Running head: OCCUPANCY MODEL EFFICIENCY

Improving the efficiency of occupancy models (something more interesting sounding?)

Lauren C. Ponisio[1,2,3], Nicholas Michaud[1], Perry de Valpine[1] (Daniel and Chris?)

1. Department of Environmental Science, Policy, and Management
University of California, Berkeley
130 Mulford Hall
Berkeley, California, USA
94720

2. Berkeley Institute for Data Science (BIDS)
University of California, Berkeley
190 Doe Library

3. 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 iterativly.
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 (maybe?). (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, mutli 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 iterativly 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 iterativly 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 (Pacifici *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 used uncritically

by practitioners. Readily available and theoretically sound alternative methods are thus greatly needed.

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

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 autocorrelation, 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.

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^{th} visit to site i. We assumed that occupancy at the i^{th} site is a Bernoulli random variable $z_i \sim \text{Bern}(\psi_i)$ with probability ψ_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

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

Where λ is the exponential decay constant and σ^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 to 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^{th} site in the j^{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

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

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^{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^{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 spatial 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^{th} species at the j^{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 detected ($x_{i,j,k} = 0$) in the k^{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^{th} species was detected. Both site occupancy and detection were influence by habitat and survey characteristics (Zipkin *et al.*, 2010). Specifically, occurrence 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 modeled as

$$
\begin{aligned}
\text{logit}(\psi_{i,j}) &= uCATO_i(Ind_j) + uFCW_i(1 - Ind_j) + \alpha 1_i UFC_j + \alpha 2_i UFC_j^2 + \alpha 3_i BA_j + \alpha 4_i BA_j^2 \\
\alpha 1 &\sim N(\mu_{\alpha 1}, \sigma^2_{\alpha 1}) \\
\alpha 2 &\sim N(\mu_{\alpha 2}, \sigma^2_{\alpha 2}) \\
\alpha 3 &\sim N(\mu_{\alpha 3}, \sigma^2_{\alpha 3}) \\
\alpha 4 &\sim N(\mu_{\alpha 4}, \sigma^2_{\alpha 4})
\end{aligned}
\tag{3}
$$

Similarly, detection depended on survey location and the date:

$$
\begin{aligned}
\text{logit}(p_{i,j,k}) &= vCATO_i(Ind_j) + vFCW_i(1 - Ind_j) + \beta 1_i date_j + \beta 2_i date_j^2 \\
\beta 1 &\sim N(\mu_{\beta 1}, \sigma^2_{\beta 1}) \\
\beta 2 &\sim N(\mu_{\beta 2}, \sigma^2_{\beta 2})
\end{aligned}
\tag{4}
$$

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, $\text{Norm}(0, 1000)$ for the means of the distributions of the hyperparameters and $\text{Unif}(0, 100)$ 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., Titterington, 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.

Hooten, M.B. & Hobbs, N.T. (2014) A guide to Bayesian model selection for ecologists. *Ecological Monographs*, pp. in press, online early.

Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diversity: emerging methods to estimate species diversity. *Trends in ecology & evolution*, **29**, 97–106.

Kery, M. & Schaub, M. (2011) *Bayesian Population Analysis using WinBUGS: A hierarchical perspective*. Academic Press, Boston, 1 edition edition.

MacKenzie, D., Nichols, J., Royle, J., Pollock, K., Bailey, L. & Hines, J. (2006) *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier, Burlington, Massachusetts, USA.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. & Teller, E. (1953) Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**, 1087–1092.

M'Gonigle, L., Ponisio, L., Cutler, K. & Kremen, C. (2015) Habitat restoration promotes pollinator persistence and colonization in intensively-managed agriculture. *Ecol Appl*, **25**, 1557–1565.

Neal, R.M. (2003) Slice sampling. *Annals of Statistics*, pp. 705–741.

OpenBUGS (????) OpenBUGS. http://www.openbugs.net/w/FrontPage.

Pacifici, K., Zipkin, E.F., Collazo, J.A., Irizarry, J.I. & DeWan, A. (2014) Guidelines for a priori grouping of species in hierarchical community models. *Ecology and evolution*, **4**, 877–888.

Plummer, M. *et al.* (2003) Jags: A program for analysis of bayesian graphical models using gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March*, pp. 20–22.

Robins, J., van der Vaart, A. & Ventura, V. (2000) Asymptotic distribution of p values in composite null models. *Journal of the American Statistical Association*, **95**, 1143–1156.

Royle, A. & Kéry, M. (2007a) A bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–1823.

Royle, A. & Kéry, M. (2007b) A bayesian state-space formulation of dynamic occupancy models. *Ecology*, **88**, 1813–1823.

Turek, D., de Valpine, P. & Paciorek, C.J. (2016) Efficient markov chain monte carlo sampling for hierarchical hidden markov models. *arXiv preprint arXiv:160102698*.

WinBUGS (????) WinBUGS. http://www.mrc-bsu.cam.ac.uk/software/bugs/.

Zipkin, E.F., Royle, J.A., Dawson, D.K. & Bates, S. (2010) Multi-species occurrence models to evaluate the effects of conservation and management actions. *Biological Conservation*, **143**, 479–484.

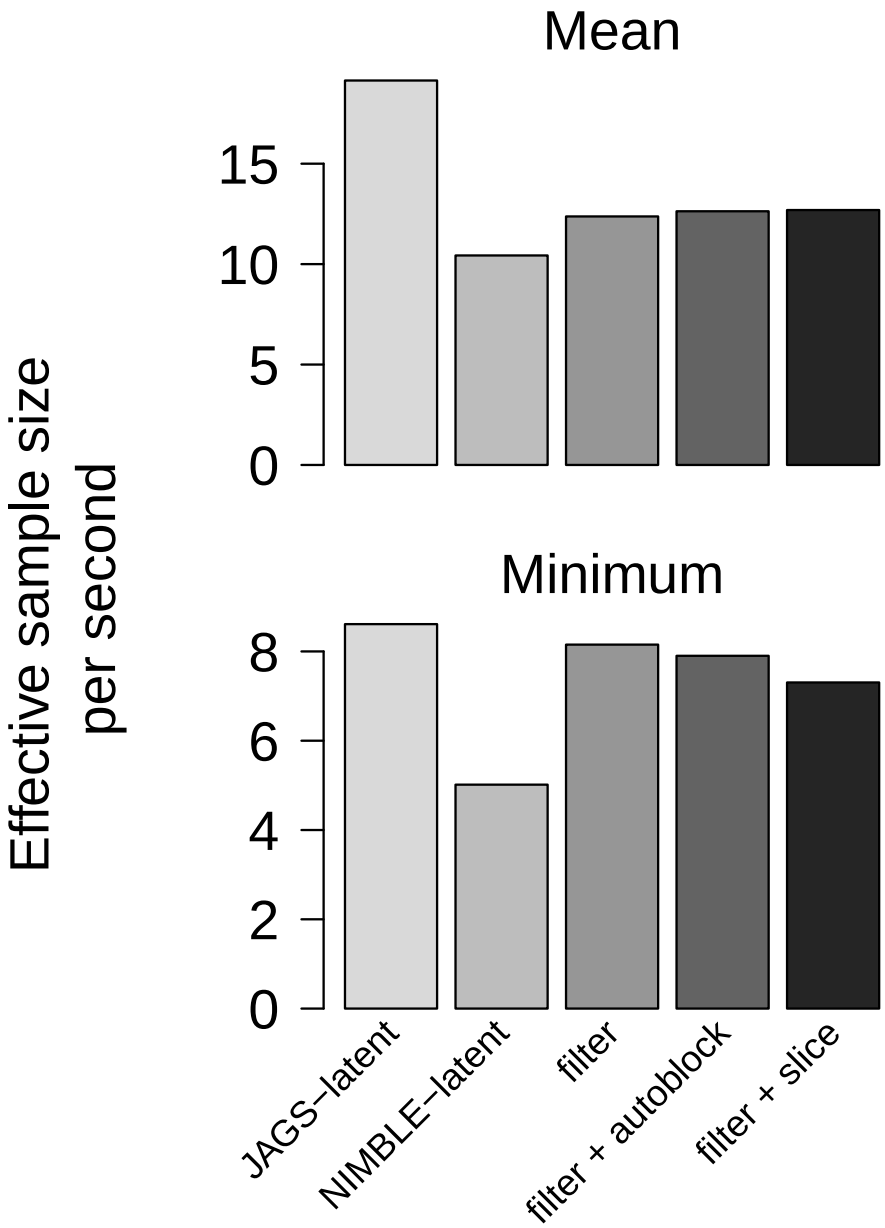

Figure 1: The mean and minimum efficiency of different MCMC samplers for the single species, multi season occupancy model.

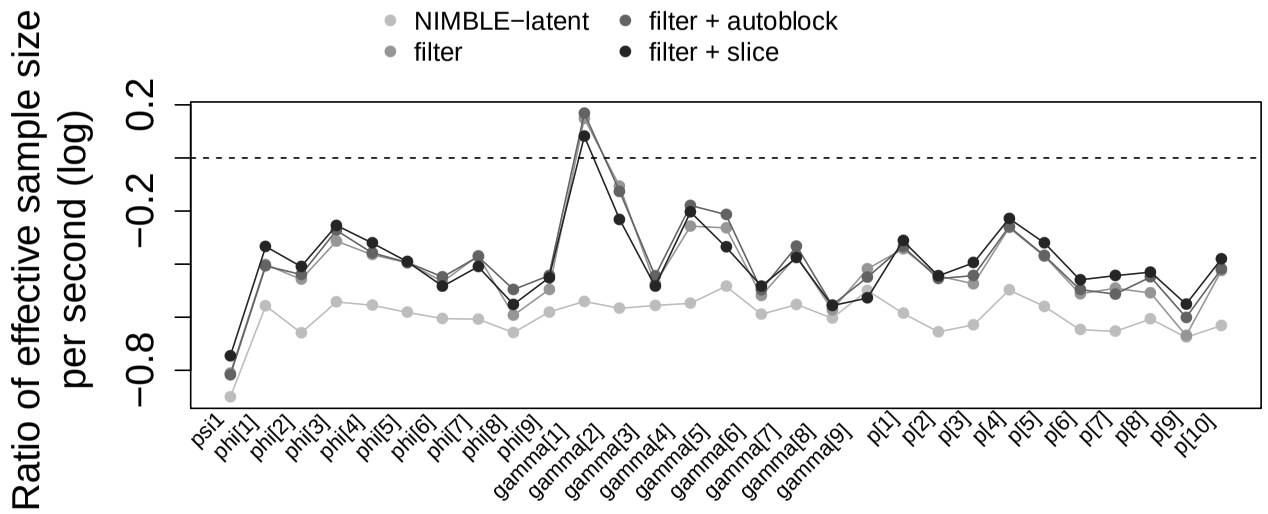

Figure 2: The log ratio of the NIMBLE samplers in comparison to the JAGS sampler for the single species, multi season occupancy model.

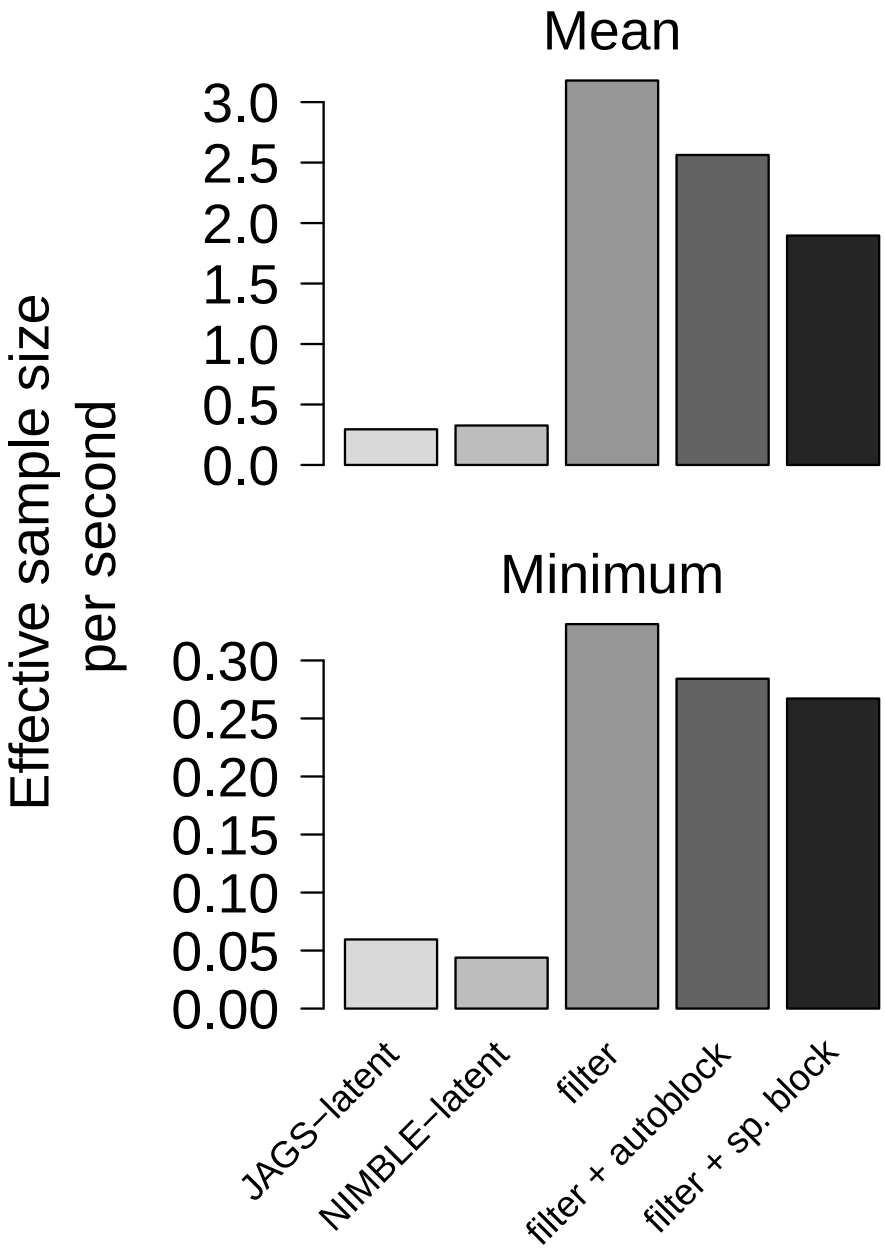

Figure 3: The mean and minimum efficiency of different MCMC samplers for the multi species, single season occupancy model.

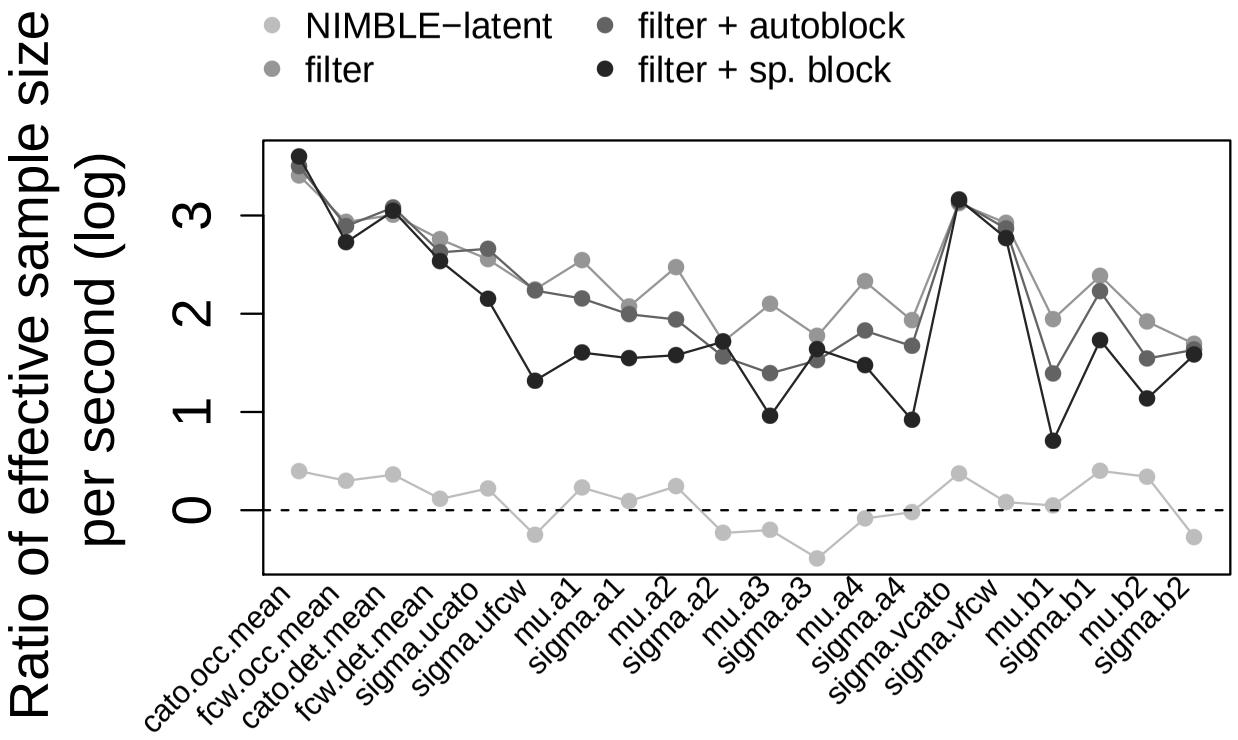

Figure 4: The log ratio of the NIMBLE samplers in comparison to the JAGS sampler for the multi species, single season occupancy model.