Specifying, Assessing and Selecting N-mixture Models in a Bayesian Framework

Fabian R. Ketwaroo, Eleni Matechou
Statistical Ecology @ Kent, School of Mathematics,
Statistics and Actuarial Science,
University of Kent

Abstract

Using only spatially replicated counts from unmarked individuals, N-mixture models provide an attractive framework to obtain estimates of population size by accounting for imperfect detection. The robustness of N-mixture models has been examined in detail in a classical inference framework. However, to our knowledge, only a small number of such studies have been carried out on N-mixture models in a Bayesian setting. In this paper, we consider fitting N-mixture models within a Bayesian framework. To aid implementation, we apply a new proper objective prior distribution to N-mixture models. Using simulated data, we compare this new proper objective prior to approximations of the popular objective prior, Jeffreys prior, and find that these prior distributions perform similarly in terms of model inference. Importantly, we find that when the detection probability is small, using priors that are concentrated at zero, even with large variance, expected population size can be considerably underestimated. Large estimates of expected population size were also found, evident by the bimodal density of posterior medians obtained for simulated data. Additionally, we consider an extensive class of N-mixture models and investigate model selection using the Watanable-Akaike Information Criterion (WAIC) in a wide range of scenarios to examine the sensitivity of WAIC to likelihood specification. We find that WAIC computed from the conditional likelihood produces misleading results favoring more complicated models than the true model. Contrary, WAIC computed using the marginal likelihood correctly selects the true model with a high probability. Hence, model selection of N-mixture models should be obtained from WAIC using the marginal likelihood, not the conditional likelihood. We demonstrate the usefulness/importance of employing these methods in two real datasets. Hence, this work can be considered a template for how to specify and select N-mixture models in a Bayesian context. We briefly investigate parameter identifiability of N-mixture models using Data cloning.

1 Introduction

A fundamental objective of many wildlife population monitoring programs and ecological studies is to estimate the size of a population. This is essential for the development and communication of management practices and guidelines. However, monitoring wildlife populations is challenging and costly, as the probability of detecting individuals in the monitored population is typically less than one. Survey sampling, which involves counting unmarked individuals in a given area over a specified period of time is relatively lower in cost and effort in comparison with other sampling methods, such as capture-recapture sampling and removal sampling.

Using count data from survey sampling, N-mixture models (Royle, 2004) are a class of hierarchical models that accounts for imperfect detection, allowing estimation of population size in a cost-effective way. N-mixture models have been used for a number of purposes, including evaluation of conservation actions (Romano et al., 2017), understanding population size and population dynamics (Studds et al., 2017), population prediction to conservation scenarios (Ladin et al., 2016) and to forecast shifts in species distributions (Hunter et al., 2017).

The performance of N-mixture models in a classical setting has been investigated in detail. Dennis et al. (2015) showed that infinite estimates of population size can arise when the probability of detection and the number of times the population is sampled are small. Barker et al. (2018) demonstrated the inability of count data to discriminate between different hierarchical models, even when these models yield substantially different estimates of population size. Knape et al. (2018) highlighted that estimated population size can be severely i) underestimated if the fitted model does not account for over-dispersion in the population process, when that is present or ii) overestimated if the fitted model does not account for over-dispersion in the detection process, when that is present. However, to our knowledge, only a small number of studies have investigated N-mixture models in a Bayesian framework (see for example Toribio et al., 2012; Link et al., 2018, who studied the robustness of the N-mixture model in a Bayesian setting). Thus, we consider fitting an extensive class of N-mixture models in a Bayesian framework, specifically focusing on prior specification and model selection, which are key aspects of Bayesian modelling.

An important question in Bayesian model building is how does one choose a prior distribution $p(\theta)$ for parameter θ? One can either be subjective: choosing priors that reflect some subjective opinion about θ (before data are collected) or objective: finding prior distributions that formally express ignorance about θ. Subjective priors have the appeal of using prior information to increase estimation precision without compromising accuracy (Morris et al., 2015), resulting in larger effective sample sizes and saved resources. However, care needs to be taken about how prior information is incorporated into the prior distribution, especially where there is limited prior information as, in the case of sparse data, which is often true in ecological applications, the prior can have a strong effect on the posterior distribution. Additionally, it can be difficult to quantify prior effects in practice.

Contrarily, objective and vague priors are two classes of priors that allow Bayesian inference when information about θ is not available. These priors aim to avoid bias in parameter estimation by placing less emphasis on prior beliefs and more emphasis on the data. Based on mathematical properties, objective priors are designed to reflect minimal information, and have certain mathematical properties, discussed in this

chapter. On the other hand, vague priors are deliberately chosen to convey no prior knowledge about the parameter being estimated, such as a flat prior or one with a very long tail, but without necessarily exhibiting the same mathematical properties as an objective prior. As a result, an objective prior may be more appropriate to express prior ignorance. Uniform distributions or normal distributions with large variances are common examples of vague priors. The Jeffreys prior (Jeffreys, 1946) is a popular objective prior designed to be invariant under reparameterization.

Notably, the majority of objective priors are improper (Leisen et al., 2018). A proper prior is a well-defined probability distribution as it integrates to 1 over the support of the parameter, whereas an improper prior has an infinite integral over the support of the parameter. In general, improper priors are not a problem as long as the resulting posterior is a proper probability distribution, from which one can derive moments such as the posterior mean. However, as of present, general results that allow one to assess if a given improper prior results in a proper posterior are yet to be developed (Leisen et al., 2018). Hence, caution is needed when using improper objective priors as spurious inference can be obtained. In addition, improper prior distributions cannot be applied in mixture models and model selection via Bayes factors (Leisen et al., 2018). Thus, this limits the use of many objective priors.

Banner et al. (2020) highlighted the use of priors in ecology and found vague priors are more often used in ecology than subjective priors. Both vague and objective priors have been used in N-mixture models: Royle (2015) demonstrated N-mixture models in a Bayesian framework by using priors that are approximations to Jeffreys prior, Link et al. (2018) used improper objective priors to study the robustness of N-mixture models, Toribio et al. (2012) used vague priors on the log and logit scale, which in turn resembled the Jeffreys prior on the original scale, to study the robustness of a Bayesian approach to fitting N-mixture models for pseudo-replicated count data. McCaffery et al. (2016) also used vague priors on the log and logit scale to analyze Lek count data.

In this paper, using a recently developed proper objective prior (Walker and Villa, 2021) and vague priors that are approximations to the Jeffreys prior, we test these priors and investigate the effect of prior choice in N-mixture models via an extensive simulation study.

N-mixture models can be relatively easily built in a Bayesian setting, but different models can result in substantially different estimates of population size (Ketwaroo, 2019). Therefore, it is imperative to have measures that allow one to compare models. Predictive accuracy measures can be used to compare models. Predictive accuracy measures simply compute how well a model estimated from available data generalises to out-of-sample data. However, the availability of out-of-sample data is often limited. One common way to overcome this deficiency is to use the sample data twice; once to fit the statistical model and again to test its predictive power. The issue here is that this can lead to over-fitting. Hence, predictive accuracy measures that use the data twice need to account for over-fitting. One such predictive accuracy measure is the Watanable-Akaike information criterion (WAIC, Watanabe, 2010). WAIC is often used in popular software such as NIMBLE (de Valpine et al., 2017) and Stan (Carpenter et al., 2017). Importantly, Ariyo et al. (2020) recently showed via an extensive simulation study that the marginal likelihood (averaging over latent variables) is superior to the conditional likelihood (given latent variables) when using WAIC to select the true longitudinal model. In addition, Millar (2018) showed using

over-dispersed count data that WAIC computed using the conditional likelihood is an unreliable tool for model selection and recommended using WAIC computed using the marginal likelihood. Thus, in this paper, we investigate whether WAIC can be used to select among the different N-mixture models considered and whether WAIC for N-mixture models is sensitive to the likelihood specification in a wide range of scenarios.

We fit N-mixture models considered using Markov Chain Monte Carlo (MCMC) methods provided by the R package NIMBLE (de Valpine et al., 2017) version 0.10.0.

Finally, we consider two real data sets, yellow-bellied toads (Ketwaroo, 2019) and Swiss great tits (Royle, 2015), and we investigate the usefulness/importance of employing these methods in each case.

The paper is organised as follows: Section 2 provides a detailed description of the different N-mixture models considered, prior specification, and model selection. Simulation results are presented in Section 3 and the results for the two case studies are presented in Section 4. Section 5 concludes the paper and provides ideas for potential future directions.

2 Materials and Methods

Assuming population closure, N-mixture models estimate population size and account for imperfect detection using only replicated counts at multiple sites. N-mixture models are composed of two key processes: a population size process describing the spatial variation in the number of individuals among sites and a detection process describing the detection of individuals at each site. Count data (hereafter C_{ij}) are obtained at $i = 1, \ldots, M$ sites with $j = 1, \ldots, J$ sampling occasions at each site.

For the population size process, it is assumed that the local population size at site i (hereafter N_i) is an independent random variable with a chosen discrete probability function g. That is,

$$\mathrm{N}_i \sim g(\mathrm{N}; \lambda_i, \gamma)$$

where λ_i represents the expected population size at site i and γ represents an optional parameter for over-dispersion in the population size process. In order to avoid over-parametrization, λ_i may be common to all sites, or it may be expressed as a function of site-specific covariates. In this paper, we consider the options introduced and considered in Ketwaroo (2019) for both the population size process and the detection process. Specifically, the Poisson and Negative binomial distributions for g as well as the lesser-known Discrete Weibull distribution (Nakagawa and Osaki, 1975).

The Discrete Weibull (DW) distribution developed by Nakagawa and Osaki (1975) is the discrete form of the continuous Weibull distribution that is popular in survival analysis and failure time studies (Peluso et al., 2019). In this paper,

The work in this chapter is a continuation of the same author's MSc project, but an extension of it in a number of ways. Specifically, the N-mixture models considered were introduced in the MSc thesis, and the analysis of the yellow-bellied toads data set was first presented in the MSc thesis. However, the introduction of the new proper objective prior within the context of N-mixture models, the comparison between this prior and approximations to Jefferys prior using simulation and real data, as well as the model selection discussion and results using conditional and marginal WAIC correspond to new work presented in this chapter.

we focus on the type 1 DW distribution, the most commonly used type in the literature. Let Y be a random variable that follows a (type 1) DW distribution, the cumulative distribution is defined as:

$$F(y; q, b) = \begin{cases} 1 - q^{(y+1)^b} & \text{for } y = 0, 1, 2, 3, \ldots, \\ 0 & \text{otherwise} \end{cases}$$

and the probability mass function is defined as:

$$f(y; q, b) = \begin{cases} q^{y^b} - q^{(y+1)^b} & \text{for } y = 0, 1, 2, 3, \ldots, \\ 0 & \text{otherwise} \end{cases}$$

where $0 < q < 1$ and $b > 0$.

Importantly, Kalktawi (2017) highlighted the flexibility of the DW distribution to model count data; relative to the Poisson distribution, the DW distribution can be used to model over-, under-, and equi-dispersed data. Regarding the parameters (q, b) of the DW distribution, Peluso et al. (2019) show that if:

1. $0 < b \leq 1$ there is over-dispersion, regardless of the value of q,
2. $b \geq 3$ there is under-dispersion, regardless of the value of q and
3. $1 < b < 3$, depending on the value of q there is under-dispersion or over-dispersion.

For the detection process, it is assumed

$$\mathrm{C}_{ij} \sim h(\mathrm{C}; \mathrm{N}_i, p_{ij}, \rho) \tag{1}$$

where h is a discrete probability distribution, p_{ij} represents the probability of detecting an individual at site i and sampling occasion j and ρ represents an optional parameter for over-dispersion in the detection process. We consider the Binomial and the Beta-Binomial (BB) distributions for h. The Binomial distribution is most commonly used to describe the detection process, assuming independence of detection. Using a Binomial detection process, p_{ij} can be assumed to be constant across all sites and sampling occasions, or in a logistic regression framework, it can be expressed as a function of site and sampling occasion specific covariates.

Martin et al. (2011) showed that the BB distribution can serve as a detection process for modelling the correlating behaviour of individuals, thus relaxing the assumption of independent detection of individuals by the Binomial distribution. The BB detection process accomplishes this by modelling

$$p_{ij} \sim \mathrm{Beta}(\alpha, \beta)$$

for $\alpha, \beta > 0$. Therefore, the BB detection process can also be used to model heterogeneity in detection probabilities (Martin et al., 2011; Ketwaroo, 2019). In addition, ρ represents the degree to which individual behaviours or site attributes correlate with each other, which could affect detection (Martin et al., 2011), and is defined as

$$\rho = \frac{1}{\alpha + \beta + 1}$$

Notably, the BB distribution does not allow the distinction between correlations in individual behaviour and attributes of the site that could affect detection.

Assuming N_i are independent random variables with discrete probability function $g(\mathrm{N}_i; \lambda_i, \gamma)$, and C_{ij} are conditionally dependent on N_i with discrete probability function $h(\mathrm{N}_i, p_{ij}, \rho)$, the marginal likelihood can be written as:

$$L(p_{ij}, \lambda_i, \rho, \gamma; \mathrm{C}_{ij}) = \prod_{i=1}^{M} \left\{ \sum_{\mathrm{N}_i = \max_j \mathrm{C}_{ij}}^{\infty} \left(\prod_{j=1}^{J} h(\mathrm{C}_{ij}; \mathrm{N}_i, p_{ij}, \rho) \right) g(\mathrm{N}_i; \lambda_i, \gamma) \right\}. \quad (2)$$

This marginal likelihood takes into account all values for population size at each site and in reality, an upper bound can be chosen when fitting N-mixture models using the marginal likelihood. In a Bayesian setting, N_i can be treated as a latent variable that can be sampled via MCMC methods, hence avoiding the need for the infinite sum or truncation. The full conditional likelihood for N-mixture models can then be written as:

$$L(p_{ij}, \rho; \mathrm{N}_i, \mathrm{C}_{ij}) = \prod_{i=1}^{M} \left(\prod_{j=1}^{J} h(\mathrm{C}_{ij}; \mathrm{N}_i, p_{ij}, \rho) \right). \quad (3)$$

where there is no longer the need to marginalise over N_i, as in equation (2), and now g from equation (2) serves as the prior distribution for N_i in this conditional model. Bayesian inference using the marginal likelihood has the appeal of being similar to the maximum likelihood approach and in some cases, faster than sampling latent variables using the conditional likelihood (Ponisio et al., 2020). However, for N-mixture models, Ponisio et al. (2020) showed that in a Bayesian setting, marginalization is generally less computationally efficient than sampling N_i. This is possibly due to the computational cost of summing over the range of possible values of N_i when the chosen upper bound is large.

Table 1: N-mixture models developed/implemented in Ketwaroo (2019) considered in this paper.

N-mixture model	Model for population size process	Model for detection process
P-B	Poisson (λ)	Binomial(N_i, p)
DW-B	Discrete Weibull (q, b)	Binomial (N_i, p)
NB-B	Negative Binomial (r, s)	Binomial (N_i, p)
P-BB	Poisson (λ)	Beta - Binomial (N_i, p_{ij}, ρ)
DW-BB	Discrete Weibull (q, b)	Beta - Binomial (N_i, p_{ij}, ρ)

Table 1 displays the list of N-mixture models investigated in this paper. We assume λ_i to be constant for all sites for all models and for models with a Binomal detection process, we assume p_{ij} to be constant across sites and sampling occasions. The P-B model is one of the most popular N-mixture models and it assumes equi-dispersion in the population size and detection processes. The NB-B model is also popular as it accounts for over-dispersion in the population size process relative to the Poisson distribution. The DW-B model offers more flexibility by accounting for over-, under-, and equi-dispersion in the population size process relative to the Poisson distribution. The P-BB model accounts for over-dispersion in the detection process,

and the DW-BB model has the advantage of accounting for over-dispersion in the detection process as well as under -, equi-, or under-dispersion in the population size process relative to the Poisson distribution.

2.1 Objective Prior distributions

Jeffreys Prior (Jeffreys, 1946) - An obvious candidate for an objective prior is to use a flat prior $p(\theta) \propto c$, $c > 0$ such that $\int p(\theta)d\theta = \infty$. This flat prior is an improper prior and not transformation invariant. Instead, Jeffreys (1946) derived prior distributions that are transformation invariant. The Jeffreys prior is the most popular objective prior and can be defined as:

$$p(\theta) \propto \sqrt{|I(\theta)|}$$

where $I(\theta) = -E\left[\frac{\partial^2 logp(x|\theta)}{\partial\theta\partial\theta^T}|\theta\right]$ is the Fisher information where $p(x|\theta)$ denotes the likelihood. For a Poisson distribution with mean λ, the Fisher information $I(\lambda) = \frac{1}{\lambda}$, and so the Jeffreys prior is the *improper prior*, $p(\lambda) \propto \frac{1}{\lambda^2}$. This prior can be approximated by a Gamma(ϵ_1, ϵ_2) where $\epsilon_1, \epsilon_2 \approx 0$ such as Gamma$(0.5, 0.00001)$ (Spiegelhalter et al., 2003). The Jeffreys prior yields sensible posterior distributions in scenarios where there is only one parameter of interest. However, it produces posteriors with poor performance when the parameter space has two or more dimensions (Leisen et al., 2018).

Walker and Villa (2021) recently developed a novel proper objective (OB) prior for continuous parameters by considering the connection between information, divergence and scoring rules. Let $\Theta = (0, \infty)$ be the parameter space of interest such that $\theta \in \Theta$. For some constant $a > 0$, the OB prior can be defined as

$$p(\theta) = \frac{a}{(a+\theta)^2}.$$

Setting $a = 1$ results in a heavy-tailed distribution as shown in Fig. 1. This distribution shape allows it to behave similarly to standard improper objective priors such as Jeffreys priors and reference priors (Berger et al., 2009), where a reference prior is an objective prior designed to maximize some measure of distance between the posterior and prior to allow the data to have maximum effect on the posterior. Measures such as the Kullback-Leibler divergence (Kullback and Leibler, 1951) or the Hellinger distance (Beran, 1977) can be used to construct reference priors. Reference priors and Jeffreys priors are only equivalent for one-dimensional parameters.

Walker and Villa (2021) showed that this novel objective prior performed almost equivalently to the Jeffreys prior on simulated data. Unlike improper objective prior distributions, this novel objective prior distribution is proper, guaranteeing a proper posterior distribution.

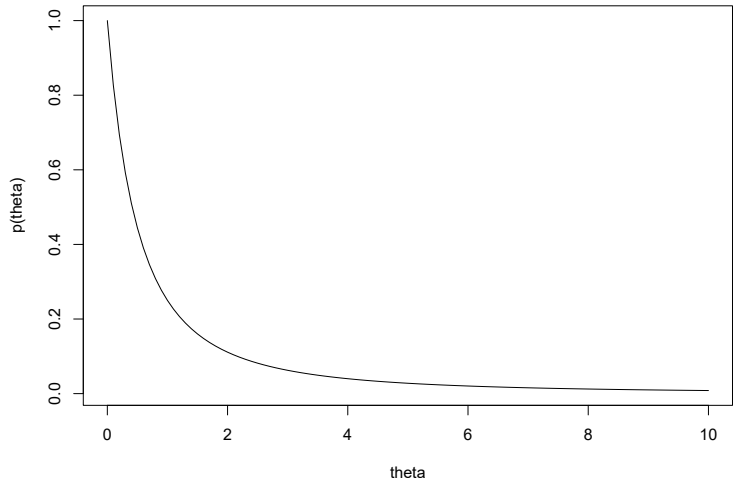


Figure 1: The OB prior $p(\theta) = 1/(1+\theta)^2$ for a parameter defined in $(0, \infty)$.

2.2 Model Selection via WAIC

WAIC, also called the "widely available information" criterion, is a fully Bayesian predictive accuracy measure estimator based on the log posterior predictive distribution (Watanabe, 2010). To mathematically define WAIC, let θ represent all model parameters, $y_1, \ldots, y_n$ denote the sample data, f be the true model, $\tilde{y}$ be the future data that could be observed, and $p_{post}(\tilde{y}) = \int p(\tilde{y}_i|\theta)p(\theta|y)d\theta$ be the posterior predictive distribution where $\tilde{y}_i$ denotes future data point i. Since the future $\tilde{y}_i$ is unknown, the expected log predicted density(elpd) can be used as a measure of predictive accuracy (Gelman et al., 2014):

$$\text{elpd} = E_f(\log p_{post}(\tilde{y}_i)) = \int \log p_{post}(\tilde{y}_i) f(\tilde{y}_i) d\tilde{y}_i$$

For the n new data points, elpd is computed for each data point to establish the predictive accuracy measure of that data set:

$$\text{Expected log pointwise predicted density (elppd)} = \sum_{i=1}^{n} E_f(\log p_{post}(\tilde{y}_i))$$

However, the log posterior predictive density is unknown as the likelihood $p(\tilde{y}_i|\theta)$ cannot be computed. For this reason, the prediction accuracy of a fitted model can be summarised using the log pointwise predictive density(lppd):

$$\text{lppd} = \log \prod_{i=1}^{n} p_{post}(y_i) = \sum_{i=1}^{n} \log \int p(y_i|\theta)p(\theta|y)d\theta$$

In practice, draws from the posterior distribution can be used to evaluate lppd. Let θ^s, for $s = 1, \ldots, S$ be the draws from the posterior distribution, then the computed lppd ($\widehat{\text{lppd}}$) can be defined as:

$$\widehat{\text{lppd}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right)$$

Accordingly, WAIC estimates the expected log pointwise predictive density elppd as the log pointwise predictive distribution lppd with a bias adjustment $\widehat{\text{elppd}}_{\text{WAIC}} = \widehat{\text{lppd}} - p_{\text{WAIC}}$. Two estimates of the bias adjustment have been proposed in the literature (Gelman et al., 2014). In this paper, we use the following bias adjustment:

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{post}(\log p(y_i|\theta)), \tag{4}$$

which can be computed by:

$$\text{computed } \ p_{\text{WAIC}} = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(y_i|\theta^s))$$

where $V_{s=1}^{S}$ represents the posterior sample variance. Thus, p_{WAIC} can be easily computed by summing the posterior variance of the log predictive density over all data points y_i. See Gelman et al. (2014) for more information on the other bias adjustment.
Hence, WAIC can be generally expressed as

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}). \tag{5}$$

Specifically, conditional WAIC (cWAIC) and marginal WAIC (mWAIC) can be expressed as

$$\text{cWAIC} = -2(\text{lppd}_c - p_{\text{cWAIC}}) \tag{6}$$

$$\text{mWAIC} = -2(\text{lppd}_m - p_{\text{mWAIC}}) \tag{7}$$

where $\text{lppd}_c, p_{\text{cWAIC}}$ are computed using the conditional likelihood (equation (3)) and $\text{lppd}_m, p_{\text{mWAIC}}$ are computed using the marginal likelihood (equation (2)). Both cWAIC and mWAIC can be computed by using MCMC samples from the fitted conditional model, and this is the approach employed in this work.

Notably, WAIC (equation (5)) is on the deviance scale, making it comparable with other measures of deviance such as the Akaike information criterion (AIC), and the Deviance information criterion (DIC). The model with the lowest WAIC is considered the best model considering all models. In addition, as opposed to conditioning on a single point as is done in AIC and DIC, WAIC has the advantage of averaging over the entire posterior distribution, making it more appropriate for Bayesian models and particularly useful for complex models with many parameters. The notable weakness of WAIC is that its calculation depends on the independence assumption of data given the parameters, making it unclear how to compute for structured data settings such as time series, spatial, and network data.

As WAIC is an information criterion, we assess the strength of evidence for each model using delta WAIC and Akaike weights. Assuming there are M candidate models, delta WAIC for the m^{th} candidate model (Δ_m) can be computed as $\Delta_m = \text{WAIC}_m - \text{WAIC}^*$ where WAIC^* is the minimum WAIC among the M candidate models.

Akaike weights, denoted by ω_m, can be computed as:

$$\omega_m = \frac{\exp(-0.5\Delta_{\text{m}})}{\sum_{i=1}^{M} \exp(-0.5\Delta_{\text{i}})}.$$

That is, ω_m, is the ratio of a candidate model's delta WAIC relative to the sum of the delta WAICs for all candidate models.

3 Simulation study

We consider two extensive simulation cases: 1) to investigate model performance when using the OB prior and priors that are approximations to the Jeffreys prior for λ in the P-B N-mixture model, and 2) to investigate whether WAIC is a reliable tool for model selection of N-mixture models and whether its performance depends on which likelihood calculation, conditional or marginal, is employed.

In both cases, we fit models using MCMC methods provided by R package NIMBLE (de Valpine et al., 2017) version 0.10.0 and use the full conditional N-mixture model (Equation (2)) as it was found to be more computationally efficient than the marginalized N-mixture model (Equation (3))(Ponisio et al., 2020). cWAIC and mWAIC are computed using MCMC samples from the fitted conditional model. To evaluate inference quality, we use the posterior median for each parameter since the conditional posterior distributions for λ and p were found to be skewed, and use $\hat{\lambda}$ and $\hat{p}$ to denote the median of the posterior medians over the simulation set. We also calculate 95% posterior credible interval coverage (Cov_θ), residual mean square error $\left(RSME_\theta = \frac{\sqrt{\sum_{i=1}^{\text{nsim}}(\hat{\theta}_i-\theta)^2/\text{nsim}}}{\theta}\right)$, and median relative bias $\left(B_\theta = \text{median}\left(\frac{\hat{\theta}-\theta}{\theta}\right)\right)$, where θ is the true parameter value, $\hat{\theta}$ is the posterior median and nsim is the number of simulation runs.

3.1 Case 1 - Comparison of prior distributions

For λ, we use the OB prior, and the following approximations to the Jeffreys priors: Gamma$(0.001, 0.001)$ and Gamma$(0.5, 0.00001)$, and for p we use a Uniform$(0, 1)$ prior. We set $M = 20$, $J = 5$, and perform 100 simulation runs for each scenario: $\lambda = (5, 100, 500)$ and $p = (0.1, 0.25, 0.6)$. For $\lambda = (5, 100), p = (0.1, 0.25)$, we run 515000 MCMC iterations with burn-in of 15000 and thinning of 10 for 1 chain. For $\lambda = (5, 100), p = 0.6$, we run 115000 MCMC iterations with burn-in of 15000 and thinning of 10 for 1 chain. We run 815000 MCMC iterations with burn-in of 105000 and thinning of 20 for 1 chain for $\lambda = 500, p = (0.1, 0.25, 0.6)$. Different MCMC settings were chosen so that the effective sample size was similar between the different simulation scenarios.

Table 2: Simulation results using the OB prior.

λ	p	$\hat{\lambda}$	Cov_λ	$RMSE_\lambda$	B_λ	$\hat{p}$	Cov_p	$RMSE_p$	B_p
5	0.1	3.215	92	0.399	−0.357	0.149	92	0.798	0.494
5	0.25	4.503	94	0.273	−0.099	0.284	97	0.284	0.135
5	0.6	4.934	96	0.154	−0.013	0.594	96	0.087	−0.009
100	0.1	57.000	89	0.441	−0.423	0.175	89	1.004	0.747
100	0.25	86.739	92	0.284	−0.133	0.291	91	0.369	0.167
100	0.6	100.326	96	0.115	0.003	0.605	97	0.096	0.008
500	0.1	315.925	90	0.415	−0.368	0.160	91	0.859	0.601
500	0.25	466.172	94	0.326	−0.067	0.268	93	0.322	0.073
500	0.6	498.572	97	0.108	−0.002	0.601	96	0.098	0.002

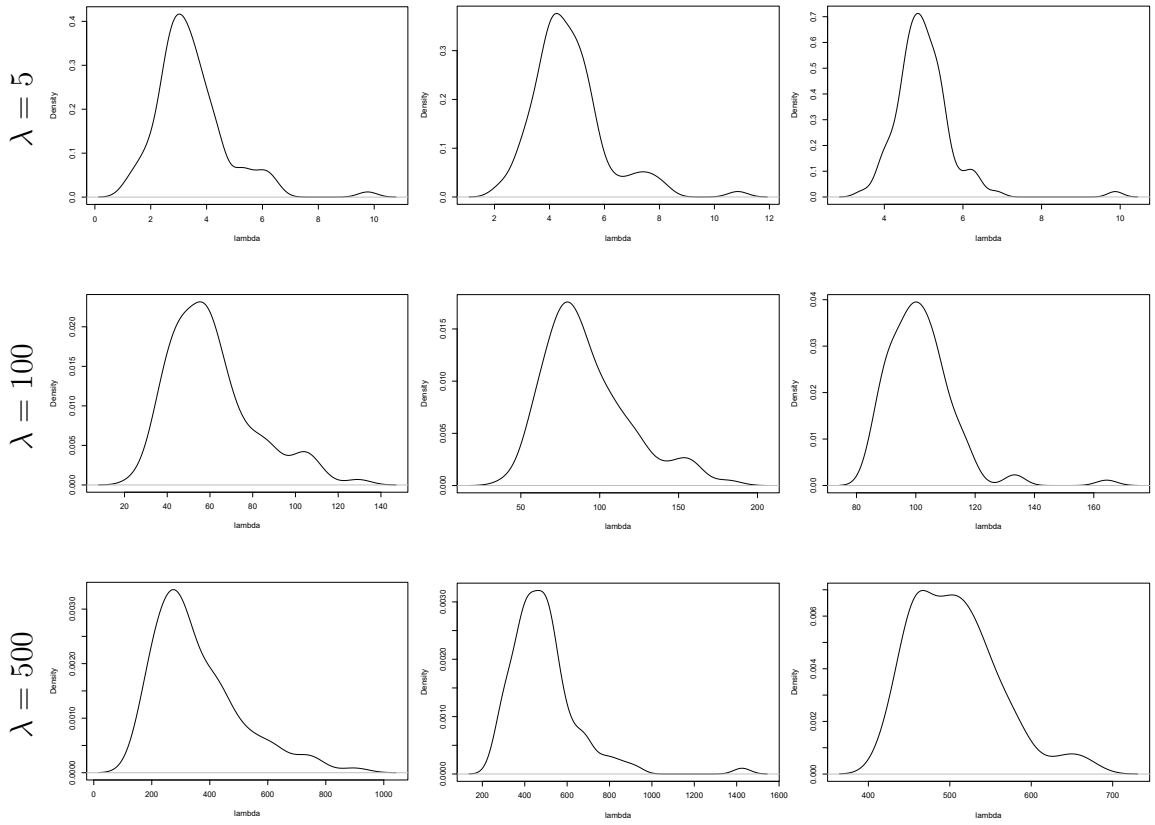
Table 3: Simulation results using the Gamma$(0.001, 0.001)$ prior.

λ	p	$\hat{\lambda}$	Cov_λ	$RMSE_\lambda$	B_λ	$\hat{p}$	Cov_p	$RMSE_p$	B_p
5	0.1	3.862	95	0.468	−0.227	0.124	96	0.706	0.237
5	0.25	4.857	95	0.407	−0.028	0.268	97	0.297	0.072
5	0.6	5.001	96	0.169	0.000	0.591	96	0.092	−0.016
100	0.1	69.914	99	0.407	−0.300	0.138	98	0.778	0.380
100	0.25	92.502	92	0.358	−0.075	0.276	91	0.349	0.103
100	0.6	101.664	96	0.122	0.016	0.598	96	0.101	−0.003
500	0.1	348.831	96	0.367	−0.302	0.143	95	0.727	0.426
500	0.25	472.884	94	0.293	−0.054	0.263	94	0.297	0.052
500	0.6	501.903	96	0.109	0.004	0.598	95	0.099	−0.003

Table 4: Simulation results using the Gamma$(0.5, 0.00001)$ prior.

λ	p	$\hat{\lambda}$	Cov_λ	$RMSE_\lambda$	B_λ	$\hat{p}$	Cov_p	$RMSE_p$	B_p
5	0.1	5.187	95	1.055	0.037	0.092	98	0.669	−0.084
5	0.25	5.175	96	0.689	0.035	0.249	95	0.334	−0.003
5	0.6	5.045	96	0.181	0.009	0.589	96	0.095	−0.017
100	0.1	89.843	99	0.702	−0.101	0.115	99	0.639	0.145
100	0.25	102.855	94	0.549	0.028	0.249	95	0.365	−0.002
100	0.6	102.726	96	0.152	0.027	0.592	96	0.111	−0.013
500	0.1	503.416	97	1.151	0.007	0.099	97	0.558	−0.007
500	0.25	567.462	92	0.972	0.135	0.221	90	0.366	−0.113
500	0.6	507.927	95	0.124	0.016	0.591	95	0.107	−0.015

From Tables 2, 3, and 4 it can be seen that the OB prior and Gamma priors perform similarly in terms of inference at high and low levels of detection probability. Notably, when p is small and priors for λ are concentrated at zero, as is the case for all prior distributions considered here, λ can be severely underestimated, as can also be seen in Fig. 2, which displays the density plots of the posterior medians of λ from the 100 runs for the OB prior. In addition, looking at Fig. 2, we can see that large estimates of λ are also obtained when p is low, evident in the tails/ bi-modal density of the distribution of posterior medians. This corroborates the results found by Dennis et al. (2015) in a classical setting, who found that the maximum likelihood estimates of population size can tend to infinity when detection probability is small. Additionally, looking at Fig.2, it can also be seen that there are cases when λ is estimated well. Hence, the results demonstrate that the distribution of posterior medians obtained for λ has two or maybe even three modes, for the first time demonstrating the substantial risk of underestimating λ when detection probability is small.



(a) $p = 0.1$ (b) $p = 0.25$ (c) $p = 0.6$

Figure 2: Density plots of the posterior medians of λ obtained using the OB prior.

3.2 Case 2 - Model Selection via WAIC

We consider four simulation scenarios:

- Scenario 1: Over-dispersion in the population size process; the true model is the Discrete Weibull Binomial (DW-B) N-mixture model.
- Scenario 2: Over-dispersion in the detection process; the true model is the Poisson Beta-Binomial (P-BB) N-mixture model.
- Scenario 3: Equi-dispersion in the population size process; the true model is the Poisson Binomial (P-B) N-mixture model.
- Scenario 4: Under-dispersion in the population size process; the true model is the DW-B N-mixture model.

In each scenario, 100 data sets were simulated from the true model, and the class of N-mixture models considered in this paper (Table 1) are fitted to each data set. Setting $M = 50,\ J = 5$, data generating model parameters for each case are shown in Table 5. For cases $1, 3$ and 4 data were simulated using $p = (0.25, 0.6)$ to investigate model selection when p is high and low. Similarly, for case 2, the Beta-Binomial parameters $\alpha = (3, 1)$ and $\beta = (2, 3)$ were chosen such that the mean detection probability is $0.25, 0.6$ respectively. For scenarios 1 and 4, data were generated

with an expected population size of 4.325 and 9.564 respectively. Parameters in the parameter space $(0, \infty)$ were assigned the OB prior, and parameters in the parameter space $(0, 1)$ were assigned a Uniform$(0, 1)$ prior. MCMC settings for each scenario are given in section 6.1 of the Supplementary material. In each scenario, we compute the cWAIC and mWAIC for each N-mixture model, report the proportion of times each model was selected %WAIC, median Δ WAIC and median WAIC weights (ω_{WAIC}) for both cWAIC and mWAIC. We use expected population size (λ) and p to compare inference quality between models. For simplicity, we let p represent the mean detection probability for BB models. We define 'Best by mWAIC' and 'Best by cWAIC' to be inferences of models selected by mWAIC and cWAIC respectively.

Table 5: Data generating model parameters for each model.

Scenario	Model	Parameters
1	DW-B	$q = 0.75, b = 0.95$
2	P-BB	$\lambda = 5$
3	P-B	$\lambda = 5$
4	DW-B	$q = 0.9999, b = 4$

3.2.1 Scenario 1- Over-dispersion in the population size process

As can be seen from Tables 6 and 7, when there was over-dispersion in the population size process, cWAIC strongly favoured the more complicated model, the P-BB model, which gave poor inference, instead of the true model. On the other hand, mWAIC selected the correct model with higher probability and better inference. The ability of mWAIC to select the true model was reduced with low p, but it selected a similar model that accommodates overdispersion in the population size process and produced good inference. Model inference results (Table 7) also agree with the findings of Knape et al. (2018), that is, models that do not accommodate overdispersion in the population size process, when overdispersion is present, underestimate expected population size.

Table 6: Scenario 1 model selection results when the true model is the DW-B N-mixture model.

p	Model	%cWAIC	%mWAIC	ΔcWAIC	Δ mWAIC	ω_{cWAIC}	ω_{mWAIC}
0.6	P-B	0	0	25.133	203.843	0	0
	DW-B	0	87	31.056	0	0	0.549
	NB-B	0	8	30.497	0.419	0	0.419
	P-BB	98	0	0	203.388	1	0
	DW-BB	2	5	162.001	135.517	0	0
0.25	P-B	0	0	35.974	26.222	0	0
	DW-B	0	29	41.345	0.205	0	0.421
	NB-B	0	64	40.645	0	0	0.453
	P-BB	100	1	0	16.553	0.999	0
	DW-BB	0	6	16.938	2.909	0	0.102

Table 7: Scenario 1 model inference results when the true model is the DW-B N-mixture model.

p	Model	$\hat{\lambda}$	Cov_λ	$RMSE_\lambda$	B_λ	$\hat{p}$	Cov_p	$RMSE_p$	B_p
0.6	P-B	3.005	7	0.334	−0.305	0.668	36	0.126	0.114
	DW-B	4.450	93	0.144	0.029	0.600	95	0.071	0.0003
	NB-B	3.377	68	0.276	−0.220	0.603	95	0.070	0.006
	P-BB	2.907	5	0.355	−0.328	0.683	18	0.145	0.138
	DW-BB	4.963	62	0.217	0.147	0.633	84	0.075	0.055
	Best by mWAIC	4.259	93	0.146	−0.015	0.603	95	0.070	0.005
	Best by cWAIC	2.908	5	0.355	−0.327	0.681	19	0.144	0.136
0.25	P-B	2.005	1	0.537	−0.537	0.416	8	0.679	0.663
	DW-B	4.356	90	0.312	0.007	0.254	96	0.278	0.016
	NB-B	3.159	81	0.353	−0.269	0.267	95	0.275	0.067
	P-BB	1.828	0	0.579	−0.577	0.440	2	0.777	0.762
	DW-BB	3.335	76	0.256	−0.228	0.358	66	0.490	0.433
	Best by mWAIC	3.467	78	0.424	−0.198	0.254	96	0.289	0.055
	Best by cWAIC	1.828	0	0.579	−0.577	0.440	2	0.777	0.762

3.2.2 Scenario 2 - Over-dispersion in the detection process

Looking at Tables 8 and 9, it can be seen that the cWAIC again strongly favoured the more complicated model, ie the DW-BB model, whilst mWAIC selected the correct model at least 3 times more and produced better inference than models selected by cWAIC. In addition, models that did not accommodate over-dispersion in the detection process over-estimated expected population size, agreeing with Knape et al. (2018).

Table 8: Scenario 2 model selection results when the true model is the P-BB N-mixture model.

p	Model	%cWAIC	%mWAIC	ΔcWAIC	Δ mWAIC	ω_{cWAIC}	ω_{mWAIC}
0.6	P-B	0	1	134.983	18.899	0	0
	DW-B	0	0	132.142	12.376	0	0.002
	NB-B	0	3	135.862	11.597	0	0.002
	P-BB	26	83	2.742	0	0.202	0.643
	DW-BB	74	13	0	1.600	0.798	0.291
0.25	P-B	0	0	154.726	38.776	0	0
	DW-B	0	0	141.092	25.334	0	0
	NB-B	0	0	139.177	20.955	0	0
	P-BB	23	61	8.565	0	0.014	0.644
	DW-BB	77	39	0	1.185	0.986	0.357

Table 9: Scenario 2 model inference results when the true model is the P-BB N-mixture model.

p	Model	$\hat{\lambda}$	Cov_λ	RMSE_λ	B_λ	$\hat{p}$	Cov_p	RMSE_p	B_p
0.6	P-B	7.644	6	0.655	0.528	0.388	1	0.367	−0.353
	DW-B	9.603	0	1.044	0.921	0.337	0	0.436	−0.437
	NB-B	8.687	2	0.930	0.737	0.342	0	0.440	−0.429
	P-BB	4.953	96	0.086	−0.009	0.591	96	0.053	-0.015
	DW-BB	5.852	25	0.208	0.170	0.593	96	0.051	−0.012
	Best by mWAIC	5.182	84	0.129	0.036	0.591	92	0.077	−0.015
	Best by cWAIC	5.736	41	0.188	0.147	0.593	96	0.050	−0.0108
0.25	P-B	11.280	1	1.798	1.608	0.111	7	0.568	−0.556
	DW-B	26.937	0	5.226	4.387	0.046	0	0.812	−0.814
	NB-B	38.048	0	7.619	6.609	0.031	0	0.864	−0.874
	P-BB	4.774	96	0.283	−0.045	0.263	96	0.228	0.051
	DW-BB	5.563	86	0.220	0.113	0.273	96	0.206	0.093
	Best by mWAIC	4.251	96	0.232	−0.149	0.267	97	0.198	0.071
	Best by cWAIC	4.219	94	0.244	−0.156	0.277	96	0.213	0.109

3.2.3 Scenario 3: Equi-dispersion in the population size process

Tables 10 and 11 show cWAIC strongly favouring the more complicated model, the DW-BB model which, compared to the true model, gave poorer inference while mWAIC selected the true model more often, in favour of models that fit the data best. In addition, models selected by mWAIC produced better inference than models selected by cWAIC.

Table 10: Scenario 3 model selection results when the true model is the P-B N-mixture model.

p	Model	%cWAIC	%mWAIC	ΔcWAIC	Δ mWAIC	ω_{cWAIC}	ω_{mWAIC}
0.6	P-B	0	63	28.854	0	0	0.360
	DW-B	0	14	25.960	1.514	0	0.179
	NB-B	0	17	29.521	0.360	0	0.320
	P-BB	12	4	2.715	3.722	0.205	0.060
	DW-BB	88	2	0	4.559	0.795	0.041
0.25	P-B	0	70	35.211	0	0	0.463
	DW-B	0	9	31.197	2.201	0	0.171
	NB-B	0	6	39.293	2.513	0	0.144
	P-BB	2	5	13.148	4.392	0.001	0.059
	DW-BB	98	10	0	4.034	0.998	0.006

Table 11: Scenario 3 model inference results when the true model is the P-B N-mixture model.

p	Model	$\hat{\lambda}$	Cov_λ	RMSE_λ	B_λ	$\hat{p}$	Cov_p	RMSE_p	B_p
0.6	P-B	5.036	94	0.095	0.007	0.593	93	0.062	−0.012
	DW-B	5.944	33	0.212	0.188	0.599	93	0.059	0.000
	NB-B	5.050	94	0.097	0.010	0.590	92	0.064	−0.016
	P-BB	4.635	90	0.098	−0.073	0.640	82	0.082	0.067
	DW-BB	5.525	61	0.133	0.104	0.647	72	0.090	0.080
	Best by mWAIC	5.168	84	0.119	0.034	0.598	91	0.065	−0.002
	Best by cWAIC	5.422	65	0.126	0.084	0.645	70	0.090	0.080
0.25	P-B	4.786	96	0.197	−0.043	0.259	94	0.214	0.036
	DW-B	5.317	94	0.195	0.063	0.282	92	0.260	0.128
	NB-B	6.056	96	0.476	0.211	0.209	94	0.295	−0.165
	P-BB	4.774	96	0.283	−0.045	0.262	96	0.228	0.051
	DW-BB	4.145	74	0.189	−0.171	0.273	96	0.206	0.093
	Best by mWAIC	4.845	92	0.177	−0.031	0.260	93	0.212	0.039
	Best by cWAIC	4.185	75	0.193	−0.165	0.263	95	0.202	0.090

3.2.4 Scenario 4: Under-dispersion in the population size process

As shown in Tables 12 and 13, cWAIC once again favoured the more complicated model instead of the true whereas mWAIC had a stronger preference for the true model and a better preference for models with good inference than models selected by cWAIC.

Table 12: Scenario 4 model selection results when the true model is the DW-B N-mixture model.

p	Model	%cWAIC	%mWAIC	ΔcWAIC	Δ mWAIC	ω_{cWAIC}	ω_{mWAIC}
0.6	P-B	0	79	36.733	0	0	0.581
	DW-B	0	16	33.225	2.103	0	0.216
	NB-B	0	0	38.395	4.846	0	0.054
	P-BB	6	2	8.202	4.560	0.016	0.062
	DW-BB	94	3	0	5.111	0.984	0.045
0.25	P-B	0	84	42.759	0	0	0.618
	DW-B	0	5	41.055	3.900	0	0.083
	NB-B	0	5	45.617	2.958	0	0.155
	P-BB	0	1	25.645	8.039	0	0.011
	DW-BB	100	5	0	5.174	1	0.047

Table 13: Scenario 4 model inference results when the true model is the DW-B N-mixture model.

p	Model	$\hat{\lambda}$	Cov_λ	RMSE_λ	B_λ	$\hat{p}$	Cov_p	RMSE_p	B_p
0.6	P-B	8.807	77	0.112	−0.079	0.584	96	0.064	−0.025
	DW-B	9.409	94	0.068	−0.016	0.600	96	0.058	0.001
	NB-B	8.979	86	0.102	−0.061	0.578	92	0.073	−0.034
	P-BB	7.795	31	0.185	−0.185	0.657	77	0.103	0.095
	DW-BB	8.450	51	0.118	−0.116	0.673	57	0.125	0.122
	Best by mWAIC	8.818	78	0.111	−0.078	0.594	92	0.074	−0.009
	Best by cWAIC	8.369	47	0.124	−0.125	0.710	6	0.182	0.184
0.25	P-B	9.084	94	0.271	−0.050	0.238	93	0.238	−0.047
	DW-B	5.562	54	0.415	−0.418	0.305	86	0.311	0.219
	NB-B	12.116	91	0.635	0.266	0.179	85	0.356	−0.285
	P-BB	6.149	62	0.363	−0.357	0.349	75	0.477	0.397
	DW-BB	5.747	7	0.401	−0.399	0.445	7	0.814	0.781
	Best by mWAIC	8.630	86	0.324	−0.097	0.239	84	0.373	−0.040
	Best by cWAIC	5.747	7	0.401	−0.399	0.445	7	0.814	0.781

From this extensive simulation study, it can be seen that mWAIC selected the correct model with a high probabiliy while cWAIC favoured the more complicated model that often gave poor inference. Hence, model selection via WAIC for N-mixture models should be performed using the marginal likelihood as cWAIC can favour unnecessarily complicated models. Importantly, these scenarios demonstrate that one can select between different N-mixture models with different model inferences using mWAIC.

4 Case studies

We consider two case studies: yellow-bellied toads and Swiss great tits. We apply all N-mixture models defined in Table 1 to both data, assuming the expected population size to be constant across sites and detection probability for Binomial models to be constant across sites and sampling occasions.

We fit models using the conditional likelihood (Equation (3)) and using MCMC samples from the fitted model, we perform model selection using both cWAIC and mWAIC. We choose the OB prior for continuous parameters with parameter space $(0, \infty)$, whereas parameters with parameter space $(0, 1)$ are assigned a Uniform$(0, 1)$ prior. Additionally, for the yellow-bellied toad, we investigate the prior sensitivity of parameters with parameter space $(0, \infty)$ by using an approximation to the Jeffreys prior, Gamma$(0.001, 0.001)$. To assess model fit, we use posterior predictive goodness of fit: we define $\tau_i = \sum_{j=1}^{J} \text{C}_{i,j}$ and using MCMC samples, we simulate counts, and hence τ_i, from our model and compare these to the observed data. A model fits the data well if it produces similar τ_i values to the observed data. MCMC settings used for both case studies are presented in section 6.2 of the Supplementary material. We assess convergence using Gelman and Rubin's convergence diagnostic (Gelman et al., 1992).

4.1 Yellow-bellied Toads

In 2018, survey sampling of five populations of yellow-bellied toads (*Bombina variegata*) was conducted at 27 sites from the end of May to the beginning of July. Each site was sampled 4 times during the period of study. Sites were represented by ponds or tanks located in a variety of habitats, mainly vineyards, and meadows, in the Italian Alps.

With convergence achieved for all model parameters, it can be seen from Table 14 that all models considered produce different estimates of expected population size and detection probability, highlighting the need to select the correct model to avoid erroneous inference. Notably, both cWAIC and mWAIC were in agreement strongly favoring the DW-BB model with cWAIC and mWAIC weights of 1.0 and the least support to the P-B model with cWAIC and mWAIC weights of 0. Additionally, the OB prior and the Jeffreys prior approximation (Gamma$(0.001, 0.001)$) give similar model inference and WAIC values. Posterior predictive goodness of fit indicated all models except the P-BB fitted the data well. Fig. 3 displays the GOF plot for the DW-BB model and it can be seen that the true value is captured between the 5th and 95th quantile for all sites for the DW-BB model. The P-BB model lack of fit is evident in the large estimates of expected population size.

Table 14: Model results from analysing yellow-bellied toads data. Values within the brackets represent the 95% posterior credible interval. For BB models, detection probability represents mean detection probability.

Prior	Model	cWAIC	mWAIC	Detection Probability	Expected population size	p correlation.
OB						
	P-B	542.328	905.735	$0.637(0.585, 0.686)$	$10.759(9.375, 12.218)$	-
	DW-B	461.211	528.078	$0.186(0.051, 0.360)$	$38.825(16.730, 145.290)$	-
	NB-B	461.535	529.078	$0.201(0.054, 0.371)$	$34.574(15.847, 130.823)$	-
	P-BB	491.621	654.837	$0.038(0.010, 0.085)$	$238.333(83.459, 653.727)$	$0.049(0.013, 0.109)$
	DW-BB	338.121	502.857	$0.662(0.555, 0.743)$	$11.111(7.360, 18.410)$	$0.182(0.085, 0.289)$
Gamma(0.001,0.001)						
	P-B	544.649	907.358	$0.635(0.579, 0.688)$	$10.818(9.401, 12.341)$	-
	DW-B	461.215	528.126	$0.193(0.054, 0.365)$	$37.553(16.39, 137.21)$	-
	NB-B	461.377	529.158	$0.197(0.053, 0.366)$	$45.425(15.977, 132.666)$	-
	P-BB	494.398	654.739	$0.018(0.003, 0.057)$	$672.845(122.529, 1951.403)$	$0.023(0.004, 0.073)$
	DW-BB	344.243	507.252	$0.649(0.511, 0.738)$	$11.328(7.460, 19.100)$	$0.163(0.061, 0.280)$

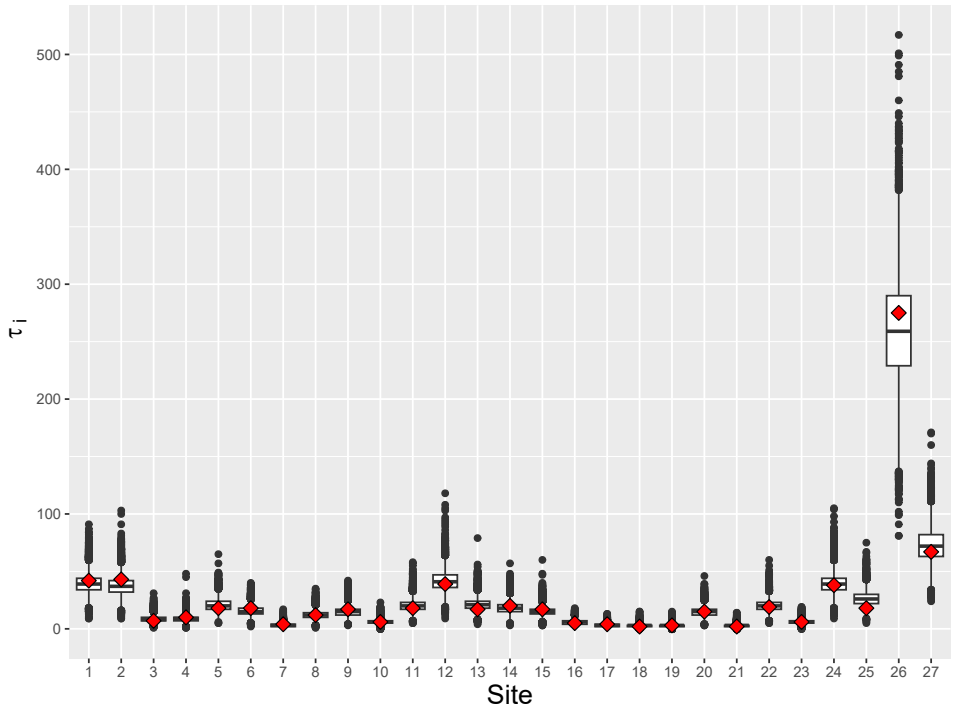


Figure 3: GOF plot for the DWBB model. Red diamonds present the observed values and boxplots represent the simulated values.

4.2 Swiss great tits

The Swiss great tits data were collected in the Swiss breeding bird survey MHB from 2013. The Swiss common bird breeding survey MHB is based on a sample of 267 1-km^2 areas. Volunteers survey a quadrant-specific route, composed of 263 sites, three times during the breeding season.

The Swiss great tits data was analysed by Kéry and Royle (2017) where they highlighted the good-fit-bad-prediction dilemma. Using covariates on both expected population size and detection probability, they analysed this data set using three models: P-B, ZIP-B and NB-B where they found that the best-fitted model (NB-B) via AIC produced unrealistic estimates of population size. To come to this conclusion, they performed residual diagnostic checks. They found that the residual diagnostic checks for the P-B and ZIP-B models looked much better than those of the NB-B model, despite the much better fit (GOF test) and predictive ability (measured by AIC) of the NB-B model. Thus, we investigate this good-fit-bad-prediction dilemma in a Bayesian framework using methods considered in this paper.

Convergence was achieved for all model parameters. From Table 16, we also find that the NB-B model was favored by both cWAIC and mWAIC over the P-B model, and the NB-B also produced large values of expected population size. To evaluate model fit, we simulate data from the model and compute the 95% coverage of $\tau_{1:M}$. For the P-B model, 33.34% of the sites captured the observed values while for the NB-B model, 72.09% of the sites captured the observed values. Contrary to Royle (2015), we consider BB models, and as can be seen from Table 16, the DW-BB model was strongly supported as the best model amongst all models by both cWAIC and mWAIC with weights of 1.0. The DW-BB model produced inference similar to the P-B model but fitted the data well with 72.09% of the sites capturing the true

value. This motivates the use of the DW-BB model as the good-fit-bad prediction dilemma observed by Royle (2015) may be due to the violation of the independence detection assumption in the Binomial detection process. Hence, our findings in both case studies suggest that model selection and model fit are in agreement with model selection favouring the model with the better fit.

Table 15: Model results from analysing Swiss great tits data. Values within the brackets represent the 95% credible interval. For BB models, detection probability represents mean detection probability.

Model	cWAIC	mWAIC	Detection Probability	Expected population size	p correlation
P-B	3689.985	6020.880	$0.641(0.621, 0.661)$	$10.142(9.679, 10.616)$	-
DW-B	2968.517	3600.062	$0.063(0.016, 0.132)$	$123.412(53.72, 460.90)$	-
NB-B	2954.897	3579.046	$0.045(0.011, 0.103)$	$202.140(61.714, 624.950)$	-
P-BB	2663.500	4311.279	$0.298(0.262, 0.335)$	$20.131(18.184, 22.558)$	$0.476(0.425, 0.526)$
DW-BB	2606.225	3564.125	$0.416(0.249, 0.560)$	$17.799(12.550, 30.460)$	$0.054(0.015, 0.116)$

Table 16: Model results from analysing Swiss tits data. Values within the brackets represent the 95% credible interval.

Model	cWAIC	mWAIC	Detection Probability	Expected population size	p correlation
P-B	3689.985	6020.880	$0.641(0.621, 0.661)$	$10.142(9.679, 10.616)$	-
DW-B	2968.517	3600.062	$0.063(0.016, 0.132)$	$123.412(53.72, 460.90)$	-
NB-B	2954.897	3579.046	$0.045(0.011, 0.103)$	$202.140(61.714, 624.950)$	-
P-BB	2663.500	4311.279	$0.298(0.262, 0.335)$	$20.131(18.184, 22.558)$	$0.476(0.425, 0.526)$
DW-BB	2606.225	3564.125	$0.416(0.249, 0.560)$	$17.799(12.550, 30.460)$	$0.054(0.015, 0.116)$

5 Discussion

As N-mixture models provide an attractive framework to gain inference on population size by using only replicated counts from unmarked individuals. A large number of studies have been carried out on N-mixture models in a classical setting, resulting in the identification of issues such as computational aspects of model fitting, model selection, sensitivity to overdispersion, etc. However, to our knowledge, few studies have been conducted in a Bayesian setting to investigate N-mixture models. N-mixture models have also become easier to fit in a Bayesian framework with the advent of software such as NIMBLE (de Valpine et al., 2017) and Stan (Carpenter et al., 2017). Hence, in this paper, we considered fitting an extensive class of N-mixture models in a Bayesian framework to corroborate and extend issues concerning N-mixture models obtained in a classical framework.

Moreover, we have performed extensive simulation studies to investigate the choice of prior distributions and model selection in N-mixture models. We implemented a novel proper objective prior, the OB prior, and compared its performance to approximations of the popular Jeffreys priors. We found these priors performed similarly in terms of inference. Importantly, when p is small, we found that λ can be considerably underestimated in addition to well-known cases of λ being overestimated, a finding we believe to be previously unknown. We further investigated model selection via WAIC, considering both the conditional and marginal WAIC criteria, cWAIC and mWAIC, respectively. We found that cWAIC can lead to misleading results that favour the more complicated model while mWAIC selected

the true model with a high probability. Hence, mWAIC should be used instead of cWAIC to select between competing N-mixture models.

Finally, we considered these methods in two case studies. We found the OB prior and a Jeffreys prior approximation produced similar inference results and model selection results as cWAIC and mWAIC were in agreement for the case study considered. In addition, contrary to the good-fit-bad-prediction highlighted by Kéry and Royle (2017), we find model selection via WAIC to be in agreement with the model goodness of fit when the DW-BB model is considered in the model list. Future work can be focused on developing Bayesian goodness-of-fit measures to check model assumptions of N-mixture models.

Notably, Vehtari et al. (2017) highlighted checks that can be done to investigate the stability of WAIC. Ariyo et al. (2022) highlighted WAIC sensitivity to the choice of prior in Bayesian linear mixed models for longitudinal data. Thus, a possible avenue for future work can be to investigate the stability of cWAIC and mWAIC and WAIC sensitivity to the choice of prior in N-mixture models.

Another important avenue for future work is the identifiability of N-mixture models in a Bayesian framework as identifiability issues have been found in a classical setting (Dennis et al., 2015; Barker et al., 2018). Thus, future work can be focused on investigating identifiability of N-mixture models in a Bayesian setting using methods such as data cloning (Lele et al., 2007).

The identifiability of N-mixture models in a Bayesian framework is another important avenue for future work. Non-identifiability is the scenario where models can be fitted to data without all model parameters being estimable. Identifiability issues have been found with N-mixture models in a classical setting. Dennis et al. (2015) showed that when the probability of detection and the number of sampling occasions are small, infinite estimates of population size can be obtained. Barker et al. (2018) highlighted that compared to capture-recapture surveys, the loss of individual information resulting from count surveys is critical and causes problems in estimated parameters in Binomial N-mixture models. Kéry (2018) responded to some of these problems of parameter identifiability in a classical framework and called for more research to be done on the parameter identifiability of N-mixture models.

Thus, we investigated parameter identifiability of the set of N-mixture models considered in this thesis using data cloning (DC) (Lele et al., 2007). DC is a statistical computing method introduced by Lele et al. (2007). Cloning the data K times, DC takes advantage of the computational simplicity of the MCMC algorithms that are used in a Bayesian framework to provide maximum likelihood point estimates and their standard errors for complex hierarchical models. Importantly, Lele et al. (2010) proved that for estimable parameters in the model, the scale posterior variance should be approximately $1/K$. If parameters do not follow this trend then parameters are non-identifiable. This is primarily a method of detecting extrinsic parameter identifiability, that is, this method is used to detect parameter identifiability for a specific data set.

Consequently, an important component in using DC to investigate parameter identifiability is the choice of K. Ponciano et al. (2012) showed that if parameters are weakly estimable, a large number of clones is needed as the parameters mean and variance may increase at the beginning but as the number of clones increases, the variance will converge to zero. Parameters that are weakly estimable produce likelihoods that are relatively flat resulting in parameter estimation with large vari-

ance.

To determine whether DC can be used to assess parameter identifiability in the P-B N-mixture model we compare DC to the covariance diagnostic proposed by Dennis et al. (2015). We simulate data with $p = 0.1,\ \lambda = 5,\ M = 20,\ J = 3$ and select data sets such that the P-B model is identifiable for 10 data sets ("identifiable cases") and non-identifiable for 10 data sets ("non-identifiable case") according to the covariance diagnostic. At the same time, we also investigate the prior effects on the performance of DC. Three types of priors were investigated: the OB prior, an approximation to Jeffreys prior (Gamma(0.5,0.00001)), and an informative prior (Gamma(5,1)). For p, a Uniform(0,1) prior was assigned. We focus on the identifiability of λ and set $K = 10$. For $K = 1$, we run 505000 MCMC iterations with a burn-in of 40000 and thinning of 5 for 2 chains. For $K \geq 1$, we run 705000 MCMC iterations with a burn-in of 50000 and thinning of 5 for 2 chains.

For the "identifiable cases", DC indicated the identifiability of λ in all data sets. Fig. 4 displays 4 such DC plots indicating parameter identifiability. For the "non-identifiable cases", DC indicated the non-identifiability of λ in all data sets. Fig. 5 displays DC plots for 4 data sets indicating non-identifiability. In this case, λ was severely overestimated giving unrealistic estimates of population size. Additionally, from these Figs., it can be seen that DC results are similar for the different types of prior considered for both "identifiable cases" and "non-identifiable cases", indicating DC is not sensitive to prior specification in this scenario.

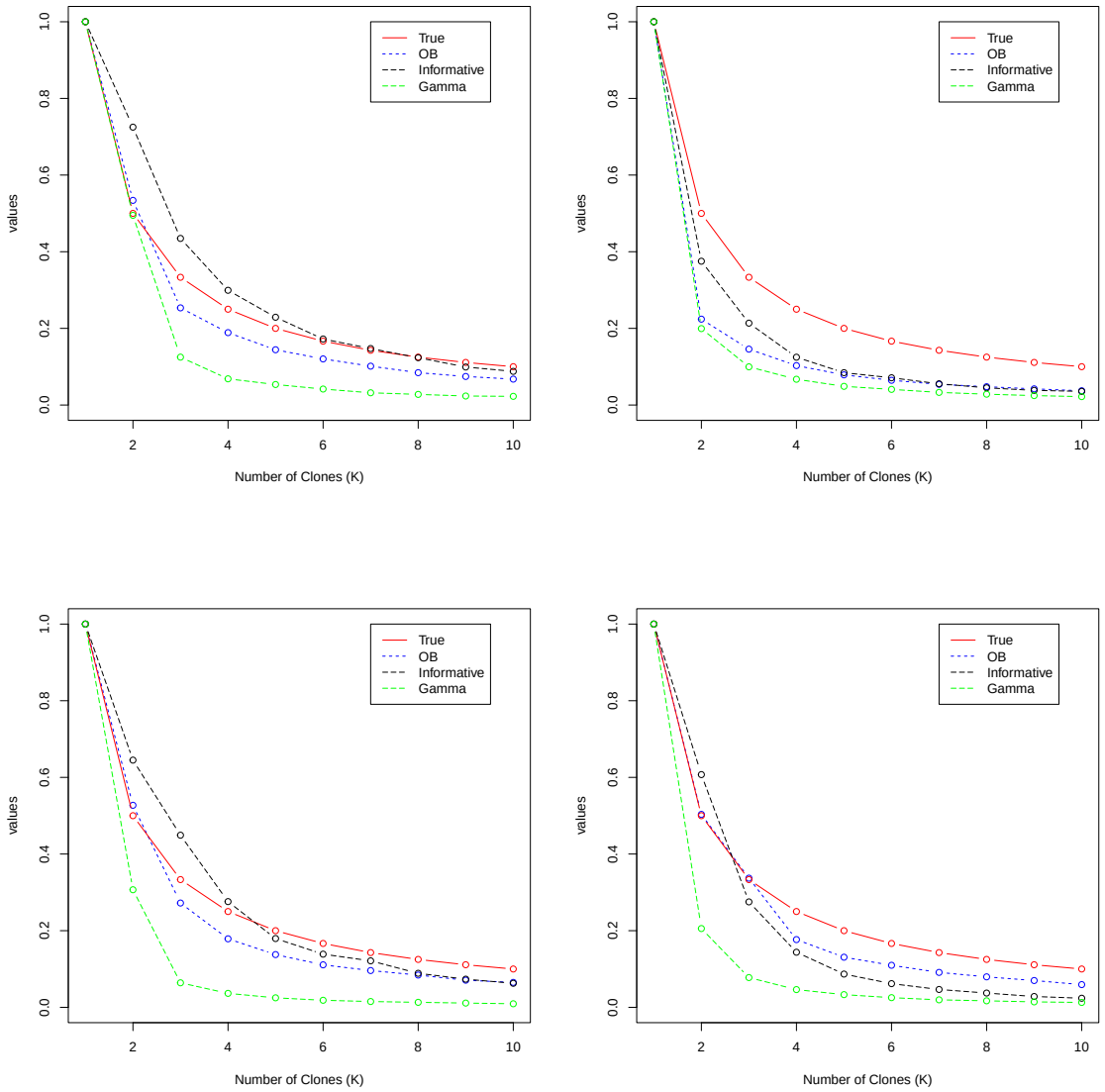


Figure 4: Data cloning identifiability diagnostic plots for 4 “identifiable cases”.

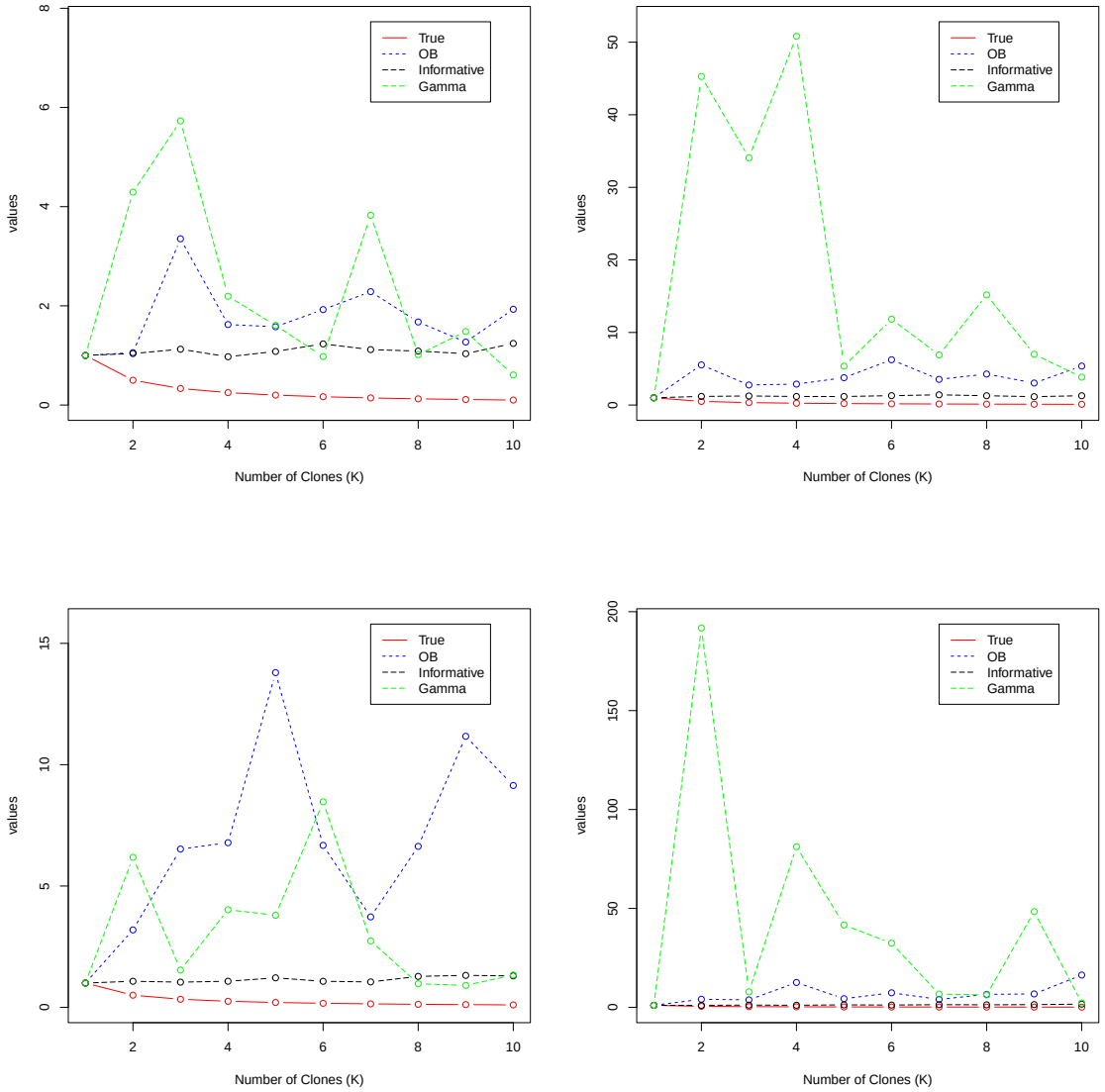


Figure 5: Data cloning identifiability diagnostic plots for 4 " non-identifiable cases".

We further investigated the identifiability of over-dispersion N-mixture models using DC. We perform 10 simulation runs for each N-mixture model: DW-B, NB-B, P-BB and DW-BB for $p = 0.1$, $\lambda = 20$, $M = 20, J = 3, K = 20$. The OB prior was assigned to parameters in the parameter space $(0, \infty)$, and a Uniform$(0, 1)$ prior was assigned to parameters in the parameter space $(0, 1)$.

For the NB-B model, 8/10 datasets DC indicated non-identifiability issues for the size parameter of the NB distribution. These estimates of the size parameter were unrealistic large estimates but estimates of expected population size and p were realistic indicating identifiability. For the DW-B model, 10/10 datasets DC indicated parameter identifiability with realistic inference. For the P-BB model, 8/10 datasets DC indicated the identifiability of all parameters. Two datasets indicated the non-identifiability of λ and β, where these were over-estimated and the mean detection probability and ρ were underestimated suggesting non-identifiability. For the DW-BB model, 6/10 datasets DC indicated the identifiability of all parameters.

In the other 4 datasets, there were identifiability issues for β as it was severely underestimated. However, there were no obvious signs of non-identifiability as inference on mean detection probability and expected population size was not unrealistic.

All in all, these results show that DC can be a valuable tool for investigating the identifiability of the P-B N-mixture model in a Bayesian setting. However, for over-dispersion N-mixture models, parameter identifiability via DC was not straightforward as in this case DC can indicate that either one or both parameters of the distribution for N are non-identifiable, but inference on N itself is reliable, suggesting that perhaps there exist several combinations of values or ranges of values for these parameters that yield similar inference for N. Dennis et al. (2015) also proposed two diagnostics to identify identifiability issues in the NB-B N-mixture model but these were found to be unreliable when used singly or in combination. Hence, future work is needed to investigate parameter identifiability in N-mixture models.

Overall, N-mixture models are a powerful tool for estimating population size. However, like any tool, care must be taken. This work highlights that in a Bayesian framework, care needs to be taken with the choice of prior distributions and advocates the use of mWAIC to select between models.

References

Ariyo, O., Lesaffre, E., Verbeke, G., and Quintero, A. (2022). Model selection for bayesian linear mixed models with longitudinal data: sensitivity to the choice of priors. *Communications in statistics-simulation and computation*, 51(4):1591–1615.

Ariyo, O., Quintero, A., Muñoz, J., Verbeke, G., and Lesaffre, E. (2020). Bayesian model selection in linear mixed models for longitudinal data. *Journal of Applied Statistics*, 47(5):890–913.

Banner, K. M., Irvine, K. M., and Rodhouse, T. J. (2020). The use of bayesian priors in ecology: The good, the bad and the not great. *Methods in Ecology and Evolution*, 11(8):882–889.

Barker, R. J., Schofield, M. R., Link, W. A., and Sauer, J. R. (2018). On the reliability of n-mixture models for count data. *Biometrics*, 74(1):369–377.

Beran, R. (1977). Minimum hellinger distance estimates for parametric models. *The annals of Statistics*, pages 445–463.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

Dennis, E. B., Morgan, B. J. T., and Ridout, M. S. (2015). Computational aspects of n-mixture models. *Biometrics*, 71(1):237–246.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016.

Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.

Hunter, E., Nibbelink, N., and Cooper, R. (2017). Divergent forecasts for two salt marsh specialists in response to sea level rise. *Animal Conservation*, 20(1):20–28.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Kalktawi, H. S. (2017). *Discrete Weibull regression model for count data*. PhD thesis, Brunel University London.

Kéry, M. (2018). Identifiability in n-mixture models: A large-scale screening test with bird data. *Ecology*, 99(2):281–288.

Kéry, M. and Royle, J. A. (2017). Applied hierarchical modeling in ecology: analysis of distribution, abundance and species richness in r and bugs. *(No Title)*.

Ketwaroo, F. R. (2019). N-mixture models from a bayesian perspective. *Unpublished MSc. Thesis*.

Knape, J., Arlt, D., Barraquand, F., Berg, Å., Chevalier, M., Pärt, T., Ruete, A., and Żmihorski, M. (2018). Sensitivity of binomial n-mixture models to overdispersion: The importance of assessing model fit. *Methods in Ecology and Evolution*, 9(10):2102–2114.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Ladin, Z. S., D'Amico, V., Baetens, J. M., Roth, R. R., and Shriver, W. G. (2016). Predicting metapopulation responses to conservation in human-dominated landscapes. *Frontiers in Ecology and Evolution*, 4:122.

Leisen, F., Villa, C., Walker, S. G., et al. (2018). On a class of objective priors from scoring rules. *Bayesian Analysis*.

Lele, S. R., Dennis, B., and Lutscher, F. (2007). Data cloning: easy maximum likelihood estimation for complex ecological models using bayesian markov chain monte carlo methods. *Ecology letters*, 10(7):551–563.

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492):1617–1625.

Link, W. A., Schofield, M. R., Barker, R. J., and Sauer, J. R. (2018). On the robustness of n-mixture models. *Ecology*, 99(7):1547–1551.

Martin, J., Royle, J. A., Mackenzie, D. I., Edwards, H. H., Kery, M., and Gardner, B. (2011). Accounting for non-independent detection when estimating abundance of organisms with a bayesian approach. *Methods in Ecology and Evolution*, 2(6):595–601.

McCaffery, R., Nowak, J. J., and Lukacs, P. M. (2016). Improved analysis of lek count data using n-mixture models. *The Journal of Wildlife Management*, 80(6):1011–1021.

Millar, R. B. (2018). Conditional vs marginal estimation of the predictive loss of hierarchical models using waic and cross-validation. *Statistics and Computing*, 28(2):375–385.

Morris, W. K., Vesk, P. A., McCarthy, M. A., Bunyavejchewin, S., and Baker, P. J. (2015). The neglected tool in the bayesian ecologist's shed: A case study testing informative priors' effect on model accuracy. *Ecology and Evolution*, 5(1):102–108.

Nakagawa, T. and Osaki, S. (1975). The discrete weibull distribution. *IEEE Transactions on Reliability*, 24(5):300–301.

Peluso, A., Vinciotti, V., and Yu, K. (2019). Discrete weibull generalized additive model: an application to count fertility data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3):565–583.

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing parameter identifiability in phylogenetic models using data cloning. *Systematic biology*, 61(6):955–972.

Ponisio, L. C., de Valpine, P., Michaud, N., and Turek, D. (2020). One size does not fit all: Customizing mcmc methods for hierarchical models using nimble. *Ecology and evolution*, 10(5):2385–2416.

Romano, A., Costa, A., Basile, M., Raimondi, R., Posillico, M., Roger, D. S., Crisci, A., Piraccini, R., Raia, P., Matteucci, G., et al. (2017). Conservation of salamanders in managed forests: Methods and costs of monitoring abundance and habitat selection. *Forest ecology and management*, 400:12–18.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60(1):108–115.

Royle, M. K. J. (2015). *Applied hierarchical modeling in ecology : analysis of distribution, abundance and species richness in R and BUGS. Volume 1, Prelude and static models*. Academic Press.

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). Winbugs user manual.

Studds, C. E., Kendall, B. E., Murray, N. J., Wilson, H. B., Rogers, D. I., Clemens, R. S., Gosbell, K., Hassell, C. J., Jessop, R., Melville, D. S., et al. (2017). Rapid population decline in migratory shorebirds relying on yellow sea tidal mudflats as stopover sites. *Nature communications*, 8:14895.

Toribio, S., Gray, B., and Liang, S. (2012). An evaluation of the bayesian approach to fitting the n-mixture model for use with pseudo-replicated count data. *Journal of Statistical Computation and Simulation*, 82(8):1135–1143.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432.

Walker, S. and Villa, C. (2021). An objective prior from a scoring rule. *Entropy*, 23:833.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

6 Supplementary material

6.1 Case 2-Model selection via WAIC MCMC settings

Table 17: MCMC settings for scenario 1.

p	Model	MCMC iterations	Burn-in	Thinning	Chains
0.6	P-B	25000	500	5	1
	DW-B	65000	5000	15	1
	NB-B	125000,	5000	15	1
	P-BB	155000	5000	30	1
	DW-BB	155000	5000	30	1
0.25	P-B	225000	5000	30	1
	DW-B	205000	5000	50	1
	NB-B	405000	5000	100	1
	P-BB	125000	5000	10	1
	DW-BB	205000	5000	100	1

Table 18: MCMC settings for scenario 2.

p	Model	MCMC iterations	Burn-in	Thinning	Chains
0.6	P-B	125000	5000	30	1
	DW-B	205000	5000	50	1
	NB-B	405000	5000	100	1
	P-BB	125000	5000	10	1
	DW-BB	205000	5000	100	1
0.25	P-B	225000	5000	30	1
	DW-B	205000	5000	50	1
	NB-B	405000	5000	100	1
	P-BB	125000	5000	10	1
	DW-BB	205000	5000	100	1

Table 19: MCMC settings for scenario 3.

p	Model	MCMC iterations	Burn-in	Thinning	Chains
0.6	P-B	25000	5000	5	1
	DW-B	85000	5000	20	1
	NB-B	805000	5000	200	1
	P-BB	205000	5000	100	1
	DW-BB	205000	5000	100	1
0.25	P-B	85000	5000	20	1
	DW-B	85000	5000	20	1
	NB-B	805000	5000	200	1
	P-BB	505000	5000	100	1
	DW-BB	505000	5000	100	1

Table 20: MCMC settings for scenario 4.

p	Model	MCMC iterations	Burn-in	Thinning	Chains
0.6	P-B	125000	5000	30	1
	DW-B	405000	5000	100	1
	NB-B	805000	5000	200	1
	P-BB	805000	5000	200	1
	DW-BB	805000	5000	200	1
0.25	P-B	125000	5000	30	1
	DW-B	405000	5000	100	1
	NB-B	805000	5000	200	1
	P-BB	805000	5000	200	1
	DW-BB	805000	5000	200	1

6.2 Case Studies

Table 21: MCMC settings used for analyzing yellow-bellied toads.

Model	MCMC iterations	Burn-in	Thinning	Chains
P-B	25000	5000	5	2
DW-B	900000	40000	25	2
NB-B	700000	50000	15	2
P-BB	605000	50000	15	2
DW-BB	605000	50000	15	2

Table 22: MCMC settings used for analyzing Swiss great tits.

Model	MCMC iterations	Burn-in	Thinning	chains
P-B	25000	5000	5	2
DW-B	900000	40000	25	2
NB-B	350000	50000	15	2
P-BB	605000	50000	15	2
DW-BB	305000	60000	15	2