Implementation of false positive occupancy models in program E-SURGE

Thierry Chambert & David A.W. Miller
Pennsylvania State University, Department of Ecosystem Science and Management, University Park, Pennsylvania, USA
email: thierry.chambert@gmail.com

We present implementation of the false positive occupancy models of Miller et al. (2011, 2013). We first show the steps for implementing the single-season, multiple-detection-state model. Then, we show the necessary modifications required to implement (1) the multiple-detection-method model, (2) the dynamic versions of these models, and (3) add random effects. The details of the implementation in E-SURGE are provided, as well as R code to generate data from these models. Lastly, E-SURGE parameters estimates are provided for a simulated dataset and compared to those obtained with PRESENCE (see Table 1 at the end of the document).

References
Miller, D.A., Nichols, J.D., McClintock, B.T., Grant, E.H.C., Bailey, L.L. & Weir, L.A. (2011). Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. *Ecology*, 92, 1422–1428.
Miller, D.A.W., Weir, L.A., Mcclintock, B.T., Grant, E.H.C., Bailey, L.L. & Simons, T.R. (2012). Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications*, 22, 1665–74.

1. Single-Season, Multiple-Detection-State Model

A. Model Description

We consider two true states (*occupied* and *unoccupied*) and three types of observations (events): non-detection (coded 0), ambiguous detection (coded 1), which could be a true of false detection and unambiguous detection (coded 2), which is necessarily a true detection.

In addition to the occupancy probability parameter (ψ), Miller et al. (2011) defined three parameters to describe conditional detection probabilities: p_{10} is the probability of a (false) detection given that the site is unoccupied, also referred to as 'false positive probability'; p_{11} is the probability of a detection given that the site is occupied, also referred to as 'true detection probability'; and b is the probability that a detection is classified as unambiguous, which can only occur on an occupied site.

B. Simulate Data in R

```
### POPULATION PARAMETERS
psi <- 0.6
p11 <- 0.7
p10 <- 0.1
b <- 0.4
### SAMPLE CHARACTERSITICS
K <- 5
R <- 200
```

```
### OCCUPANCY STATES
Z <- rbinom(R, 1, psi) ; head(Z) ; length(Z)
### DATA
 Y <- matrix(NA,R,K)
  for (i in 1:R){
        pr.X0 <- Z[i]*(1-p11) + (1-Z[i])*(1-p10)
        pr.X1 <- Z[i]*(1-b)*p11 + (1-Z[i])*p10
        pr.X2 <- Z[i]*b*p11
          p <- c( pr.X0 , pr.X1 , pr.X2 )  ;  sum(p)
   for (t in 1:K){
          Y.index <- as.vector(rmultinom(1,1,p))
          Y[i,t] <- which(Y.index == 1) - 1
   } # 't'
  } # 'i'
```

C. Implementation in E-SURGE

GEPAT

Initial state

$* \quad p$

Transition

$$\begin{matrix} * & - & - \\ - & * & - \\ - & - & * \end{matrix}$$

Event

$$\begin{matrix} * & \beta_1 & - \\ * & \beta_2 & \beta_3 \\ * & - & - \end{matrix}$$

Here, subscripts are shown for convenience; in E-SURGE, one only writes a letter (e.g., β) in each cell. This detection/non-detection structure is the core of these false positive models. There is a direct correspondence between each event probability and the parameters described in Miller et al. 2011:

$\beta_1 = p_{10}$

$\beta_2 = (1 - b) \times p_{11}$

$\beta_3 = b \times p_{11}$

If needed, estimates for p_{11} and b can be derived as follows:

$\hat{p}_{11} = \hat{\beta}_2 + \hat{\beta}_3$

$\hat{b} = \hat{\beta}_3 / (\hat{\beta}_2 + \hat{\beta}_3)$

The corresponding standard errors and 95% confidence intervals can be obtained by the delta method (using the R package msm for example) or via a bootstrap procedure.

GEMACO

Initial state
  i

Transition

Event
  f.to

Once the model is fully parameterized, click the IVFV button (Initial Values Fixed Values) to set initial values, and the RUN button to launch the likelihood optimization.

2. MULTIPLE-DETECTION-METHOD MODEL

A. Description

Now, we consider a sampling design where 3 surveys were done with an ambiguous detection method (M1) and 2 surveys were done with an unambiguous detection method. Detections from method M1 are coded as event 2 ($y = 1$) and are specified as surveys 1 through 3. Detections from method M2 are coded as event 3 ($y = 2$) and specified as surveys 4 and 5. Non-detection for both survey types is coded as event 1 ($y = 0$).

B. Simulate Data

```
### POPULATION PARAMETERS
psi <- 0.6
p11 <- 0.7
p10 <- 0.2
r11 <- 0.5
### SAMPLE CHARACTERSITICS
 K <- 3
 J <- 2
R <- 200
### OCCUPANCY STATES
Z <- rbinom(R, 1, psi) ; head(Z) ; length(Z)
### DATA
 Y <- matrix(NA,R,K)
 W <- matrix(NA,R,J)

  for (i in 1:R){
    for (k in 1:K){
        p <- (Z[i]*p11) + ((1-Z[i])*p10)
          Y[i,k] <- rbinom(1,1,p)
    } # 'k'
    for (j in 1:J){
          W[i,j] <- rbinom(1,1,Z[i]*r11)*2
    } # 'j'
```

```
  } # 'i'
```

C. Implementation in E-SURGE

We use the same GEPAT structure as for the multiple-detection-state model, modifying the inputs in GEMACO and IVFV to specify that:
- β_3 (probability of unambiguous detection) is null for occasions 1 to 3;
- β_1 and β_2 are null for occasions 4 and 5. To do this, we change the GEMACO syntax for event probabilities to read **f.to.t(1 2 3 , 4 5)**. The added syntax, **t(1 2 3 , 4 5),** specifies different detection probability for method M1 (occasions 1 to 3) and method M2 (occasions 4,5).
In the IVFV module, we fix to zero the values of β_3 for occasions 1 to 3, as well as the values of β_1 and β_2 for occasions 4 and 5.

3. Dynamic Models

A. Description

We present implementation of the multi-season false positive occupancy models of Miller et al. (2013). The two types of model (multiple-detection-state and multiple-detection-method) can equally be implemented by following the same instruction as above (single-season model) for the "Event" structure in GEMACO.

B. Simulate Data

```
### POPULATION PARAMETERS
 psi <- 0.6
p11 <- 0.7
p10 <- 0.1
b <- 0.4
col <- 0.5
ext <- 0.4
### SAMPLE CHARACTERSITICS
 T <- 3          # years
 K <- 4          # surveys
R <- 200
### OCCUPANCY STATE DYNAMIC
  Z <- matrix(NA, R, T)
Z[,1] <- rbinom(R, 1, psi) ; head(Z) ; length(Z)
for (i in 1:R){
 for(t in 2:T){
 Pr.Occ <- Z[i,t-1]*col + (1-Z[i,t-1])*ext
Z[i,t] <- rbinom(1, 1, Pr.Occ)
} }
### DATA
 Y <- matrix(NA,R,T*K)
  for (i in 1:R){
    for (t in 1:T){
        pr.X0 <- Z[i,t]*(1-p11) + (1-Z[i,t])*(1-p10)
```

```
        pr.X1 <- Z[i,t]*(1-b)*p11 + (1-Z[i,t])*p10
        pr.X2 <- Z[i,t]*b*p11
          p <- c( pr.X0 , pr.X1 , pr.X2 )  ;  sum(p)

        for (k in 1:K){
          Y.index <- as.vector(rmultinom(1,1,p))
          Y[i,(t-1)*K+k] <- which(Y.index == 1) - 1
        } # 'k'
  } # 't'
 } # 'i'
```

C. Implementation in E-SURGE

Fitting dynamic models in E-SURGE only requires modifying the transition matrix, while initial state and event probabilities are the same as for the single-season model. In GEPAT, we will define the colonization (γ) and extinction (ε) probabilities. In GEMACO, we will specify at which occasions extinction and colonization can occur. We use an example with 3 primary sampling seasons and 4 secondary surveys per season (12 total occasions): state transitions can only occur from occasions 4 to 5 and 8 to 9. The values of γ and ε are thus estimated for occasion 4 and 8 and will be fixed to zero (at step IVFV) for all other occasions.

GEPAT

Initial state

$* \quad p$

Transition (the same letter can be used for colonization/extinction)

$$\begin{matrix} * & \gamma & - \\ \varepsilon & * & - \\ - & - & * \end{matrix}$$

Event

$$\begin{matrix} * & \beta_1 & - \\ * & \beta_2 & \beta_3 \\ * & - & - \end{matrix}$$

GEMACO

Initial state

i

Transition

to.t(1 2 3 5 6 7 9 10 11) + to.t(4 8)

Event

f.to

In the next step (IVFV), we fix the parameters corresponding to secondary occasions to zero to impose closure within the primary periods.

4. RANDOM EFFECTS

In E-SURGE, one can easily add random effects to any parameter of interest. For the false positive models, this is particularly attractive to deal with site heterogeneity in false positive rates (e.g., when the distribution or abundance of the entities responsible for misidentification [similar-looking species] is not uniform across sites sampled).

Random effects on parameter p_{10} are easily included by adding the syntax **f(1)>ind** for the event probability matrix in GEMACO, which specifies that individual site random effects ('**ind**') apply only ('**>**') to parameter β_1 (component '**f(1)**'). The full GEMACO syntax for event probabilities is **f.to + f(1)>ind**. Currently, no other software allows such easy and fast implementation of random effects for occupancy.

In the dynamic model, one can also include site heterogeneity (random effects) on colonization or extinction probabilities. For instance, for colonization, use **to.t(1 2 3 5 6 7 9 10 11) + to.t(4 8) + [to(2).t(4 8)]>ind** in GEMACO.

Example of Simulated Data, with random effect

```
### Functions
expit <- function(x) {1/(1 + exp(-x))}
logit <- function(p) {log(p/(1-p))}
### SAMPLE CHARACTERSITICS
 K <- 8
R <- 500
### POPULATION PARAMETERS
 psi <- 0.6
p11 <- 0.8
b <- 0.4
mu.p10 <- logit(0.3)
sd.orig <- 0.1
  sigma.p10 <- sqrt((sd.orig ^2)/( (expit(mu.p10)^2)*((1-expit(mu.p10))^2) ) )
RE <- rnorm(R, 0, sigma.p10)
 p10 <- expit(mu.p10 + RE)
mean(p10) ; sd(p10)
### OCCUPANCY STATES
Z <- rbinom(R, 1, psi) ; head(Z) ; length(Z)
### DATA
 Y <- matrix(NA,R,K)
  for (i in 1:R){
        pr.X0 <- Z[i]*(1-p11) + (1-Z[i])*(1-p10[i])
        pr.X1 <- Z[i]*(1-b)*p11 + (1-Z[i])*p10[i]
        pr.X2 <- Z[i]*b*p11
          p <- c( pr.X0 , pr.X1 , pr.X2 )  ;  sum(p)
    for (t in 1:K){
          Y.index <- as.vector(rmultinom(1,1,p))
          Y[i,t] <- which(Y.index == 1) - 1
   } # 't'
  } # 'i'
```

Table 1: Parameter estimates obtained from programs E-SURGE and PRESENCE (for comparison) for false positive occupancy models fitted to simulated datasets.

Model	Parameter	True Value	E-SURGE		PRESENCE	
			MLE	SE	MLE	SE
Single-season	ψ	0.60	0.61	0.04	0.61	0.04
Multiple State	p_{10}	0.10	0.10	0.02	0.10	0.02
	β_2	0.42	0.44	0.02	--	--
	β_3	0.28	0.30	0.02	--	--
	p_{11}	0.70	0.74*	0.03†	0.74	0.02
	b	0.40	0.40*	0.03†	0.40	0.02
Single-season	ψ	0.60	0.65	0.05	0.65	0.05
Multiple Methods	p_{10}	0.20	0.17	0.04	0.17	0.04
	p_{11}	0.70	0.65	0.03	0.65	0.03
	r_{11}	0.50	0.51	0.04	0.51	0.04
Single-season	ψ	0.60	0.59	0.02	--	--
Random effects	$\mu_{p_{10}}$	0.30	0.32	0.01	--	--
	$\sigma^2_{p_{10}}$	0.10	0.09	0.004	--	--
	β_2	0.48	0.47	0.01	--	--
	β_3	0.32	0.33	0.01	--	--
	p_{11}	0.80	0.80*	0.01†	--	--
	b	0.40	0.42*	0.02†	--	--
Multi-season	ψ	0.60	0.61	0.04	0.61	0.04
	γ	0.50	0.41	0.04	0.41	0.04
	ε	0.40	0.45	0.04	0.45	0.04
	p_{10}	0.10	0.08	0.01	0.08	0.01
	β_2	0.42	0.42	0.01	--	--
	β_3	0.28	0.26	0.01	--	--
	p_{11}	0.70	0.68*	0.02†	0.68	0.02

	b	0.40	0.38*	0.02†	0.38	0.02

*Estimates calculated as derived parameters.
†SE calculated with the delta method.