_{(t+1)}/1-\psi_{(t+1)})/(\psi_t/1-\psi_t)$. It can be interpreted as the amount that the odds of occupancy at time (t) would be multiplied by to get the odds of occupancy at time ($t+1$). For example, the final estimate in the table lists the odds of a Group 2 site being occupied in 2001 compared to in 2000. The estimate of λ' is 0.81, so non-stream sites in 2001 were only 0.81 times as likely to be occupied as those in 2000. The ratio can also be inverted (1/0.81) to make a statement that is easier to interpret: non-stream sites were 1.2 times more likely to be occupied in 2000 than in 2001.

22.7. The 'false positive' (misidentification) occupancy models

Occupancy models were initially developed to account for a one-way directional bias; species could remain undetected at sites where they are present (i.e., false negatives; Bailey *et al*. 2014, MacKenzie *et al*. 2017). This was a great improvement from the use of logistic or probit regression, which ignored this important potential source of bias. However, as discussed in section (22.3.4), occupancy models rely on the assumption, among others, that the species of interest cannot be erroneously detected at a site where it is not present. In other words, all of the occupancy models we have described so far do not allow for

false positive detections. But, as discussed in section (22.3.4), there are situations where false positive detections might occur (e.g., species misidentification). When this happens, severe biases in occupancy estimators will be induced (McClintock *et al*. 2010, Miller *et al*. 2011) unless accounted for.

What can we do about it? If possible, our best choice will always be to avoid false positive detections using design-based solutions. We could, for instance, change our detection method and use one that is more reliable. That's easy to say, but not always easy to do, especially when we start considering logistical and cost/effort constraints. If getting rid of the risk of false positives is not an option, then we will need to account for them in our modeling. Luckily, statistical developments have included models to handle datasets that contain false detections, which are now implemented in **MARK**. If false positives are a concern in your study system, just keep reading.

Before we move forward, it is worthwhile to re-emphasize the importance of avoiding false positives in the first place when possible. Why? Accounting for false positives through modeling increases model complexity (number of parameters), which makes the analysis more data hungry. In other words, it decreases the value of your dataset, rather significantly (Clement 2016). So, depending on the richness of the data, you cannot guarantee that you will obtain 'good estimates' (i.e., precise enough to be useful) when you are constrained to using the false positive models we are about to present. This is something to bear in mind. But, if despite your best efforts and intentions your dataset contains false positives, then you should use the models described next. Why? To avoid severe biases in your occupancy estimates.

First we describe the static (single-season) version of the false positive occupancy model. Then, we provide an example of the multi-season false positive model, which is straightforward now that you are familiar with the multi-season occupancy model without false positives. Let's redefine a false positive (or false detection) in probabilistic terms and in relation to the occupancy and observation processes. To do so, let's define z_i, as the true occupancy state at site i; a site is either occupied ($z_i = 1$) or unoccupied ($z_i = 0$). We use z to denote the latent (unobserved) occupancy state to make it clear that false positive detections can only occur when the species is absent. In the earlier models we presented, we considered that when a site i was unoccupied ($z_i = 0$), the only possible observational outcome at any occasion j was nondetection, $y_{ij} = 0$. Thus, we considered that when a species was observed on at least one survey ($\sum_{j=1}^{J} y_i > 0$), the site is occupied with certainty ($\Pr(z_i = 1) = 1$ and, by complementarity, $\Pr(z_i = 0) = 0$). In other words, we assume it is impossible to get a detection at an unoccupied site.

Now that we are considering the possibility of false positives, this is not true anymore. What does that change in terms of our modeling? It adds another stochastic process. If you recall, from section (22.1), we described the basic occupancy model with two stochastic processes: (1) the occupancy process, defined by probability ψ; and (2) the detection process, defined by p, which was strictly conditional on $z_i = 1$. With false positives, we must add a stochastic detection process for unoccupied sites ($z_i = 0$). Before, we did not need this process, because when $z_i = 0$, the outcome was deterministic ($y_{ij} = 0$). This means that we now have two detection parameters. One for sites that are occupied, termed the true detection probability, which we will refer to as p_{11} (subscript '11' indicating that both the observed and true state are occupied) and one for sites that are unoccupied, termed the false detection probability, which we will refer to as p_{10} (indicating that the observed state [1; occupied] is different from the true state [0; unoccupied]). This notation (p_{yz}) is straightforward and it is the one used in several of the main papers describing false positive occupancy models (Royle & Link 2006, Miller *et al*. 2011, Chambert *et al*. 2015). An important remark, though: here, these subscripts do not refer to a specific survey j. If we need to add variation among surveys j , we will use a comma to separate yz from j, such as: $p_{(yz,j)}$. For instance, the false detection probability at survey 3, would be written as $p_{10,3}$.

We now have three stochastic processes and three basic parameter types: ψ, p_{11} and p_{10} (we will, of course have more parameters as we include covariates, etc.). The next step is to link these parameters to the data through probability statements. To illustrate this, let's imagine the same simple scenario we used in section (22.1). We surveyed site i twice ($J = 2$) and observed the following encounter history:

h_i = '01'. There are two possibilities for this history: (1) the site is truly occupied (ψ) and we failed to detect the species on survey ($1 - p_{11}$), but succeeded on survey 2 (p_{11}); or (2) the site is unoccupied ($1 - \psi$) and we did not falsely detect the species on survey 1 ($1 - p_{10}$), but we made a false detection on survey 2 (p_{10}). The corresponding probability statement (including subscripts for surveys) is:

$$\Pr(h_i = \text{'01'}) = \left[\psi(1 - p_{11,1})p_{11,2}\right] + \left[(1 - \psi)(1 - p_{10,1})p_{10,2}\right].$$

Each block within brackets (on each side of the '+' sign) corresponds to one of the two possibilities we just described verbally. On the left is the piece dealing with true detections; on the right, the piece dealing with false detections. In the table below, we provide the probability statements (omitting survey-specific subscripts, for the sake of simplicity) for all possible encounter histories when $J = 2$.

history, h_i	*probability expression*
11	$\psi p_{11} p_{11} + (1 - \psi) p_{10} p_{10}$
10	$\psi p_{11}(1 - p_{11}) + (1 - \psi) p_{10}(1 - p_{10})$
01	$\psi(1 - p_{11}) p_{11} + (1 - \psi)(1 - p_{10}) p_{10}$
00	$\psi(1 - p_{11})(1 - p_{11}) + (1 - \psi)(1 - p_{10})(1 - p_{10})$

As you can see in the table, there is always an exact symmetry, on each side of the '+' sign, between the true and the false detections' pieces of the probability statements. In this simplest version of the false positive model, first described by Royle & Link (2006), this symmetry exists for all possible encounter histories. OK, but is it an issue? Yes! This symmetry in the likelihood creates a situation where there are always two equally optimal solutions (i.e., values that maximize the likelihood) for all parameters. For instance, if $\psi = 0.3, p_{11} = 0.6$ and $p_{10} = 0.1$ are maximum likelihood estimates (MLEs), then by symmetry, $\psi = 0.7, p_{11} = 0.4$ and $p_{10} = 0.9$ will also be MLEs (Note: the correspondence between the two is $1 - \theta$, where θ represents any of the model parameters).

With two equal optimums, how do we know which optimum is 'correct'? In other terms, which of these solutions provides the best estimates, given the data at hand? In the example just used, we see that the 'best estimate' of occupancy, given our data, could be either 0.3 or 0.7. Not really satisfying, right?

To solve this issue, we need additional information: either as an additional assumption or as data. In their seminal paper, Royle & Link (2006) solved this issue by imposing the constraint (or assumption) that $p_{11} > p_{10}$. In other words, they assume the probability of true detection is higher than the probability of false positive detection, which ensures a unique maximum of the likelihood. Consider the example we used a few lines above:

$$\{\psi = 0.3, p_{11} = 0.6 \text{ and } p_{10} = 0.1\} \text{ vs. } \{\psi = 0.7, p_{11} = 0.4 \text{ and } p_{10} = 0.9\}.$$

With this new constraint ($p_{11} > p_{10}$), we can easily see that the only possible MLE is now solution 1 (left hand side), as $0.6 > 0.1$ is consistent with the assumption, but $0.4 < 0.9$ is not. Although this assumption might sometimes be relevant, it is usually preferable to rely on data to solve the symmetry issue and make our model identifiable.

Miller *et al*. (2011) developed such a model and it is implemented in **MARK**. To utilize this model, our dataset needs to contain some detections that we know for sure are true detections (i.e., not false positives). So, our dataset must consist of two types of detections: (i) ambiguous (or uncertain) detections, which can either be true or false detections; and (ii) unambiguous (or certain) true detections,

which cannot be false positives. In their seminal work, Miller *et al*. (2011) considered two different ways of obtaining unambiguous and ambiguous detections, and thus, provided two different models: (1) the 'two detection states' model and (2) the 'two detection methods' model. In the first scenario (two detection states), ambiguous and unambiguous detection correspond to two different types of observations that can be made during any survey. They are thus treated as two different observation states, one being certain, the other uncertain.

For instance, when monitoring a large predator (e.g., lynx), one might consider indirect cues (e.g., scat or tracks) as being ambiguous detections (e.g., as tracks can be confounded with those of another species), but direct observations of the animal (i.e., directly seeing a lynx) to be unambiguous. In the second scenario (two detection methods), one detection method (M1), implemented on some survey occasions, provides unambiguous detections, while another method (M2), implemented on different survey occasions, provides uncertain detections. For instance, some amphibian monitoring use both aural and visual surveys. Often, aural surveys are prone to misidentification, while visual surveys provide detections that are totally unambiguous. Only the 'two detection states' model is implemented in **MARK**, but don't worry, we can easily analyze the data obtained from a 'two detection methods' study design with this same model, by using a simple trick. We will explain this below. But first, let's start with the 'two detection states' model.

The two different types of detections are considered as different states. So, they must be distinguished (i.e., coded differently) in the input data. In **MARK**, unambiguous (or certain) detections are coded as '2', while ambiguous (or uncertain) detections are coded a '1'. As always, a non-detection is '0'. The detection probabilities, for each of the 3 observation states (y), conditional on the true site occupancy status (z) look like this:

true state	$y = 0$	$y = 1$	$y = 2$
$z = 0$	$1 - p_{10}$	p_{10}	0
$z = 1$	$1 - p_{11}$	$(1 - b)p_{11}$	bp_{11}

Here, the definition of p_{11} and p_{10} is pretty much the same as before: p_{11} [p_{10}] is the probability of a detection given that the site is occupied [unoccupied].

Now that we have two types of detections (1's and 2's), we need an additional parameter: b. Parameter b represents the probability that, given the site is occupied and a detection is made, that the detection is unambiguous. By complementarity $(1 - b)$ is the probability of an ambiguous detection. This parameter only applies to occupied site detections, because by definition, any detection made at an unoccupied site is a false positive, and can only be an unambiguous detection.

This model has four basic parameters: ψ, p_{11}, p_{10} and b. Let's consider the following encounter history: h_i = '201'. Here, there were three surveys; the species was detected with certainty on survey 1, it was not detected on survey 2, and on survey 3, it was detected, but the observation was ambiguous. Because the first detection was unambiguous, we know that this site was occupied. The probability statement for this encounter history is:

$$Pr(h_i = \text{'201'}) = \psi b p_{11,1}(1 - p_{11,2})(1 - b)p_{11,3}.$$

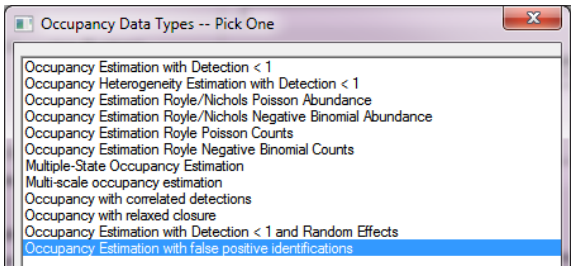
Now let's consider a similar encounter history, except that the first detection was ambiguous: h_i = '101'. Now, each of the two detections made could be either be a true detection or a false positive, so we need to consider the case of an unoccupied site, too. The probability statement is:

$$Pr(h_i = \text{'101'}) = \psi(1 - b)p_{11,1}(1 - p_{11,2})(1 - b)p_{11,3} + (1 - \psi)p_{10,1}(1 - p_{10,2})p_{10,3}.$$

Hopefully, this seems straightforward. Let's now illustrate how to implement this model in **MARK**. We will start with the single-season model. For the sake of simplicity, we will not consider covariates here, but adding covariate effects is done in the same way as for the other occupancy models that were present earlier in this chapter.

22.7.1. False positive single-season occupancy model in MARK

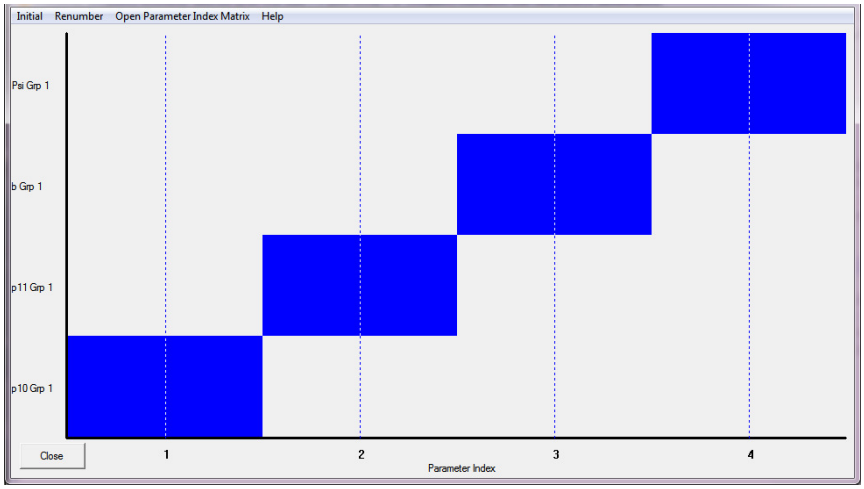
To illustrate model implementation in **MARK**, we will be analyzing data that we simulated, so we know the true value of each parameter. Start a new **MARK** session. In the '**Data Type**' window, select '**Occupancy Estimation**' and choose '**Occupancy Estimation with false positive identifications**':



Then, select the file called **FP-Example-Constant.inp** and view it. Next, specify that the data consist of 6 occasions, only 1 group and no individual covariates. Click '**OK**'.

This dataset was simulated using the following (real) parameter values: $\psi = 0.6, p_{11} = 0.65, p_{10} = 0.20, b = 0.15$. There is no time variation, and no covariate effects. Data were simulated for 100 hypothetical sites. Let's build some models and see if we can get the same estimates back.

First let's build the most simple model: the 'dot' model, where all parameters are assumed constant: $\{\psi(.), p_{11}(.), p_{10}(.), b(.)\}$. This is model is actually consistent with the 'true' process model used to simulate data, so we should expect this model to do well (both in terms of model support [AIC] and parameter estimates). Open the PIM chart. You can see the four types of parameters: `psi`, `b`, `p11` and `p10`. To build the constant model, right-click on each blue-box and select '**all constant**'. You should now have only 4 parameters left and the chart should look like this:



Next, '**Run the model**'. You can keep all default settings. In the '**Results**' browser, you should see an AIC_c value of ≈ 977. Here are the real parameter estimates we get:

FP demo

Real Function Parameters of {psi(.)p11(.)p10(.)b(.)}

Parameter	Estimate	Standard Error	95% Confidence Interval Lower	Upper
1:p10	0.1881563	0.0386482	0.1236908	0.2756516
2:p11	0.6726268	0.0369751	0.5965173	0.7406221
3:b	0.1574523	0.0264851	0.1121853	0.2165300
4:Psi	0.5507896	0.0751892	0.4032919	0.6898656

Now let's try to run some more complex models and see how they compare to this one, in terms of model support (AIC). Let's run models where p_{11} and p_{10} vary over time. For the three possible models, we get the following AIC_c values:

model	AIC
$\psi(\cdot)\, p_{11}(t)\, p_{10}(\cdot)\, b(.)$	980.98
$\psi(\cdot)\, p_{11}(\cdot)\, p_{10}(t)\, b(\cdot)$	980.85
$\psi(\cdot)\, p_{11}(t)\, p_{10}(t)\, b(\cdot)$	986.44

These three models receive less support than the constant model ($AIC = 977.01$), so we would conclude that there is no evidence for time variation in p_{11} or p_{10}. And indeed, we know this is true as we simulated data with constant parameters.

begin sidebar

Analyzing data collected with two different methods

As mentioned above, the two different types of detections (ambiguous and unambiguous) could be obtained from different detection methods (e.g., visual vs. aural surveys), each method represents different surveys. The last part of the sentence is important. Indeed, if the two methods are implemented, conjointly, at every sampling occasion, then we can simply apply the two state models, as we can have each type of detection (1's and 2's) occur at any occasion (survey). Here we consider the case where method M1 is deployed on some occasions, while method M2 is used on other sampling occasions.

Let's consider the case of a study on frogs. Suppose visual surveys (method M1), which provide unambiguous detections of the species of interest, were used on occasions 1 and 3. Aural surveys (method M2), which are prone to false positives, were done on occasions 2, 4 and 5. All detections from visual surveys can be coded as '2', given that they are unambiguous. All detections from aural surveys should be coded as '1', as they are ambiguous (they can be true or false positives). Your data would look something like this:

```
00211 1;
00011 1;
20211 1;
21201 1;
00000 1;
```

This looks very similar to data obtained under a 'two detection states' scenario, don't you think?

However, a few things are different here. First, unambiguous detections (2's) only occur on occasions 1 and 3 and no ambiguous detections (1's) occur. Conversely, on occasions 2, 4 and 5, only ambiguous occasions occur.

In our example, all sites are surveyed on each occasion, but if only a subset of sites were surveyed with the unambiguous method (M1, visual surveys) then a subset of those sites would have missing values, or '.' , in the first or third columns of their histories. Not all sites have to be surveyed equally.

Now, the probabilities of true detection (p_{11}) of each method are likely to be different. This is something important to consider, and which you are probably interested in estimating anyway. We use a superscript to differentiate the true detection parameters: p_{11}^{M1} for method M1 (here, visual surveys) and p_{11}^{M2} for method M2 (here, aural surveys). The conditional detection probabilities for each method look like this:

M1	No detection (0)	Detection (2)
$z = 0$	1	0
$z = 1$	$1 - p_{11}^{M1}$	p_{11}^{M1}

M2	No detection (0)	Detection (2)
$z = 0$	$1 - p_{10}$	p_{10}
$z = 1$	$1 - p_{11}^{M2}$	p_{11}^{M2}

First, note that the false detection parameter p_{10} only applies to method M2. As for the 'two state' model, we have three detection parameters here too: $p_{10}, p_{11}^{M1}, p_{11}^{M2}$. To better see the link with the 'two state' model, we can write the conditional detection probabilities as:

true state	$y = 0$	$y = 1$	$y = 2$
$z = 0$	$1 - p_{10}$	p_{10}	0
$z = 1$	$1 - (p_{11}^{M2} + p_{11}^{M2})$	p_{11}^{M2}	p_{11}^{M1}

On occasions where visual surveys (M1) were used, we simply set $p_{11}^{M2} = 0$ and $p_{10} = 0$. When aural surveys (M2) were used, we set $p_{11}^{M1} = 0$. If you have a look at the similar table we showed (above) for the 'two state' model, you can see a relationship between parameters $\{p_{11}, b\}$ (which is the parameterization used by **MARK**) and the methods-specific parameters $\{p_{11}^{M1}, p_{11}^{M2}\}$. We have:

$$p_{11} = p_{11}^{M1} + p_{11}^{M2}$$

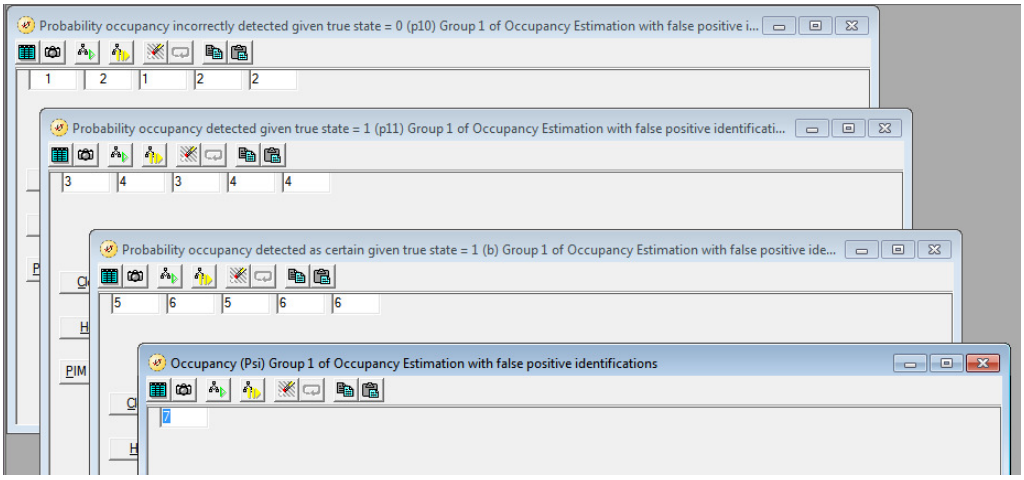
$$p_{11}^{M2} = (1 - b)p_{11}$$

$$p_{11}^{M1} = bp_{11}$$

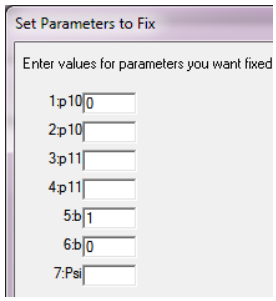
We can thus easily implement the 'two methods' model using the parameterization of the 'two states' model implemented in **MARK**. To do this, we will simply need to do the following:

1. Fix $p_{10} = 0$ for occasions when Method 1 was used (occasions 1 and 3 in our example).
2. Specify different p_{11} parameters for occasions when Method 1 and Method 2 were used. In our example, occasion $\{1, 3\}$ vs. $\{2, 4, 5\}$.
3. Fix $b = 1$ for occasions when Method 1 was used, and $b = 0$ for occasions when Method 1 was used. By doing so, **MARK** will directly provide us with method-specific p_{11} estimates. This is obvious from the formulae above: when $b = 1, p_{11} = p_{11}^{M1}$ and when $b = 0, p_{11} = p_{11}^{M2}$.

Let's illustrate this with a real example. We simulated data (see file **FP-two-methods.inp**), using the following constant parameter values: $\psi = 0.35, p_{11}^{M1} = 0.55, p_{11}^{M2} = 0.75, p_{10} = 0.10$. There are 5 sampling occasions. Method 1 (unambiguous) was used on occasions 1 and 3, while Method 2 (ambiguous) was used on surveys 2, 4 and 5. In **MARK**, open the PIMs for all 4 parameter types and number the parameters as follows:



Then, do '**Run | Current Model**' and click on '**Fix Parameters**'. We have 7 numbered parameters (see PIMs), but as we are running a constant model, there are only 4 real parameters. So, we should be fixing values for 3 parameters. Indeed, we fix the following: (i) Parameter 1: $p_{10} = 0$, (ii) Parameter 5: $b = 1$, and (iii) Parameter 6: $b = 0$. This ensures that we define parameter 3 as p_{11}^{M1} and parameter 4 as p_{11}^{M2}.



Click '**OK**' and run the model. The result should give you an AIC_c of about 4458.46. The parameter estimates are the following:

two method

Real Function Parameters of {model 1}

Parameter	Estimate	Standard Error	95% Confidence Interval Lower	Upper	
1:p10	0.0000000	0.0000000	0.0000000	0.0000000	Fixed
2:p10	0.0905324	0.0079822	0.0760570	0.1074424	
3:p11	0.5215459	0.0213330	0.4796801	0.5631113	
4:p11	0.7626228	0.0142814	0.7335069	0.7894706	
5:b	1.0000000	0.0000000	1.0000000	1.0000000	Fixed
6:b	0.0000000	0.0000000	0.0000000	0.0000000	Fixed
7:Psi	0.3518391	0.0165462	0.3201301	0.3849108	

These are indeed very close to the values we used to simulate the data. Note that, as expected, parameter 3 (p11) corresponds to p_{11}^{M1} (real value = 0.55) and parameter 4 (p11) corresponds to p_{11}^{M2} (real value = 0.75).

This was for the constant model. But what if you want to assess time variation in the probability detection of either Method 1, or Method 2, or both? It is actually very easy. Let say you want to model occasion-specific variation for the detection parameter of Method 1 (p_{11}^{M1}) which was employed on occasions 1 and 3. To do this, open the PIM of p11, and change the numbering as follows: `p11: 3 5 4 5 5`.

Now, p_{11}^{M1} corresponds to two parameters (3 and 4), one for each occasion that Method 1 was used. p_{11}^{M2} is now parameter 5 and is constant over time. '**`Renumber without overlap`**' to get the correct numbering for the other parameters (the numbering now goes from 1 to 8). Because the numbering has changed, we now need to fix the following parameters : (i) Parameter 1: $p_{10} = 0$, (ii) Parameter 6: $b = 1$, (iii) Parameter 7: $b = 0$.

Run the model. This model gets an AIC_c of 4,460.20 and the two estimates for p_{11}^{M1} are 0.51 (parameter 3) and 0.53 (parameter 4) for occasion 1 and 3, respectively. All other parameter estimates are virtually the same as before.

To run a model where only p_{11}^{M2} varies over time, but p_{11}^{M1} is constant, simply change the PIM numbering to: `p11: 3 4 3 5 6`. Parameters 4, 5 and 6 will provide you estimates of p_{11}^{M2} for occasions 2, 4 and 5, respectively. p_{11}^{M1} is simply parameter 3.

To run a model where both p_{11}^{M1} and p_{11}^{M2} vary, simply make p11 all-varying. So, change the PIM numbering to: `p11: 3 5 4 6 7`. Parameters 3 and 4 will provide you estimates of p_{11}^{M1} for occasions 1 and 3, respectively. Parameters 4, 5 and 6 will provide you estimates of p_{11}^{M2} for occasions 2, 4 and 5, respectively. Each time you change the PIM numbering, remember to (1) renumber without overlap and (2) modify and correctly assign fixed parameters' values for the b parameter(s).

end sidebar

22.7.2. The dynamic occupancy model with false positives

If you understand (1) the 'classic' dynamic occupancy model (see section 22.6), and (2) the single-season false-positives model we just presented, it should be relatively straightforward for you to understand (and implement) the dynamic false-positives occupancy model. As you might expect, this model simply combines both processes:

(1) Dynamic Occupancy Process: We now allow occupancy to change between seasons (i.e., primary occasions). To model the dynamic occupancy process, we use the 3 basic parameters: (i) ψ_1 is the initial occupancy probability (i.e., occupancy in season 1), (ii) ϵ is the probability of site extirpation, and (iii) γ is the probability of site colonization.

These parameters are exactly the same as those defined in section (22.6). Remember that data should be collected following a robust design, with multiple surveys in each season. Occupancy can only change between seasons, and within any season, we assume closure, so occupancy does not change between surveys within a season.

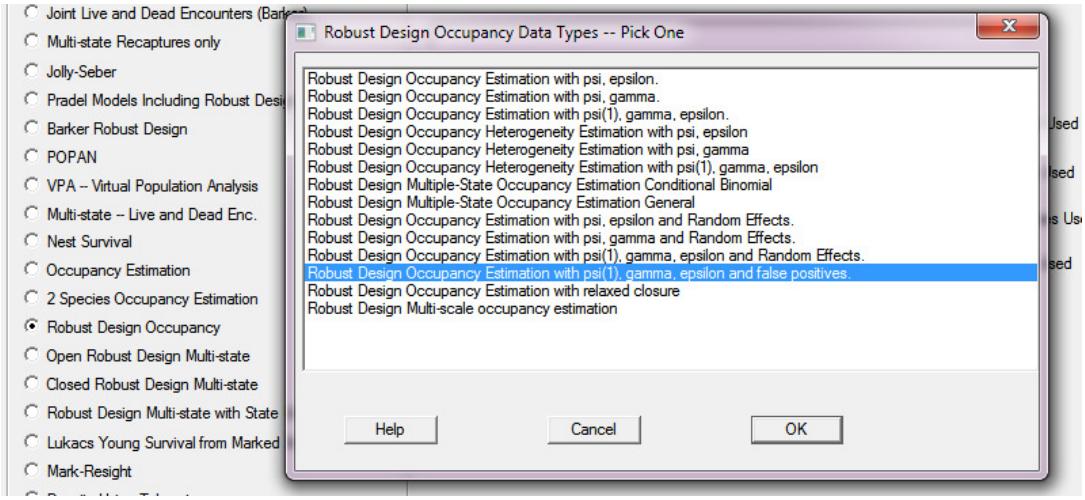
(2) Detection Process: We consider the possibility that both false negative and false positive detections may occur at any occasion. To model this detection process, we still use the 3 parameters: p_{10}, p_{11} and b. Here, again, it is important that our datasets has two types of detections: *ambiguous* and *unambiguous*. We can implement the 'two detection-states' or the 'two detection-methods' versions of the model, using the same procedure we explained earlier in this section.

Let's consider how we run this model in **MARK**. Here again, we simulated data and created a data file example. We considered 4 seasons, with 3 surveys in each season. We used the following constant

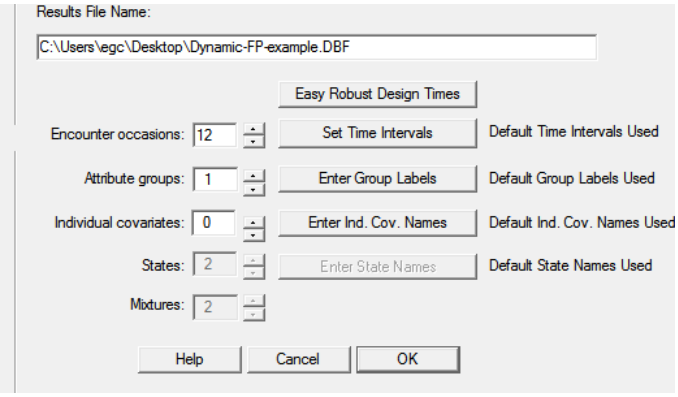
parameter values to simulate the data:

```
psi = 0.40          # Pr(Initial Occupancy)
epsilon = 0.10      # Pr(Extirpation)
gamma = 0.25        # Pr(Colonization)
p10 = 0.20           # Pr(y=1|z=0), i.e., False Positive Probability
p11 = 0.65      # Pr(y=1|z=1), i.e., True Detection Probability
Probability  b = 0.7  # Proportion of "certain" detection (see Miller et al. 2011).
```

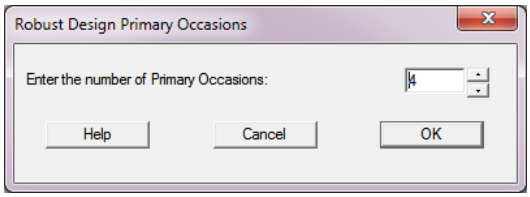
Open **MARK**, select the '**Robust Design Occupancy**' data type and choose '**Robust Design Occupancy False Positives Estimation with psi(1), gamma, epsilon**'.



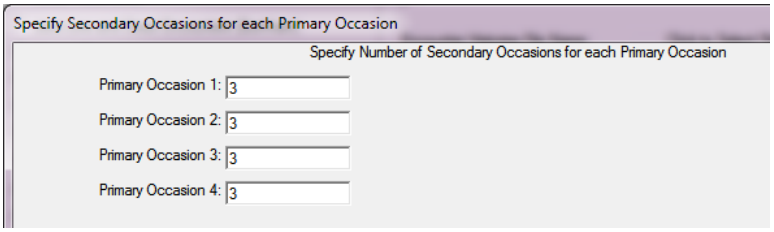
Click '**OK**'. To use our simulated data file, select the file called **Dynamic-FP-example.inp**. Look at it if you wish. You can see that there are 12 occasions (i.e., 4 seasons × 3 surveys/season). Next, specify the total number of encounter occasions (12).



Then click on the '**Easy Robust Design**' button to specify that there are 4 primary occasions (seasons).



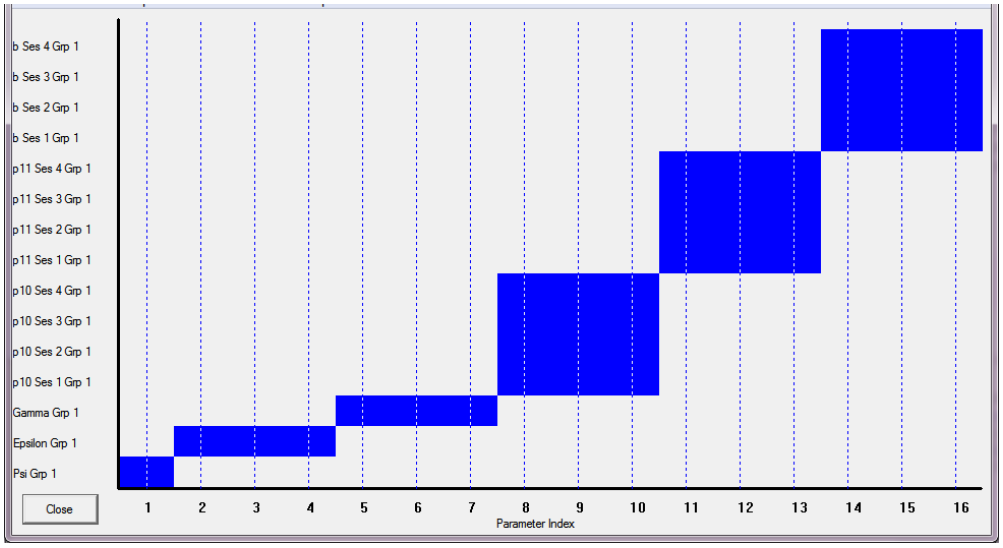
In the window that pops up (see below), specify 3 surveys (secondary occasions) for each season (primary occasion). These values have defaulted in so leave as is. If, in some years, you had more or less occasions you could adjust this here. Then, click '**OK**'.



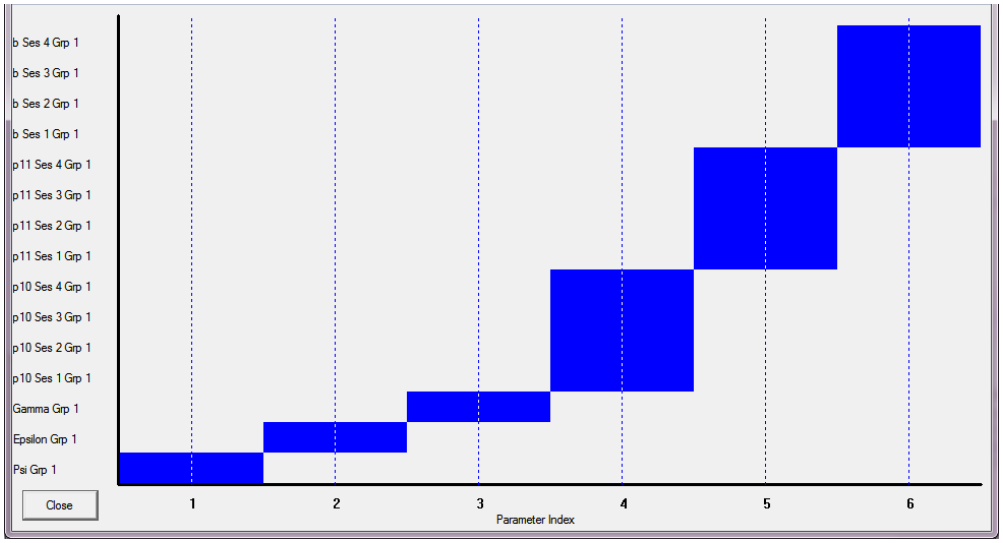
Click '**OK**' again in the major window to get started with the analysis.

Have a look at the PIM chart. Remember that in **MARK** seasons are referred to as 'sessions'. You can see from the PIM chart that, by default, **MARK** assumes different 'detection' parameters (`p10`, `p11` and `b`) for each season. For instance, you can read on the y-axis of the chart: '`p10 Ses 1 Grp 1`', this is `p10` for season 1 (`Ses 1` = Season 1); '`p10 Ses 2 Grp 1`', this is `p10` for season 2 (`Ses 2` = Season 2), and so on.

Let's first build a model where all basic parameters are constant. This is in fact the model we used to simulate data. In the PIM chart, we first stack the different seasons (sessions), for parameters `p10`, `p11` and `b`, just like in the picture below. This specifies that these parameters are equal across seasons. Remember to '**`renumber with overlap`**'. It should look like the following:



Next, we want set all these 'blue boxes' as '**constant**' over surveys. Use right click, and '**constant**' to do this. Renumber with overlap. It should look like this:



Six parameters remain, which is indeed what we want for this simple model. One parameter for each of the six basic parameter: `psi`, `epsilon`, `gamma`, `p10`, `p11` and `b`. Now, simply go ahead and just run the model. It gets an AIC_c value of 22,061.5940. Let's look at the parameter estimates:

Dynamic False Positive Example

Real Function Parameters of {model 1}

Parameter	Estimate	Standard Error	95% Confidence Interval Lower	Upper
1:Psi	0.3953672	0.0177044	0.3612447	0.4305403
2:Epsilon	0.0896572	0.0120253	0.0687145	0.1161864
3:Gamma	0.2708277	0.0133774	0.2454184	0.2978294
4:p10 Session 1	0.2058186	0.0059984	0.1943103	0.2178243
5:p11 Session 1	0.6461338	0.0073063	0.6316859	0.6603197
6:b Session 1	0.7005843	0.0080732	0.6845254	0.7161647

As expected, these are very close to the real values we used to simulate the data: $\psi = 0.4$, $\epsilon = 0.1$, $\gamma = 0.25$, $p10 = 0.2$, $p11 = 0.65$, $b = 0.7$. You can then go ahead and proceed as usual to run any other model.

22.8. Summary

The occupancy models described in this chapter represent only a subset of the occupancy models available in program **MARK**. You've likely noticed the variety of options (models) when selecting 'Occupancy' data types. We believe that the single-season and dynamic versions of the multi-state occupancy models (Nichols *et al*. 2007, MacKenzie *et al*. 2009), models involving multiple scales (e.g., Nichols *et al*. 2008, Hines *et al*. 2010), and species-interaction models (2 Species Occupancy Estimation models, Richmond *et al*. 2010 – see also Chapter 23 in this book) are all extremely useful with many applications in the literature. Currently, we refer readers to the help files in **MARK** for details about

these models and associated references in the primary literature.

Relative to mark-recapture information, occurrence data is relatively easy to collect, and historic data with replication is often available. The combined flexibility of occupancy models and the availability of detection-nondetection data has led to investigators to consider occupancy models for their applications. For any occupancy-based study, we believe investigators should clearly define the following elements as applied to their objectives: sample units (i.e. sites), the time period over which occurrence is assumed to be static (i.e. season(s)), replicate surveys, and the criteria that constitute 'detection' (Bailey *et al*. 2014). Relating these definitions of sites, surveys, and season to the study's biological questions is the foundation of any good study design and influences the interpretation of resulting model parameters.

For example, in their summary paper Bailey *et al*. (2014) used a single host-pathogen system to demonstrate how study design and focal scale change depending on the motivating biological question(s). In their range of examples, biological interest focused on investigating factors influencing pathogen prevalence in a single host population, or estimating pathogen occurrence across multiple host populations or across various habitat types. Accordingly, the appropriate definition of 'site' varied from an individual amphibian, to host populations, to potential amphibian breeding habitats. Season and survey definitions ranged from a single visit (season) with replicate qPCR surveys (qPCR wells), to a defined amphibian breeding season where surveys may include a mixture of pathogen samples from individual hosts (e.g., swabs) and environmental surveys (e.g., water filter samples). Different biological questions lead to unique study designs and associated definitions of site, season, and surveys. In closing, we encourage investigators to think carefully about their study design and the associated model assumptions when considering occupancy models to address their own biological questions.

22.9. References

Bailey, L. L., Reid, J. A., Forsman, E. D., and Nichols, J. D. (2009) Modeling co-occurrence of northern spotted and barred owls: accounting for detection probability differences. *Biological Conservation*, **142**, 2983-2989.

Bailey, L. L., MacKenzie, D. I., and Nichols, J. D. (2014) Advances and applications of occupancy models. *Methods in Ecology and Evolution*, **5**, 1269-1279.

Chambert, T., Miller, D. A. W., and Nichols, J. D. (2015) Modeling false positive detections in species occurrence data under different study designs. *Ecology*, **96**, 332-339.

Chambert, T., Kendall, W. L., Hines, J. E., Nichols, J. D., Pedrini, P., Waddle, J. H.,Tavecchia, G., Walls, S. C., and Tenan, S. (2015) Testing hypotheses on distribution shifts and changes in phenology of imperfectly detectable species. *Methods in Ecology and Evolution*, **6**, 638-647.

Clement, M. J. (2016) Designing occupancy studies when false-positive detections occur. *Methods in Ecology and Evolution*, **7**, 1538-1547.

Farris, Z. J., Gerber, B. D., Valenta, K., Rafaliarison, R., Razafimahaimodison, J. C., Larney, E., Rajaonarivelo, T., Randriana, Z., Wright, P. C., and Chapman, C. A. (2017) Threats to a rainforest carnivore community: A multi-year assessment of occupancy and co-occurrence in Madagascar. *Biological Conservation*, **210**, 116-124.

Gerber, B. D., Karpanty, S. M., and Randrianantenaina, J. (2012) The impact of forest logging and fragmentation on carnivore species composition, density and occupancy in Madagascar's rainforests. *Oryx*, **46**, 414-422.

Gerber, B. D., Williams, P. J., and Bailey, L. L. (2014) Primates and cameras. *International Journal of Primatology*, **35**, 841-858.

Gimenez, O., and Choquet, R. (2010) Individual heterogeneity in studies on marked animals using numerical integration: capture-recapture mixed models. *Ecology*, **91**, 951-957.

Hines, J. E., Nichols, J. D., Royle, J. A.,MacKenzie, D. I., Gopalaswamy, A., Kumar, N. and Karanth, K. (2010) Tigers on trails: Occupancy modeling for cluster sampling. *Ecological Applications*, **20**, 1456-1466.

Kendall, W. L. (1999) Robustness of closed capture-recapture methods to violations of the closure assumption. *Ecology*, **80**, 2517-2525.

Kendall, W. L., and White, G. C. (2009) A cautionary note on substituting spatial subunits for repeated temporal sampling in studies of site occupancy. *Journal of Applied Ecology*, **46**, 1182-1188.

Lachish, S., Gopalaswamy, A. M., Knowles, S. C., and Sheldon, B. C. (2012) Site-occupancy modelling as a novel framework for assessing test sensitivity and estimating wildlife disease prevalence from imperfect diagnostic tests. *Methods in Ecology and Evolution*, **3**, 339-348.

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L. L., and Hines, J. E. (2017) Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. 2nd Edition. Academic Press, Burlington, MA, USA. 648 pp.

MacKenzie, D. I., Nichols, J. D., Seamans, M., and Gutierrez, R. (2009) Modeling species occurrence dynamics with multiple states and imperfect detection. *Ecology*, **90**, 823-835.

Martin, D. J., White, G. C., and Pusateri, F. M. (2007) Occupancy rates by swift foxes (*Vulpes velox*) in eastern Colorado. *The Southwestern Naturalist*, **52**, 541-551.

McClintock, B. T., and White, G. C. (2009) A less field-intensive robust design for estimating demographic parameters with mark-resight data. *Ecology*, **90**, 313-320.

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simons, T. R. (2010) Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, **91**, 2446-2454.

Miller, D. A. W., Nichols, J. D., Gude, J. A., Rich, L. N., Podruzny, K. M., Hines, J. E., and Mitchell, M. S. (2013) Determining occurrence dynamics when false positives occur: estimating the range dynamics of wolves from public survey data. *PLOSOne*, **8**,[e65808].

Miller, D. A. W., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., and Weir, L. A. (2011) Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology*, **92**, 1422-1428.

Miller, D. A. W., Weir, L. A., McClintock, B. T., Grant, E. H. C., Bailey, L. L., and Simons, T. R. (2012) Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications*, **22**, 1665-1674.

Nichols, J. D., Hines, J. E., MacKenzie, D. I., Seamans, M., and Gutierrez, R. (2007) Occupancy estimation and modeling with multiple states and state uncertainty. *Ecology*, **88**, 1395-1400.

Nichols, J. D., Bailey, L. L., O'Connell, A., Talancy, N., Grant, E., Gilbert, A., Annand, E., Husband, T., and Hines, J. E. (2008) Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology*, **45**, 1321-1329.

Otto, C. R., Bailey, L. L., and Roloff, G. J. (2013) Improving species occupancy estimation when sampling violates the closure assumption. *Ecography*, **36**, 1299-1309.

Pledger, S., and Phillpot, P. (2008) Using mixtures to model heterogeneity in ecological capture-recapture studies. *Biometrical Journal*, **50**, 1022-1034.

Pillay, R., Miller, D. A. W., Hines, J. E., Joshi, A. A., and Madhusudan, M. D. (2014) Accounting for false positives improves estimates of occupancy from key informant interviews. *Diversity and Distribu-*

tions, **20**, 223-235.

Richmond, O., Hines, J. E., and Beissinger, S. (2010) Two-species occupancy models: a new parameterization applied to co-occurrence of secretive rails. *Ecological Applications*, **20**, 2036-2046.

Rodda, G. H., Dean-Bradley, K., Campbell, E. W., Fritts, T. H., Lardner, B., Yackel Adams, A. A., and Reed, R. N. (2015) Stability of detectability over 17 years at a single site and other lizard detection comparisons from Guam. *Journal of Herpetology*, **49**, 513-521.

Rota, C. T., Fletcher Jr., R. J., Dorazio, R. M., and Betts, M. G. (2009) Occupancy estimation and the closure assumption. *Journal of Applied Ecology*, **46**, 1173-1181.

Ruiz-Gutiérrez, V., Zipkin, E. F., and Dhondt, A. A. (2010) Occupancy dynamics in a tropical bird community: unexpectedly high forest use by birds classified as non-forest species. *Journal of Applied Ecology*, **47**, 621-630.

Royle, J. A., and Nichols, J. D. (2003) Estimating abundance from repeated presence-absence data or point counts. *Ecology*, **84**, 777-790.

Royle, J. A., and Link, W. A. (2006) Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, **87**, 835-841.

Simons, T. R., Alldredge, M. W., Pollock, K. H., and Wettroth, J. M. (2007) Experimental analysis of the auditory detection process on avian point counts. *The Auk*, **124**, 986-999.

Weir, L. A., Royle, J. A., Nanjappa, P., and Jung, R. E. (2005) Modeling anuran detection and site occupancy on North American Amphibian Monitoring Program (NAAMP) routes in Maryland. *Journal of Herpetology*, **39**, 627-639.

White, G. C., and Cooch, E. G. (2017) Population abundance estimation with heterogeneous encounter probabilities using numerical integration. *Journal of Wildlife Management*, **81**, 322-336.