Fitting occupancy models with E-SURGE:

Hidden Markov modelling of presence-absence data

Olivier Gimenez[1], Laetitia Blanc[1], Aurélien Besnard[1], Roger Pradel[1], Paul F. Doherty Jr[2], and Rémi Choquet[1]

[1]Centre d'Ecologie Fonctionnelle et Evolutive, UMR 5175, campus CNRS, 1919 Route de Mende, 34293 Montpellier Cedex 5, France.

[2]Department of Fish, Wildlife and Conservation Biology, Colorado State University, Fort Collins, CO 80523-1474, USA

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

Abstract.

1. Occupancy – the proportion of area occupied by a species – is a key notion for addressing important questions in ecology, biogeography and conservation biology. Occupancy models allow estimating and inferring about species occurrence while accounting for false absences (or imperfect species detection).

2. Most occupancy models can be formulated as hidden Markov models (HMM) in which the state process captures the Markovian dynamic of the actual but latent states while the observation process consists of observations that are made from these underlying states.

3. We show how occupancy models can be implemented in program E-SURGE, which was initially developed to analyse capture-recapture data in the HMM framework. Replacing individuals by sites provides the user with access to several features of E-SURGE that are *not available altogether* or *just not available* in standard occupancy software: i) user-friendly model specification through a SAS/R-like syntax without having to write custom code, ii) decomposition of the observation and state processes in several steps to provide flexible parameterisation, iii) up-to-date diagnostics of model identifiability and iv) advanced numerical algorithms to produce fast and reliable results (including site random effects).

4. To illustrate E-SURGE features, we provide simulated data and the details of the implementation on the analysis of several occupancy models. These detailed examples are gathered in a companion wiki platform http://occupancyinesurge.wikidot.com/.

Key words: capture-recapture; detectability; detection-nondetection; E-SURGE; hidden Markov models; presence-absence; species occurrence.

Introduction

Occupancy models allow estimating and inferring about species occurrence while accounting for false absences or imperfect species detection (MacKenzie *et al.* 2006). These models have been extensively used to address important questions in fields as diverse as conservation biology, biogeography, wildlife epidemiology, metapopulation dynamics and community ecology (review in (Bailey, MacKenzie, & Nichols 2013)).

Following the seminal work of MacKenzie and colleagues (MacKenzie *et al.* 2006), it was soon realized that occupancy models could be formulated as hidden Markov models (HMMs) in which two time series run in parallel: the state process captures the Markovian dynamic of the actual but latent states (e.g., site occupied vs. unoccupied) while the observation process consists of observations that are made from these underlying states (e.g., species detected vs. undetected) (e.g., Royle & Kéry 2007).

There is an intimate connection between occupancy and capture-recapture models that can be realized by interchanging individuals and sites. Interestingly, the formulation of capture-recapture models as HMMs was also witnessed in the capture-recapture literature (Pradel 2005; Gimenez *et al.* 2012).

Several software are available to fit occupancy models, either in the Frequentist framework with programs PRESENCE (Hines 2013), MARK (White & Burnham 1999) and the R package Unmarked (Fiske & Chandler 2011), or in the Bayesian framework using WinBUGS (Kéry & Schaub 2011). WinBUGS requires writing custom code and knowledge about the Bayesian theory. Programs PRESENCE and MARK often require the construction of design matrices to specify models, a process that can be error-prone. PRESENCE, MARK and Unmarked do not incorporate random effects.

Here, by exploiting the equivalence between occupancy and capture-recapture models, we illustrate how occupancy models can be implemented in program E-SURGE (Choquet,

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

Rouan, & Pradel 2009) which was initially developed to analyse capture-recapture data in the HMM framework. We aim at providing the user with access to features of E-SURGE that are *not implemented altogether* in available occupancy software, namely i) user-friendly model specification through a SAS/R-like syntax without having to write custom code, ii) advanced numerical algorithms to produce fast and reliable results, including the incorporation of site random effects, and several other features that are simply *not implemented* in these software, namely iii) decomposition of the observation and state processes in several steps to provide flexible parameterisation and iv) up-to-date diagnostics of model identifiability, in other words a reliable way of counting the number of parameters entering the calculation of the Akaike Information Criterion.

HIDDEN MARKOV MODELLING OF OCCUPANCY DATA

In Figure 1, we provide the HMM formulation of the general dynamic occupancy models to carry out inference about occurrence and how extinction and colonization drive changes in occurrence.

[FIGURE 1 AROUND HERE]

The parameters of interest are the probability of local extinction ε and of colonization γ as well as the detection probability p and the probability of initial occupancy ψ_1 where we have assumed all parameters constant across periods and sites for simplicity. A HMM is built around three pieces of information: the vector of initial state probabilities, the matrix of transition probabilities linking states in successive sampling occasions and the matrix of observation probabilities linking observations and states at a given occasions. At the first

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

sampling occasion $t = 1$, with the first state being 'unoccupied' and the second 'occupied', the vector of initial state probabilities is:

$$\begin{bmatrix} 1-\psi_1 & \psi_1 \end{bmatrix} \tag{1}$$

Then, the states are distributed as a first-order Markov chain governed by the transition matrix with states unoccupied and occupied at t in rows and states unoccupied and occupied at $t + 1$ in columns:

$$\begin{bmatrix} 1-\gamma & \gamma \\ \varepsilon & 1-\varepsilon \end{bmatrix} \tag{2}$$

The observation process conditional on underlying occupancy states is summarized by a matrix with unoccupied and occupied states at t in rows and undetected and detected observations at visits j in columns:

$$\begin{bmatrix} 1 & 0 \\ 1-p & p \end{bmatrix} \tag{3}$$

Single-season occupancy models can be reformulated as HMMs and fitted in E-SURGE by imposing no extinction ($\varepsilon = 0$) and no colonisation ($\gamma = 0$) in the dynamic model. The extension to multiple states with uncertainty (Nichols *et al.* 2007) is illustrated with breeding states. We consider the states a site is unoccupied, occupied by non-breeders and occupied by breeders, while the observations are species undetected (coded 0), species detected without young (coded 1) and species detected with young (coded 2). We use ψ^1

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

(resp. ψ^2) the probability that the site is occupied by non-breeders (resp. by breeders), p^1 (resp. p^2) the detection probability of non-breeders (resp. of breeders). There is also a possibility to accommodate uncertainty on a state, here for example on the breeder state to acknowledge that even though reproduction occurs on a site, young might be missed. We introduce δ the probability of detecting evidence of reproduction, given the site is occupied with young. Then, the vector of initial state probabilities is:

$$\begin{bmatrix} 1-\psi^1-\psi^2 & \psi^1 & \psi^2 \end{bmatrix} \tag{4}$$

while the transition matrix is the identity matrix. The main modifications are in the observation matrix which can be written as a product of two matrices, highlighting the successive processes of detection and breeding state ascertainment:

$$\begin{bmatrix} 1 & 0 & 0 \\ 1-p^1 & p^1 & 0 \\ 1-p^2 & 0 & p^2 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1-\delta & \delta \end{bmatrix} \tag{5}$$

MAIN FEATURES OF E-SURGE

To illustrate the main features of E-SURGE, we go through the workflow provided in Figure 2. We refer to the E-SURGE manual (Choquet & Nogué 2013) as well as (Choquet 2008) and (Choquet, Rouan, & Pradel 2009) for more details.

[FIGURE 2 AROUND HERE]

Preliminary steps

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

We assume that the user has started a new session, loaded a data file and selected the Occupancy option. In the ‘Data status’ section of the main window, the ‘Modify’ button allows to specify the characteristics of your model. The number of age classes should be fixed to 1 here. Age is the time elapsed since first detection, which is equivalent to time if all sites start being monitored at $t = 1$. If site-specific covariates are to be used, then the number of individual covariates should be amended accordingly (recall that the sites are the equivalent of individuals in capture-recapture analyses). The number of events is the number of observations (e.g., undetected and detected), while the number of states should always be the number of states to be used (e.g., unoccupied, occupied by non-breeders, occupied by breeders) plus one. Indeed, because E-SURGE was initially developed to estimate demographic parameters, it always considers the absorbing state ‘dead’ which is useful in a capture-recapture context to have individuals die but of little interest in an occupancy analysis.

In the ‘Advanced numerical options’ section, the 'Compute C-I' option is deactivated by default to save time by avoiding calculating the parameters’ confidence intervals of each model. Just tick this option to get standard errors and confidence intervals, e.g., for the best fitting model. Yet in the same section, in the pull-down menu ‘Initial values’, E-SURGE offers the possibility to use several sets of initial values randomly chosen or use estimates of a simpler model (from last model) as initial values, which might be useful to avoid picking up local minima in the deviance, an issue often encountered in HMMs.

Model building

The model construction is served in E-SURGE by two modules called GEPAT and GEMACO (Choquet 2008). In GEPAT, you specify the structure of the vector of initial state probabilities, the matrix of transitions governing the state process and the matrix of observations (conditional on the states) governing the observation process. To specify a parameter that will be estimated (i.e., that will be assigned an effect in GEMACO, see below), you can use any letter. The minus sign '-' means that the parameter corresponding to this cell will always be set to 0 while the star '*' means the complementary of the sum of all the other parameters on the same row. This GEPAT step is very useful as it deactivates the relevant matrix elements once and for all without having to fix values every time a model is fitted.

Note that in the vector of initial state probabilities, which corresponds in the capture-recapture context to the state of individual at first encounter, the 'dead' state is always removed as individuals are all alive when marked. In the transition and observation matrices, this state is present and needs to be accounted for.

Of practical interest, the three elements of a HMM can be specified through a multistep process that proves very useful in accommodating state uncertainty for example (eqn. 5). After entering the size of the matrix, the default matrix options (diagonal, full or empty matrix) can be used as a starting point to specify a matrix.

In GEMACO, we specify the effects (*sensu* the design matrix in programs MARK and PRESENCE) using a R-like syntax: for example, a season effect will be specified by 't' for time, a group effect by 'g' while a constant effect will be 'i' for intercept. If the effect of a site covariate needs to be investigated, we use 'i + xind' where xind specifies the slope of the relationship. The matrices defined at the GEPAT step can be manipulated using the syntax 'from' for rows and 'to' for columns. For example, f(1).to(2) will pick the element in row 1 and column 2 of the corresponding matrix, and if a time effect is required on this element, then the syntax will be f(1).to(2).t. If the entire first row of a matrix with five columns needs to be selected, then we use f(1).to(1,2,3,4,5). A colon ':' is useful to lump categories together while the ampersand symbol '&' aggregates parameters corresponding to levels of different

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

factors. Additive and interactive effects can be specified with the plus sign '+' and dot '.' respectively.

Shortcuts can be defined to assign a name to a given syntax, hence simplifying the formulas in GEMACO.

Final steps

The last steps consist of specifying initial values, using the values by default, or fixing parameters to some values if needed (IVFV step), and then running E-SURGE to fit the current model (RUN step). Standard numerical results (e.g., maximum likelihood estimates, AIC, confidence intervals) can be obtained in a text file or an Excel file. In particular, E-SURGE provides a reliable number of parameters via an algorithm described in (Choquet & Cole 2012), which is crucial in particular with parameter-redundant models, and is one of the key steps for correct model selection using the AIC.

CASE STUDIES IN E-SURGE

To illustrate the use of E-SURGE for fitting occupancy models, we simulated data and used existing simulated datasets. Results from fitting the three models described above to these data are provided in Table 1. Estimates for all parameters were close to the true values and the 95% confidence intervals covered the true values in all cases.

[TABLE 1 AROUND HERE]

We go through the most important steps of the implementation in E-SURGE and we refer to the companion wiki website http://occupancyinesurge.wikidot.com/ for full details.

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

Dynamic models

We start with the dynamic occupancy models described in (1), (2) and (3) above. This mathematical formulation of the model can be translated for E-SURGE as follows. In GEPAT, the vector of initial state probabilities is specified as

```
* Ψ
```

which corresponds to (1). Recall that the dead state is not displayed at this step. Then, the transition matrix in (2) is specified as

```
* γ -
ε * -
- - *
```

where the last row and column are for the dead state. Finally, the observation matrix in (3) is coded as:

```
* -
* p
* -
```

where the last row corresponds to the dead state. In GEMACO, we write `i` at the Initial state and Event steps to impose a constant parameter. Regarding the Transition step, the data were simulated with 3 seasons (primary sessions) and 3 visits within each season (secondary session), therefore we use `to.t(1 2 4 5 7 8)+to.t(3 6)` (or `to.t(1 2 4 5 7 8,3 6)`). The `to` makes the columns of the matrix different, resulting here in distinguishing the colonization and extinction parameters (note that `from` would produce exactly the same result

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

by differentiating the rows). To handle the robust design, `t(1 2 4 5 7 8)` puts the intervals between secondary occasions together, and the corresponding parameters will be fixed to 0 at the IVFV step to impose closure within primary session (neither extinction nor colonization). The term `t(3 6)` puts together the intervals between the last visit in a season and the first visit of the next one (between primary sessions).

Single-season models

We start with the simplest model as described above. In GEPAT, the specification for the Initial state and the Event steps is exactly the same as for the dynamic model. To satisfy the closure assumption, we impose neither colonization nor extinction between sampling occasions by specifying at the Transition step:

```
* - -
- * -
- - *
```

In GEMACO, we use `i` for all steps to obtain constant parameters.

The extension of this model to multiple states with uncertainty is obtained as follows in E-SURGE. In GEPAT, the initial state probabilities in (4) are:

```
* Ψ Ψ
```

while the Transition step is the same as before. The originality lies in the specification of the observation process that is accomplished in two steps to match (5) in GEPAT: step 1 and the detection matrix is:

```
* - - -
* p - -
* - p -
* - - -
```

while step 2 gives the assignment matrix and is:

```
* - -
- * -
- * δ
* - -
```

In GEMACO, we use `to` at the Initial state step to distinguish the occupancy probabilities according to states, `from` for the step 1 of the Event step to distinguish the detection probabilities according to states and `i` for step 2 to have a constant assignment probability.

FURTHER E-SURGE CAPABILITIES

E-SURGE offers the possibility to include heterogeneity in the detection using finite mixtures, to fit multiple species models can also be fitted in E-SURGE, incorporate covariates measured at the site or season level as well as site random effects. The companion wiki website http://occupancyinesurge.wikidot.com/ presents such examples.

CONCLUSIONS

Although initially developed for capture-recapture data, E-SURGE can be efficiently used to build and analyse a variety of occupancy models via the HMM framework. E-SURGE includes a user-friendly syntax for specifying models without having to write custom code and used advanced numerical algorithms to produce fast and reliable results. By making the link between the two fast growing user communities of capture-recapture and occupancy, E-

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

SURGE has the potential to provide a unified framework for the construction and analysis of hidden-Markov models in ecology.

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

References

Bailey, L.L., MacKenzie, D.I. & Nichols, J.D. (2013) Advances and applications of occupancy models. *Methods in Ecology and Evolution*.

Choquet, R. (2008) Automatic generation of multistate capture recapture models. *The Canadian Journal of Statistics*, **36**, 43–57.

Choquet, R. & Cole, D.J. (2012) A Hybrid Symbolic-Numerical Method for Determining Model Structure. *Mathematical Biosciences*, **236**, 117–125.

Choquet, R. & Nogué, E. (2013) *E-SURGE 1.8 User's Manual*. Montpellier.

Choquet, R., Rouan, L. & Pradel, R. (2009) Program E - SURGE : A Software Application for Fitting Multievent Models. *Environmental and Ecological Statistics* (eds D.L. Thomson, E.G. Cooch & M.J. Conroy), pp. 845–865. Springer US.

Fiske, I.J. & Chandler, R.B. (2011) unmarked : An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. *Journal Of Statistical Software*, **43**, 1–23.

Gimenez, O., Lebreton, J.-D., Gaillard, J.-M., Choquet, R. & Pradel, R. (2012) Estimating demographic parameters using hidden process dynamic models. *Theoretical Population Biology*, **82**, 307–316.

Hines, J.E. (2013) PRESENCE 5.9 – Software to estimate patch occupancy and related parameters.

Kéry, M. & Schaub, M. (2011) *Bayesian Population Analysis Using WinBUGS: A Hierarchical Perspective*. Academic Press.

MacKenzie, D.I., Nichols, J.D., Royle, J.A., Pollock, K.H., Bailey, L.L. & Hines, J.E. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*. Academic Press.

PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

Nichols, J.D., Hines, A.J.E., Mackenzie, D.I., Seamans, M.E. & Gutiérrez, R.J. (2007) Occupancy estimation and modeling with multiple states and state uncertainty. *Ecology*, **88**, 1395–1400.

Pradel, R. (2005) Multievent: an extension of multistate capture-recapture models to uncertain states. *Biometrics*, **61**, 442–7.

Royle, J.A. & Kéry, M. (2007) A Bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–23.

White, G.C. & Burnham, K.P. (1999) Program MARK: survival estimation from populations of marked animals. *Bird Study*, **46**, 120–139.

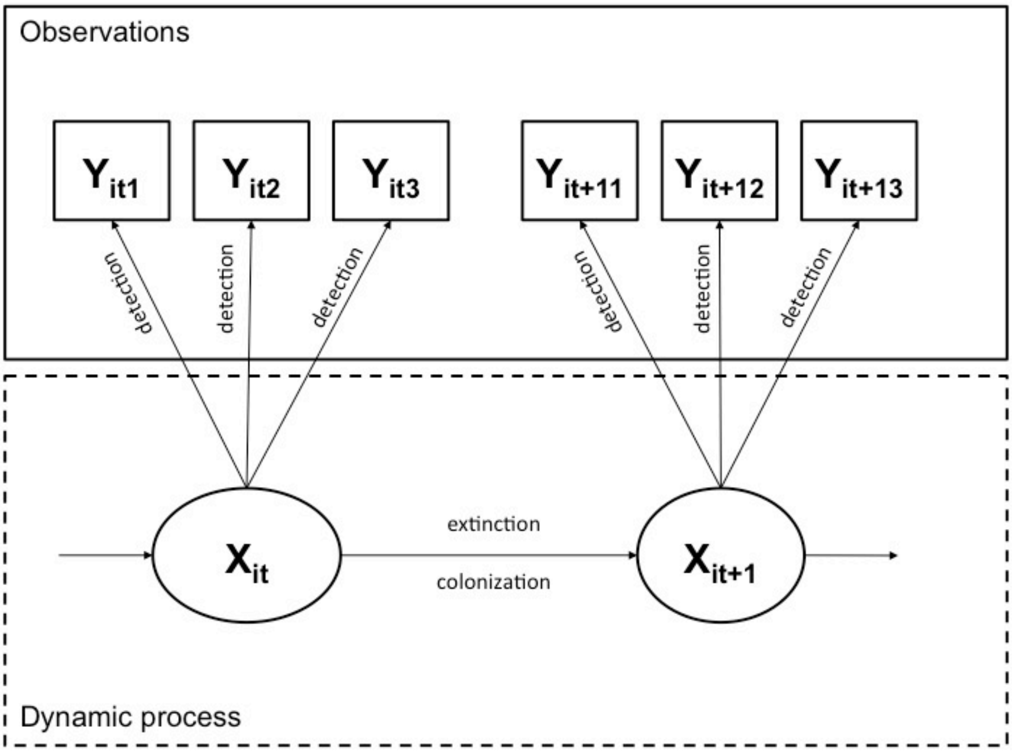
PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

Table 1. Estimates of the parameters in the occupancy models fitted to the simulated data set in E-SURGE. For each parameter, the true value, the maximum likelihood estimate and the 95% confidence interval are provided. See text for details.

Model	**Parameter**	**True value**	**Estimate**	**95% confidence interval**
Single-season	ψ	0.8	0.8	(0.70, 0.87)
	p	0.5	0.5	(0.44, 0.56)
Multistate with uncertainty	ψ^1	0.3	0.30	(0.23, 0.37)
	ψ^2	0.5	0.50	(0.43, 0.57)
	p^1	0.5	0.52	(0.43, 0.61)
	p^2	0.7	0.70	(0.65, 0.76)
	δ	0.8	0.79	(0.73, 0.85)
Dynamic model	ψ_1	0.6	0.59	(0.53, 0.65)
	γ	0.3	0.32	(0.27, 0.39)
	ε	0.5	0.53	(0.47, 0.59)
	p	0.7	0.70	(0.67, 0.73)

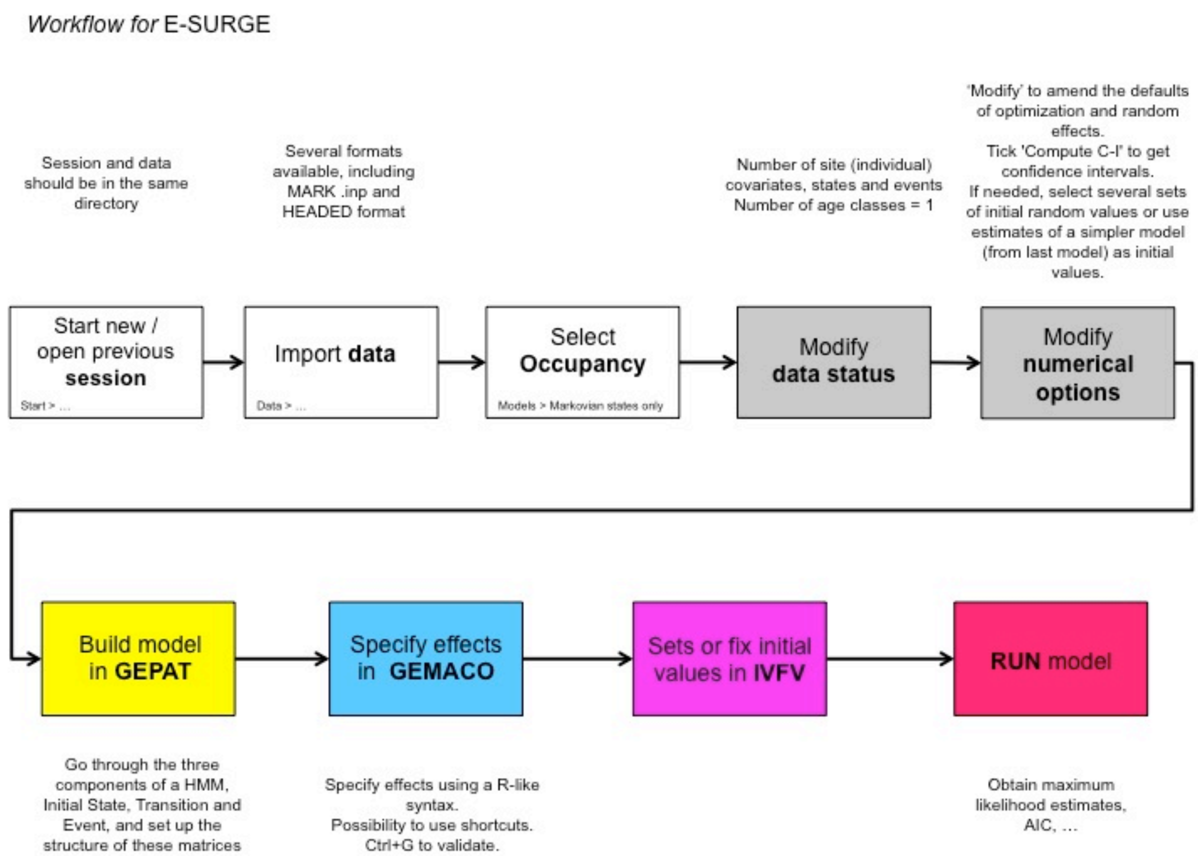
PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

Figure 1. Schematic hidden Markov representation of a dynamic occupancy model with imperfect detection. We consider site *i* between seasons (primary periods) *t* and *t* + 1 and three visits (secondary periods) within each primary period. Each site is closed (i.e. no change in its occupancy status) within primary periods but open (i.e., allowing for changes in occupancy status) between primary periods. The first layer (circles) is a succession of hidden states or latent states (0 for unoccupied vs. 1 for occupied) of site *i* at season *t +1* (X_{it+1}) depending on its states at time *t* (X_{it}). The dynamic of the states is driven by transition probabilities, here probabilities of colonization $\gamma = \Pr\left(X_{it+1} = 1 \middle| X_{it} = 0\right)$ and local extinction $\varepsilon = \Pr\left(X_{it+1} = 0 \middle| X_{it} = 1\right)$. The second layer (squares) corresponds to the detection ($Y_{itj} = 1$) or not ($Y_{itj} = 0$)of the target species on site *i* at visit *j* = 1, 2 or 3 conditional on site *i* being in state X_{it}. These events are driven by the species detection probability $p = \Pr\left(Y_{itj} = 1 \middle| X_{it} = 1\right)$. The initial state probabilities $\psi_1 = \Pr\left(X_{i1} = 1\right)$ and $1 - \psi_1$ are not represented.



PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1

Figure 2. Workflow diagram for E-SURGE. We describe the successive steps of a typical analysis in E-SURGE, from data input to model fitting through model building and effects specification. Steps that need to be accomplished through pull-down menus are in white boxes, the others can be done directly from the main interface. We provide details on key steps in using E-SURGE in the text above or below the boxes.



PeerJ PrePrints | https://peerj.com/preprints/84v1/ | v1 received: 20 Oct 2013, published: 20 Oct 2013, doi: 10.7287/peerj.preprints.84v1