

# Something occupancy models

Lauren C. Ponisio<sup>1,2</sup>, Nicholas Michaud<sup>1</sup>, Perry de Valpine<sup>1</sup> (Daniel and Chris?)

1. Department of Environmental Science, Policy, and Management  
University of California, Berkeley  
130 Mulford Hall  
Berkeley, California, USA  
94720
2. Department of Entomology  
University of California, Riverside  
417 Entomology Bldg.  
Riverside, California, USA  
92521

## Abstract

1. Something expressing the sentiment: occupancy models are everywhere, but model fitting and assessment are extremely computationally intensive
2. Because models are so computationally intensive, users often forgo model assessment (determining if a model provides an adequate fit to a particular dataset) and model selection (choosing the best model out of a set of models) — methods that generally involve simulating from and refitting the model iteratively.
3. Using the open source modeling software NIMBLE, we develop combined computational approaches including user-defined and automatic blocking of parameters for MCMC, filtering over latent states, and customized MCMC samplers for specific parameters to improve efficiency. We test these approaches using three representative occupancy models of varying levels of complexity including a single species model with spatial auto-correlation, a single species dynamic (multi-season) model, and a multi-species model. We also develop and implement methods for calculating calibrated predictive posterior  $p$ -values to assess model fit and cross validation for model selection within NIMBLE.
4. These computation approaches lead to an improvement in MCMC sampling efficiency over, particularly with models including random effects. (more results once they are available)
5. *Implications:* Ours results highlight the need for more customizable approaches to MCMC to fit and assess hierarchical models in order to ensure occupancy and other hierarchical models are accessible to practitioners. By implementing MCMC procedures and model assessment and selection techniques in open source software, we have made progress toward this aim.

NIMBLE, Markov chain Monte Carlo, latent states, block sampling, dynamic occupancy, multi species occupancy, spatial occupancy, JAGS

## Introduction

Estimating the proportion of sites occupied by a species is common challenge for many sub-disciplines in ecology and evolution including meta-population, endangered species and invasion biology. Greater acceptance of the biases introduced by imperfect detection has lead to the development and proliferation of occupancy models where the occurrence of a species at a site is modeled as a latent state layered underneath a detection process (e.g., MacKenzie *et al.*, 2006; Royle & Kéry, 2007a). Now only a little over a decade after occupancy models were introduced to ecology, they are being used to model the occurrence of everything from bees (M'Gonigle *et al.*, 2015) to tigers (Hines *et al.*, 2010) in an endless variety of complexity.

Occupancy models are part of a larger class of models known as Hidden Markov Models. In discrete Hidden Markov Models like occupancy models where a species is either present or absent from a site, likelihood calculation involves summing over the distribution of latent states. Because estimating the effect of explanatory variables on site occupancy or shared variation in occupancy across species is often of greatest interest to biologists (e.g., Iknayan *et al.*, 2014), the Hidden Markov Model is often embedded within a hierarchical model. In such cases, practitioners generally rely on Markov chain Monte Carlo (MCMC) to perform a Bayesian analysis. Standard MCMC software will include the latent state variables in MCMC sampling (e.g., Plummer *et al.*, 2003; WinBUGS; OpenBUGS). Such models are computationally intensive, and large models requiring hundreds or thousands of dimensions which require MCMC can be intractable.

In addition, fitting these models is such a challenge that users often forgo any additional computation to asses model fit. A common idea behind evaluating whether a model provides an adequate fit to a dataset is that if data is simulated from the model, the simulated

data should resemble the observed data. This is the basis of posterior predictive  $p$ -values, which compare the distribution of summary statistics calculated from simulated datasets to the observed statistic. Posterior predictive  $p$ -values alone, however, often fail to reject poor-fitting models (Bayarri & Berger, 2000; Robins *et al.*, 2000; Hjort *et al.*, 2006). Methods for correcting posterior predictive  $p$ -values for better performance by refitting the model via MCMC iteratively have been proposed (e.g., calibrated posterior predictive  $p$ -values, Hjort *et al.*, 2006), but no methods are available in open source software. In addition, given fitting an occupancy model just once can be a time consuming task, efficient methods for MCMC are necessary to ensure methods for assessment are feasible for these models.

Beyond assessing the fit of a model, choosing between models is one of the most widely used applications of statistics by practitioners. Though many methods for Bayesian model selection have been developed (Hooten & Hobbs, 2014), but they, like model assessment, are computationally intensive. For example, cross-validation, one of the most fundamental procedures in model selection, requires iteratively re-fitting the model. A typical need for model selection arises when a practitioner is choosing whether to include a specific layer of hierarchy (i.e., random effect). This is often the case with so called “multi-species” occupancy models, where the occupancy of many species is estimated simultaneously in a model with a random effect of species (reviewed in, Iknayan *et al.*, 2014). Ecologists are often interested in whether there is some variability in the response of species to an explanatory variable such that a random effect of species accounts for that variability (Pacifi *et al.*, 2014). Currently, the Deviance Information Criteria (DIC), originally derived to mimic AIC for Bayesian, non-hierarchical models, is now commonly used by scientists to evaluate hierarchical models. Though the limitations of DIC for hierarchical model selection are widely recognized by statisticians (Celeux *et al.*, 2006; Hooten & Hobbs, 2014), because it is built into open-source software such as WinBUGS, it can be

77 used uncritically by practitioners. Readily available and theoretically sound alternative  
78 methods are thus greatly needed.

79 Luckily, methods to improve MCMC efficiency of Hidden Markov Models have been  
80 developed such filtering over latent states to calculate model likelihoods in order to limit  
81 MCMC sampling to top-level parameters dynamic blocking or parameters (Turek *et al.*,  
82 2016). A synergistic strategy is to assign specific MCMC samplers to different parameters  
83 depending on the nature of those nodes (i.e., discrete versus continuous). Though these  
84 methods are available in isolation in application-specific software, they cannot be used in  
85 combination for any arbitrary model structure.

86 Using the open source modeling software NIMBLE, we develop combined computational  
87 approaches including user-defined and automatic blocking of parameters for MCMC,  
88 filtering over latent states, and customized MCMC samplers for specific parameters to  
89 improve efficiency. We test these approaches using three representative occupancy mod-  
90 els of varying levels of complexity including a single species model with spatial auto-  
91 correlation, a single species dynamic (multi-season) model, and a multi-species model.  
92 We also develop and implement methods for calculating calibrated predictive posterior  $p$ -  
93 values to assess model fit and cross validation for model selection within NIMBLE.

## Materials & Methods

### Computational approaches

#### Single species, single season occupancy model with spatial auto-correlation

The first model we explore is a single species, single season occupancy model accounting for spatial auto-correlation. We let  $z_i$  denote the true occupancy of a species at site  $i$ . We then let  $x_{i,j}$  indicate whether the species was ( $x_{i,j} = 1$ ), or was not detected ( $x_{i,j} = 0$ ) in the  $j^{\text{th}}$  visit to site  $i$ . We assumed that occupancy at the  $i^{\text{th}}$  site is a Bernoulli random variable  $z_i \sim \text{Bern}(\psi_i)$  with probability  $\psi_i$ . We included the effect of an arbitrary covariate (e.g., elevation) on site occupancy. To model the spatial auto-correlation in occupancy between sites, we assume the co-variance between sites  $Y_i$  and  $Y_j$  is a function of distance between  $p_i$  and  $p_j$ . We computed the probability of occupancy at site  $i$

$$\begin{aligned}\text{logit}(\psi_i) &= \alpha + \beta * \text{elevation}_i + \rho_i \\ \rho_i &\sim \text{MVN}(0, \text{Cov}(Y_i, Y_j)) \\ \text{Cov}(Y_i, Y_j) &= \sigma^2 \exp(-\lambda \|p_i - p_j\|) .\end{aligned}\tag{1}$$

Where  $\lambda$  is the exponential decay constant and  $\sigma^2$  is SOMETHING...

We simulate data for this model and then fit it using the default settings for NIMBLE and JAGS. To improve efficiency of this model...

We implemented a procedure to calculate calibrated posterior predictive  $p$ -values (CPPP, Hjort *et al.*, 2006) to assess the fit of the model to the data. After the parameters have been fit to the model, a sample of the posterior is used to simulate data from the model. A discrepancy measure, which we chose to be the model likelihood, is then calculated, and

the posterior  $p$ -value is the number of simulated  $p$ -values that fall below the observed. To “calibrate” the distribution of posterior  $p$ -values, the MCMC is rerun on the simulated data to refit the model. If the CPPP  $< 0.05$ , the model is rejected as having an adequate fit of the data (Hjort *et al.*, 2006).

### Single species, multi season (dynamic) occupancy model

The second model we examine is a relatively simple single species occupancy model over multiple seasons (Royle & Kéry, 2007b). We let  $z_{i,j}$  denote the true occupancy of a species in year  $j$  at site  $i$ . We assumed that occupancy at the  $i^{\text{th}}$  site in the  $j^{\text{th}}$  year is a Bernoulli random variable  $z_{i,j} \sim \text{Bern}(\psi_{i,j})$ .

Letting  $\phi_{i,j}$  denote the probability the species persists at site  $i$  from years  $j$  to  $j + 1$  (provided it was present at site  $i$  in year  $j$ ,  $z_{i,j} = 1$ ) and  $\gamma_{i,j}$  denote the probability that site  $i$  is colonized in year  $j + 1$  (provided it was not present at site  $i$  in year  $j$ ,  $z_{i,j} = 0$ ), we then computed the probability of occupancy at site  $i$  in subsequent years as

$$\psi_{i,j+1} = \phi_{i,j} * z_{i,j} + \gamma_{i,j} * (1 - z_{i,j}) . \quad (2)$$

We then let  $x_{i,j,k}$  indicate whether that species was ( $x_{i,j,k} = 1$ ) or was not detected ( $x_{i,j,k} = 0$ ) in the  $k^{\text{th}}$  visit to site  $i$  in year  $j$ . We assume detection was distributed according to be a Bernoulli random variable such that  $x_{i,j,k} \sim \text{Bern}(p_j * z_{i,j})$ , where  $p_j$  is the probability that the species was detected at site  $i$  in the  $j^{\text{th}}$  year, given that it was present.

As with the spatial occupancy model, we first simulate data for this model and then fit it using the default settings for JAGS and NIMBLE where all model parameters and latent states undergo MCMC sampling (“NIMBLE-latent” and “JAGS-latent”, respectively). Following Royle & Kéry (2007b); Kery & Schaub (2011), we use uninformative,  $\text{Unif}(0, 1)$

priors for all parameters.

Next, to improve efficiency, using NIMBLE we filter over latent states to calculate model likelihoods in order to limit MCMC sampling to top-level parameters (“filter”). [Do we want to write out the likelihood?](#) We then use two additional computational approaches to improve the efficiency of this model 1) dynamic blocking of the parameters (“filter + autoblocking”, Turek *et al.*, 2016), and 2) a custom MCMC specification where slice samplers (Neal, 2003) are used for all parameters (“filter + slice”). Slice samplers are a class of methods that sample from a target distribution by using that fact that samples from any distribution can be obtained by sampling uniformly from the area under that distribution’s probability density function curve. The horizontal coordinates of these uniform samples will provide samples from the distribution of interest. Slice samplers have been shown to perform well in situations where choosing a tuning parameter for a Metropolis algorithm is difficult. When used to sample from the posterior distribution of a univariate parameter, a slice sampler proceeds at each iteration by first choosing a vertical coordinate sampled uniformly between 0 and the height of the density curve at the parameter value from the previous iteration. Then, a horizontal coordinate is chosen uniformly from the set of all possible parameter values whose density is at least as great as the chosen vertical coordinate. [other options we want to present?](#)

As with the spatail model, we use calibrated posterior predictive  $p$ -values (Hjort *et al.*, 2006) to assess the fit of the model to the data.

### **Multi species, single season occupancy model**

The last model we analyze is a multi-species, single season occupancy model examining the effect of wildlife management and habitat characteristics on bird communities (Zipkin *et al.*, 2010). The species-specific coefficients for the effect of basal tree area, understory



foliage and deer management where bound together by a common distribution with an  
 estimated variance. For species  $i$ , we let  $z_{i,j}$  denote its true occupancy state at site  $j$ . We  
 assumed that the occupancy of the  $i^{\text{th}}$  species at the  $j^{\text{th}}$  site is a Bernoulli random variable  
 $z_{i,j} \sim \text{Bern}(\psi_{i,j})$ . We then let  $x_{i,j,k}$  indicate whether species  $i$  was ( $x_{i,j,k} = 1$ ) or was not de-  
 tected ( $x_{i,j,k} = 0$ ) in the  $k^{\text{th}}$  visit to site  $j$ . We also assumed that detection was distributed  
 according to be a Bernoulli random variable such that  $x_{i,j,k} \sim \text{Bern}(p_i * z_{i,j})$ , where  $p_i$  is  
 the probability that the  $i^{\text{th}}$  species was detected. Both site occupancy and detection were  
 influence by habiat and survey characteristics (Zipkin *et al.*, 2010). Specifically, occurence  
 depended on the study area (CATO, Ind=1, or FCW, Ind=0), the basal tree area (BA) and  
 the understory foliage cover (UFC). The species-specific occupancy probabilities are mod-  
 eled as

$$\begin{aligned}
 \text{logit}(\psi_{i,j}) &= u\text{CATO}_i(\text{Ind}_j) + u\text{FCW}_i(1 - \text{Ind}_j) + \alpha_1 \text{UFC}_j + \alpha_2 \text{UFC}_j^2 + \alpha_3 \text{BA}_j + \alpha_4 \text{BA}_j^2 \\
 \alpha_1 &\sim N(\mu_{\alpha_1}, \sigma_{\alpha_1}^2) \\
 \alpha_2 &\sim N(\mu_{\alpha_2}, \sigma_{\alpha_2}^2) \\
 \alpha_3 &\sim N(\mu_{\alpha_3}, \sigma_{\alpha_3}^2) \\
 \alpha_4 &\sim N(\mu_{\alpha_4}, \sigma_{\alpha_4}^2)
 \end{aligned}
 \tag{3}$$

Similarly, detection depended on survey location and the date:

$$\begin{aligned}
\text{logit}(p_{i,j,k}) &= vCATO_i(Ind_j) + vFCW_i(1 - Ind_j) + \beta1_i date_j + \beta2_i date_j^2 \\
\beta1 &\sim N(\mu_{\beta1}, \sigma_{\beta1}^2) \\
\beta2 &\sim N(\mu_{\beta2}, \sigma_{\beta2}^2)
\end{aligned} \tag{4}$$

We first fit the using the default settings for JAGS and NIMBLE where all model parameters and latent states undergo MCMC sampling (“NIMBLE-latent” and “JAGS-latent”, respectively). We use uninformative priors, Norm(0, 1000) for the means of the distributions of the hyperparameters and Unif(0, 100) the variances.

To improve the efficiency of this model, we first filtered over latent states to calculate model likelihoods in order to limit MCMC sampling to top-level parameters (“filter”). We also vectorized all calculations that would have require for loops in JAGS. [Do we want to write out the likelihood?](#) We then applied two approaches to speed sampling of the top-level parameters 1) dynamic blocking of the parameters (“filter + autoblocking”, Turek *et al.*, 2016), and 2) a custom blocking scheme where the parameters of each species are blocked together with adaptive random walk MCMC (“Filter + species blocking”). Random walk Metropolis algorithms (Metropolis *et al.*, 1953) sample from the posterior distribution of a parameter of interest by first proposing a new parameter value from a normal proposal distribution centered at the current value, and then deciding whether to accept or reject the proposed value via the Metropolis ratio. Although common, standard univariate or multivariate normal proposal distributions can prove to be inefficient at sampling for models in which parameters are highly correlated. Adaptive random walk Metropolis sampling (Haario *et al.*, 1998) allows the correlation structure of the posterior distribution of blocks of parameters to be used to produce better proposals. For an oc-

cupancy model where parameters are highly correlated within species, this can provide much more efficient sampling than a non-adaptive Metropolis algorithm.

We also use calibrated posterior predictive  $p$ -values (Hjort *et al.*, 2006) to assess the fit of the model to the data. In multi-species occupancy models, practitioners are often interested in determining whether a model including a species random effect for explanatory variables is a better fit than a model without the random effect. We implemented a cross-validation procedure for this model where the detection data for species is left out, the model refitted, and the fitted model used to predict the occurrence of that species. The predictive error of the model included a random effect of species is then compared to a model where no species random effects were included.

## Results

Single species, single season occupancy model with spatial auto-correlation

Multi species, single season occupancy model

Single species, multi season (dynamic) occupancy model

## Discussion

## Acknowledgments

## References

- Bayarri, M. & Berger, J. (2000) P values for composite null models. *Journal of the American Statistical Association*, **95**, 1127–1142.
- Celeux, G., Forbes, F., Robert, C.P., Titterton, D.M. *et al.* (2006) Deviance information criteria for missing data models. *Bayesian Analysis*, **1**, 651–673.
- Haario, H., Saksman, E. & Tamminen, J. (1998) An adaptive metropolis algorithm. *Bernoulli*, **7**, 223–242.
- Hines, J., Nichols, J., Royle, J., MacKenzie, D., Gopalaswamy, A., Kumar, N. & Karanth, K. (2010) Tigers on trails: occupancy modeling for cluster sampling. *Ecological Applications*, **20**, 1456–1466.
- Hjort, N.L., Dahl, F.A. & Hognadottir, G. (2006) Post-processing posterior predictive p values. *Journal of the American Statistical Association*, **101**, 1157–1174.

- 216 Hooten, M.B. & Hobbs, N.T. (2014) A guide to Bayesian model selection for ecologists.  
217 *Ecological Monographs*, pp. in press, online early.
- 218 Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity:  
219 emerging methods to estimate species diversity. *Trends in ecology & evolution*, **29**, 97–  
220 106.
- 221 Kery, M. & Schaub, M. (2011) *Bayesian Population Analysis using WinBUGS: A hierarchical*  
222 *perspective*. Academic Press, Boston, 1 edition edition.
- 223 MacKenzie, D., Nichols, J., Royle, J., Pollock, K., Bailey, L. & Hines, J. (2006) *Occupancy*  
224 *estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier,  
225 Burlington, Massachusetts, USA.
- 226 Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equa-  
227 tion of state calculations by fast computing machines. *The journal of chemical physics*, **21**,  
228 1087–1092.
- 229 M'Gonigle, L., Ponisio, L., Cutler, K. & Kremen, C. (2015) Habitat restoration promotes  
230 pollinator persistence and colonization in intensively-managed agriculture. *Ecol Appl*,  
231 **25**, 1557–1565.
- 232 Neal, R.M. (2003) Slice sampling. *Annals of Statistics*, pp. 705–741.
- 233 OpenBUGS (????) OpenBUGS. <http://www.openbugs.net/w/FrontPage>.
- 234 Pacifici, K., Zipkin, E.F., Collazo, J.A., Irizarry, J.I. & DeWan, A. (2014) Guidelines for a  
235 priori grouping of species in hierarchical community models. *Ecology and evolution*, **4**,  
236 877–888.

- 237 Plummer, M. *et al.* (2003) Jags: A program for analysis of bayesian graphical models using  
 238 gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical*  
 239 *Computing (DSC 2003). March*, pp. 20–22.
- 240 Robins, J., van der Vaart, A. & Ventura, V. (2000) Asymptotic distribution of p values in  
 241 composite null models. *Journal of the American Statistical Association*, **95**, 1143–1156.
- 242 Royle, A. & Kéry, M. (2007a) A bayesian state-space formulation of dynamic occupancy  
 243 models. *Ecology*, **88**, 1813–1823.
- 244 Royle, A. & Kéry, M. (2007b) A bayesian state-space formulation of dynamic occupancy  
 245 models. *Ecology*, **88**, 1813–1823.
- 246 Turek, D., de Valpine, P. & Paciorek, C.J. (2016) Efficient markov chain monte carlo sam-  
 247 pling for hierarchical hidden markov models. *arXiv preprint arXiv:160102698*.
- 248 WinBUGS (????) WinBUGS. <http://www.mrc-bsu.cam.ac.uk/software/bugs/>.
- 249 Zipkin, E.F., Royle, J.A., Dawson, D.K. & Bates, S. (2010) Multi-species occurrence models  
 250 to evaluate the effects of conservation and management actions. *Biological Conservation*,  
 251 **143**, 479–484.