bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

An efficient extension of N-mixture models for multi-species abundance estimation

Juan Pablo Gomez[1,2,3,*], Scott K. Robinson[2], Jason K. Blackburn[3,4], and José Miguel Ponciano[1,*]

[1]Department of Biology, University of Florida, Gainesville, Florida

[2]Florida Museum of Natural History, Gainesville, Florida

[3]Spatial Epidemiology and Ecology Research Laboratory, Department of Geography, University of Florida, Gainesville Florida

[4]Emerging Pathogens Institute, University of Florida, Gainesville, Florida

[*] *Correspondence author*

Abstract

1. In this paper we propose an extension of the N-mixture family of models that targets an improvement of the statistical properties of the rare species abundance estimators when sample sizes are low, yet of typical size in tropical studies. The proposed method harnesses information from other species in an ecological community to correct each species' estimator. We provide guidance to determine the sample size required to estimate accurately the abundance of rare tropical species when attempting to estimate the abundance of single species.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2. We evaluate the proposed methods using an assumption of 50m radius plots and perform simulations comprising a broad range of sample sizes, true abundances and detectability values and a complex data generating process. The extension of the N-mixture model is achieved by assuming that the detection probabilities of a set of species are all drawn at random from a beta distribution in a multi-species fashion. This hierarchical model avoids having to specify a single detection probability parameter per species in the targeted community. Parameter estimation is done via Maximum Likelihood.

3. We compared our multi-species approach with previously proposed multi-species N-mixture models, which we show are biased when the true abundances of species in the community are less than seven individuals per 100ha. The beta N-mixture model proposed here outperforms the traditional Multi-species N-mixture model by allowing the estimation of organisms at lower densities and controlling the bias in the estimation.

4. We illustrate how our methodology can be used to suggest sample sizes required to estimate the abundance of organisms, when these are either rare, common or abundant. When the interest is full communities, we show how the multi-species approaches, and in particular our beta model and estimation methodology, can be used as a practical solution to estimate organism densities from rapid inventories datasets. The statistical inferences done with our model via Maximum Likelihood can also be used to group species in a community according to their detectabilities.

Keywords: Maximum Likelihood estimation, Sample size estimation, Community Abundance Models, Tropical Species, Hierarchical models, Data cloning.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1 Introduction

Unbiased abundance and occupancy estimates are of paramount value for making inferences about ecological processes and making sound conservation decisions (Hubbell, 2001; Leibold *et al.*, 2004; Margules & Pressey, 2000). To date, quantitative ecologists have proposed several statistical methods to estimate species' detection probabilities and use these to correct the occupancy or abundance estimates (Denes *et al.*, 2015). Our study was motivated by the attempt to use these novel models to estimate the abundance of rare species in tropical communities. In these communities, it is well-known that abundance distributions are typically characterized by long right tails with few abundant species and many rare ones (see Hubbell, 2001). Such high proportion of rare species in the overall community makes it very difficult to obtain enough detections during field censuses for appropriate estimation of both abundance and detection probability for many, if not the majority of tropical species. When we extensively tested via simulations these recent methodologies, we found persistent bias in estimates of low abundances that corresponded to abundance ranges previously not dealt with in temperate forest studies yet common in neotropical studies (see also Yamaura, 2013; Yamaura *et al.*, 2016). As an answer to this problem, in this study we present an alternative, community-based abundance estimation approach that markedly improves these estimates. Our method is widely applicable in communities with similarly abundance patterns.

The N-mixture models aim to tackling the problem of the bias in abundance estimation induced by species differences in detection probabilities (MacKenzie *et al.*, 2002; Martin *et al.*, 2005; Royle & Dorazio, 2008). It uses spatially and temporally replicated counts in which, the counts of species y are binomially distributed with N being the total number of individuals available for detection and p the probability of detecting an individual of that species (Royle, 2004). The model is hierarchical because the abundance N is assumed to be a latent, random process adopting a discrete

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

probability distribution (*e.g.* Poisson). Inferences about the abundance of the species of interest therefore rely on estimating the detection probability and the underlying parameters of the distribution giving rise to N (Royle, 2004). N-mixture models were developed to estimate occupancy/abundance while accounting for imperfect detection of single species (Royle, 2004). Multi-species models have been proposed to deal with estimating the abundance and occupancy of rare species (see Iknayan *et al.*, 2014; Denes *et al.*, 2015, for a review). These models have the advantage to "borrow information" from abundant species in the community to estimate parameters of rare ones (Zipkin *et al.*, 2009; Ovaskainen & Soininen, 2011; Yamaura *et al.*, 2016, 2011; Chandler *et al.*, 2013; Barnagaud *et al.*, 2014). Most of the research and advances in the proposition of multi-species models has focused on estimating occupancy (Iknayan *et al.*, 2014; Denes *et al.*, 2015), even though, studying the abundance and rarity of species is one of the main focuses in ecology (Yamaura *et al.*, 2016; Hubbell, 2001; McGill *et al.*, 2007).

In recent multi-species abundance models, both abundance and detection probabilities are assumed to be normally distributed random effects governed by the community's "hyper parameters" (Iknayan *et al.*, 2014). For that reason they have been named community abundance models, because they focus in describing the characteristics of the entire community from spatially and temporally replicated counts or detections(Yamaura *et al.*, 2012, 2011, 2016). The main assumption behind the community abundance models is that groups of species in the community might share characteristics that make their abundance and detection probability to be correlated (Yamaura *et al.*, 2011, 2012, 2016; Sauer & Link, 2002; Barnagaud *et al.*, 2014; Ruiz-Gutiérrez *et al.*, 2010). These type of abundance community models have been useful for estimating diversity properties of the species assemblages while accounting for imperfect detection (Yamaura *et al.*, 2011, 2012).

While the assumption of normally distributed logit transformed random effects

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

for detection probabilities of species across the community is statistically convenient, other probability distributions might have properties that relate more directly. For example, (Martin *et al.*, 2011) proposed a single species abundance estimation model that allowed individuals within a species to vary in detection probability. They assumed that detection probabilities in a species were described by a beta distribution which naturally ranges between [0-1]. The latter assumption is convenient for community abundance models as well, because it eliminates the need of the logit transformation. Further more, (Dorazio *et al.*, 2013) showed that the beta distribution can be parametrized to reflect the mean detection probability among species and their degree of similarity making the two parameters that determine the shape of the beta distribution ecologically interpretable.

In this study, we: (1) increase the simulation scenarios presented in Yamaura (2013) to provide a full baseline for the sampling design for ecologists that want to estimate the abundance of tropical organisms using N-mixture models, (2) propose and alternative multi-species abundance model that uses a beta distribution for the random effects of detection probability instead of a normal distribution and (3) propose a maximum likelihood approach for multi-species abundance estimation using data cloning (4) compare our alternative multi-species abundance model to one of the previously proposed ones.

1.1 The Model

In the following section, after summarizing the widely used N-mixture models, we develop a multi-species model extension that allows a more accurate estimation of the abundance of rare species. Our approach differs from other multi-species abundance estimation by assuming that detection probabilities in a community are product of a beta distribution instead of a logit transformation of normally distributed random effects.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Using an N-mixture model, we usually let y_{ij} be the number of individuals for a given species in the $i-th$ sampling unit (a point count) and $j-th$ replicate of the sampling unit (or visit to the point count). Let p be the individual detection probability for that species. Finally, let n_i be the fixed number of individuals available for detection in the $i-th$ sampling unit. If we assume that the counts are binomially distributed, the likelihood of the counts (y_{ij}) for a given species is

$$\mathcal{L}(y_{ij}; n_i, p) = \prod_{i=1}^{r} \prod_{j=i}^{t} \binom{n_i}{y_{ij}} p^{y_{ij}} (1-p)^{n_i - y_{ij}}.$$

for $i = 1, 2, 3 \ldots r$ and $j = 1, 2, 3 \ldots t$, where r is the total number of point counts sampled and t is the number of times each point count was visited (Royle, 2004).

The N-mixture model assumes that the number of individuals available for detection is in fact unknown and random. Thus, such number is considered to be a latent variable, modeled with a Poisson process with mean λ (the mean number of individuals per sampling unit). From here on, we write $N_i \sim \text{Pois}\,(\lambda)$, where we have used the convention that lowercase letters such as n_i denote a particular realization of the (capitalized) random variable N_i. To compute the likelihood function, one then has to integrate the binomial likelihood over all the possible realizations of the Poisson process,

$$\mathcal{L}(y_{ij}; \lambda, p) = \prod_{i=1}^{r} \sum_{N_i = \max(\mathbf{y_i})}^{\infty} \prod_{j=1}^{t} \binom{N_i}{y_{ij}} p^{y_{ij}} (1-p)^{N_i - y_{ij}} \frac{e^{-\lambda} \lambda^{N_i}}{N_i!}, \tag{1}$$

where $\mathbf{y_i} = \{y_{i1}, y_{i2}, \ldots, y_{it}\}$. If the objective is to estimate the abundance of S species, the overall likelihood is simply written as the product of all the individual species' likelihoods,*i.e.*,

$$\mathcal{L}(y_{sij}; \underline{\lambda}, \underline{p}) = \prod_{s=1}^{S} \prod_{i=1}^{r} \sum_{N_{si} = \max(\mathbf{y_{si}})}^{\infty} \prod_{j=1}^{t} \binom{N_{si}}{y_{sij}} p_s^{y_{sij}} (1-p_s)^{N_{si} - y_{sij}} \frac{e^{-\lambda_s} \lambda_s^{N_{si}}}{N_{si}!}, \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

where y_{sij} is a three dimensional array of dimensions $r \times t \times S$, and both $\underline{\lambda} = \{\lambda_1, \ldots, \lambda_S\}$ and $\underline{p} = \{p_1, \ldots, p_S\}$ are vectors of length S. In what follows, we will refer to the n_{si}'s as the latent, realized abundance and to the mean abundances, the λ_s's simply as the "abundances". To avoid the proliferation of parameters one could assume that all the p_s come from a single probability model that describes the community-wide distribution of detection probabilities (Yamaura *et al.*, 2011, 2012, 2016; Sauer & Link, 2002; Barnagaud *et al.*, 2014; Ruiz-Gutiérrez *et al.*, 2010). These community-wide detection probabilities can be modeled with a beta distribution in which we let $P_s \sim \text{Beta}(\alpha, \beta)$. The probability density function of the random detection probabilities is then $g(p_s; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_s^{\alpha-1}(1-p_s)^{\beta-1}$.

Following (Dorazio *et al.*, 2013), we parameterize the Beta distribution as $Beta(\alpha = \tau\overline{p}, \beta = \tau(1-\overline{p}))$ such that the parameters are related to biological processes. Here, $\overline{p}$ is the mean detection probability among species in the community and τ is a measurement of the similarity in detection probabilities (Dorazio *et al.*, 2013).

The overall likelihood function now integrates over all the realizations of the community-wide detection probabilities P_s:

$$\mathcal{L}(y_{sij}; \underline{\lambda}, \overline{p}, \tau) = \int_0^1 \prod_{s=1}^{S} \prod_{i=1}^{r} \sum_{N_{si}=\max(\mathbf{y_{si}})}^{\infty} \prod_{j=1}^{t} \binom{N_{si}}{y_{sij}} p_s^{y_{sij}} (1-p_s)^{N_{si}-y_{sij}} \frac{e^{-\lambda_s}\lambda_s^{N_{si}}}{N_{si}!}$$

$$\times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_s^{\alpha-1}(1-p_s)^{\beta-1} dp_s. \tag{3}$$

The usefulness of specifying the likelihood in this way is that in the case in which many species are rare, we can use the information on the abundant species to estimate the

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

detection probability, leaving the actual counts to estimate only the abundance of the species. Note that by integrating the beta process at the outmost layer of the model, we are following the sampling structure. When this approach is used and the integral is tractable, the resulting distribution is a multivariate distribution with a specific covariance structure (Sibuya *et al.*, 1964). Thus, we expect our approach to result in a multivariate distribution of counts with a covariance structure arising naturally from the sampling design and the assumed underlying beta process of detectabilities.

1.2 Maximum Likelihood Estimation

One drawback of the beta-N-mixture and other models for multi-species abundance estimation is their computational complexity, which imposes a substantial numerical challenge for Maximum Likelihood (ML) estimation. Such problem is not unique to abundance estimation but to many other hierarchical models in ecology (Lele & Dennis, 2009). For those reasons, parameter estimation in hierarchical models is usually performed under a bayesian framework (Cressie *et al.*, 2009). To date however, many numerical approximations for obtaining the Maximum Likelihood Estimates (MLEs) for hierarchical models have been proposed (de Valpine, 2012). The "Data Cloning" methodology has proven to be a reliable approach to obtain the MLEs, hypothesis testing and model selection, as well as unequivocally measuring the estimability of parameters for hierarchical models (Lele *et al.*, 2010; Ponciano *et al.*, 2012). The method proposed by Lele *et al.* (2007, 2010) uses the Bayesian computational approach coupled with Monte Carlo Markov Chain (MCMC) to compute Maximum Likelihood Estimates (MLE) of parameters of hierarchical models and their asymptotic variance estimates (Lele *et al.*, 2007). The advantage of using the data cloning protocol is that one only needs to compute means and variances of certain posterior distributions.

Data Cloning proceeds by performing a typical Bayesian analysis on a dataset

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

that consists of k copies of the originally observed data set. In other words, to implement this method, one has to write the likelihood function of the data as if, one had observed k identical copies of the data set. Then, Lele *et al.* (2007, 2010) show that as k grows large, the mean of the resulting posterior distribution converges to the MLE. In addition, for continuous parameters as $\underline{\lambda}$, $\overline{p}$, and τ, the variance covariance matrix of the posterior distribution converges to $\frac{1}{k}$ times the inverse of the observed Fisher's information matrix. Thus, the variance estimated by the posterior distribution can be used to calculate Wald-type confidence intervals of the parameters (Lele *et al.*, 2007, 2010). The advantage of data cloning over traditional Bayesian algorithms is that while in Bayesian algorithms the prior distribution might have influence over the posterior distribution, in data cloning the choice of the prior distribution does not determine the resulting estimates. In our case, the hierarchical model is

$$\begin{aligned}
\mathbf{Y} \sim \text{Binomial}\left(\underline{\mathbf{N}}, \mathbf{P}\right) &= f(y|\underline{\mathbf{N}} = n, \mathbf{P} = p) \quad \text{(Observation model)}, \\
\underline{\mathbf{N}} \sim \text{Pois}\left(\underline{\lambda}\right) &= g(\underline{\mathbf{N}}; \underline{\lambda}) \quad \text{(Process model)}, \\
\mathbf{P} \sim \text{Beta}\left(\overline{p}\tau, (1-\overline{p})\tau\right) &= h(\mathbf{P}; \overline{p}, \tau) \quad \text{(Process model)}.
\end{aligned}$$

$\underline{\mathbf{N}}$ and $\mathbf{P}$ are latent variables which are products of a stochastic process given by the Poisson and Beta distributions respectively. Furthermore, $\underline{\lambda}$, and $\overline{p}, \tau$ are seen as random variables themselves that have a posterior distribution $\pi(\underline{\lambda}, \overline{p}, \tau|\mathbf{Y})$. A typical Bayesian approach would sample from the following posterior distribution:

$$\pi(\underline{\lambda}, \overline{p}, \tau, \underline{\mathbf{N}}, \mathbf{P}|\mathbf{Y}) \propto \left[f(y|\underline{\mathbf{N}} = n, \mathbf{P} = p) g(\underline{\mathbf{N}}; \underline{\lambda}) h(\mathbf{P}; \overline{p}, \tau)\right] \pi(\underline{\lambda}, \overline{p}, \tau),$$

where $\pi(\underline{\lambda}, \overline{p}, \tau)$ is the joint prior of the model parameters. This approach would yield many samples of the vector $(\underline{\lambda}, \overline{p}, \tau, \underline{\mathbf{N}}, \mathbf{P})$ and in order to sample from the marginal

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

posterior $\pi(\underline{\lambda}, \overline{p}, \tau|\mathbf{Y})$ oneneeds to look at the samples of the subset of $\underline{\lambda}$, $\overline{p}$, and, τ. The data cloning approach proceeds similarly, except one needs to sample from the following posterior distribution:

$$\pi(\underline{\lambda}, \overline{p}, \tau, \underline{\mathbf{N}}, \mathbf{P}|\mathbf{Y})^{(k)} \propto \left[f(y|\underline{\mathbf{N}} = n, \mathbf{P} = p)g(\underline{\mathbf{N}}; \underline{\lambda})h(\mathbf{P}; \overline{p}, \tau)\right]^k \pi(\underline{\lambda}, \overline{p}, \tau).$$

The notation $^{(k)}$ on the left side of this equation does not denote an exponent but the number of times the data set was "cloned". On the right hand side, however, k is an exponent of the likelihood function. The MLEs of $\underline{\lambda}$, $\overline{p}$, and, τ are then simply obtained as the empirical average of the posterior distribution $\pi(\underline{\lambda}, \overline{p}, \tau|\mathbf{Y})^{(k)}$ and the variance of the estimates are given by $\frac{1}{k}$ times the variance of this posterior distribution.

2 Methods

2.1 Estimation for Single Species

To determine the minimum sample size required for accurate estimation of the abundance of tropical species, we used a series of simulations where we varied the number of plots (r), visits to plots (t), mean number of individuals in a 100 ha plot (λ) and detection probability (p). We varied r between 5 and 50, t between 2 and 20, $\lambda = 1, 2, 3, 4, 5, 7, 10, 15, 25, 40, 55, 65, 75, 85, 100$ and p between 0.1 and 0.9. For each combination of parameters, we simulated 170 data sets and estimated λ and p using equation 1. In each simulation, we computed the relative bias of the abundance estimate by using, $bias = \frac{\hat{\lambda}-\lambda}{\lambda}$, where $\hat{\lambda}$ is the MLE for a particular data set and λ is the true value of the parameter. Finally, we retained the mean bias for each combination of model parameters. We considered an acceptable bias to be lower than 0.1, which

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

is a 10% difference between the estimate and the true population density. All of the simulations were performed using R statistical software v.3.0.2 (R Core Team, 2013) and maximum likelihood estimation by maximizing the likelihood of eq (1) using the optim function with the Nelder-Mead algorithm. The R code used for simulations and maximum likelihood estimation is presented in the Appendix C.

2.2 Assessing the Beta N-mixture Model performance

To assess the Beta N-mixture Model performance we followed three steps: First, we simulated 1500 data sets under the model, compute the ML estimates of our model parameters each time, and then examine the distribution of the MLEs. The objective of this approach was to evaluate if the average of the distribution of ML estimates gets at the true parameter values and also, if the variability around those estimates is small. In a reality, data come from a much more complex process involving many variables and quantities. Therefore, we also tested the robustness of our model by simulating data from a complex, spatially explicit data-generating process. To do that, we simulated 500 datasets under a spatially-explicit model (see description below) and then estimated the abundances and detection probabilities using our model. We compared the performance of our model *vis-à-vis* a previously proposed multi-species abundance model (Yamaura *et al.*, 2016). From here on, we refer to Yamaura *et al.* (2016)'s approach as the Normal N-mixture model. Finally, the third step of our performance assessment consisted in estimating the abundance of 26 species of neotropical dry forest birds using a perviously non-published dataset. The objective of this step was to illustrate the use of our model with a realistic scenario.

2.2.1 Bias benchmark assessment

To evaluate the bias of the Beta N-mixture model, we simulated species counts in a 100 ha quadrant sampled using 25, 50 meter circular plots visited three times each.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We assumed that the community was composed by 15 species with mean number of individuals/100 ha of $\lambda = 1, 2, 3, 4, 5, 7, 10, 15, 25, 40, 55, 65, 75, 85, 100$. In the latter vector each value of λ represents the abundance of a single species. In each simulation we drew $N_i j$ individuals in each plot from a Poisson distribution with mean λ_i. We then simulated the detection process using a Binomial distribution with parameters $N_i j$ and p_i. We varied mean detection probability by assuming $\overline{p} = 0.25, 0.5, 0.75$ and $\tau = 4.5$ (E[p] = 0.25, 0.5, 0.75;Var[p] = 0.03, 0.04, 0.03). Even though the variance seems small, the 2.5% and 97.5% quantiles of the three distributions range over a large portion of the [0,1] interval (quantiles 2.5 and 97.5: low = (0.01,0.68); mid = (0.1,0.89); high = (0.31,0.98)). For each type of community we simulated 500 data sets, and estimated λ_i, $\overline{p}$ and τ using data cloning. To determine the number of clones required to accurately get to the Maximum Likelihood Estimates of λ_i, $\overline{p}$ and τ we used one randomly generated data set and estimated the parameters cloning the data sequentially from 1 to 64 times (Lele *et al.*, 2010). This allowed us to determine an adequate number of clones to get convergence of the $k - th$ posterior mean to the MLEs. We used rjags v. 4.2.0 (Plummer, 2014) with two Markov chains allowing each chain to run for 14000 generations sampling every 10 generations and discarded the first 4000 iterations. For each type of community we estimated the relative bias ((Estimated-True)/True) in λ_i, $\overline{p}$ and τ.

2.2.2 Comparison to other community abundance models

There are two essential differences between the Beta and Normal N-mixture models. The first one is that the Beta model treats abundance (the mean abundance, that is, see definition of the λ parameters above) as a fixed effects instead of random. As a result, the Normal N-mixture model has an extra hierarchy level than our model. Both models are hierarchical stochastic models where the binomial sampling model is the first hierarchy level, then, in both, the realized abundances (the N's) and the

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

detection probabilities are the inner hierarchies. The Normal N-mixture model goes one more level and assumes that the parameters λ governing the realized abundances N also come from a stochastic process governed itself by hyper-parameters. The second difference between our model and the Normal N-mixture model is the distributional assumption giving rise to detection probabilities. In our model p_s are assumed to be $p_s \sim Beta(\tau\overline{p}, \tau(1-\overline{p}))$ and in the Normal model, $p_s = \frac{1}{1+e^{(-(r_i))}}$ where $r_i \sim Norm(\mu, \sigma^2)$ which gives a Johnson's SB distribution between 0 and 1. Besides these two model differences, Yamaura *et al.* (2016) use a Bayesian approach to fit their hierarchical model, while we use the method of Maximum Likelihood estimation. Much discussion exists regarding the merits of each inferential approach for hierarchical models in Ecology (see for instance Lele & Dennis, 2009; Cressie *et al.*, 2009). Here we limit ourselves to compare the results from Yamaura *et al.* (2016)'s estimation approach, taken as the benchmark of a known method in the literature, to our approach. Note that using data cloning, one could compute the Maximum likelihood estimates of the model parameters in Yamaura *et al.* (2016)'s model, and their associated properties but doing so is outside the scope of our study.

We simulated 500 data sets under a spatially explicit model and for each data set we fitted the Normal N-mixture model then compared the posterior mean and mode estimates with the MLEs for the model proposed here (see Figure 2). For each simulation, we randomly drew 30 λ_i from a gamma distribution with parameters $\alpha = 0.65$, $\beta = 0.033$ and excluded λ_i values smaller than 1 individuals/100 ha, resulting in a community of 27 species. The gamma distribution used is the best fit of an observed species abundance distribution of a neotropical bird assemblage that was gathered using field intensive methods (Robinson *et al.*, 2000). Following, we randomly drew from a poisson distribution with mean λ_i the number of individuals of the $i-th$ species (N_i) present in a 100 hectares plot. We located each individual randomly across the plot and following, we randomly placed 25 circular plots with a radius of 50 meters

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the 100 ha plot that were separated by at least 150 meters. Finally, we obtained species specific detection probability (p_i) from a uniform distribution. To obtain the counts, we drew the number of individuals detected in a 50 meter-radius plot from a Binomial distribution using the number of individuals in each 50 m plot N_{ij} and the individual's detection probability p_i. We repeated the detection process three times to generate three temporal replicates of the sampling process. The R-function to simulate the described process is presented in Appendix C.

For each of the simulated data sets we estimated λ_i, $\overline{p}$ and τ under the Beta N-mixture model using maximum Likelihood estimation with Data Cloning (Lele *et al.*, 2007). We used rjags (Plummer, 2014) to build the model and run the analysis with 2 chains, with 15000 iterations in each chain and retained the parameter values every 10 generations after a burn-in period of 4000 generations. After initial parameter estimation, we sampled the posterior distribution given the estimated parameters to obtain the realized values of p_i given the data. For the Normal N-mixture model we performed bayesian parameter estimation using rjags and ran the analysis using 2 chains, with 50000 iterations and retained parameters values every 20 generations after a burn-in of 10000 generations. In the latter case, we retained the mean and mode of λ_i, p_i for comparison with the beta N-mixture model.

2.3 Example Using Real Data

Finally, we used a data set that consisted of 94 point counts, located in three dry forest patches in Colombia. Bayesian and Maximum likelihood estimation for the Normal and Beta N-mixture models respectively were performed in the same way as described in the previous section. Details of the sampling procedure the R code and jags models used are presented in the Appendix (Appendix B, C)

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

3 Results

3.1 Estimation for Single Species

We found that the required minimum sample size needed to accurately estimate the abundance of tropical organisms decreased with increasing both λ and p (Figure 1). For the sample sizes evaluated, there is no combination of point counts and replicates that allows the estimation of abundances with less than 7 individuals/100ha using single species N-mixture models (Figure A1). In the 7 ind/100 ha threshold, the effort required is very high. For example, for species with a probability of detection of 0.5 the required sample size to obtain a bias lower than 0.1 is around 50 points and more than 6 replicates of each point count or around 40 point counts with more than 10 replicates (Figure 1,A1). As λ increases the sample size required to estimate appropriately the abundance of species decreases.

3.2 Assessing the Beta N-mixture Model performance

3.2.1 Bias Benchmark assessment

We found that the parameters of the Beta N-mixture model are fully identifiable since the relative magnitude of the first eigenvalue of the parameter variance-covariance matrix decreased very closely at a rate of $1/k$ ($eigenvalue = -0.066 + 1.019(1/k)$; $r^2 = 0.98$). This result also identified that 20 clones were sufficiently large to guarantee convergence to the MLEs. The Beta model tends to slightly overestimate the abundance of rare species and underestimate the abundance of abundant species but this tendency decreases with increasing detection probability (Figure A2). This is evidenced by the slopes estimated by the relationship between estimated and true λ. The relationship for $p = 0.25$ resulted was $\hat{\lambda} = 5.8 + 0.7\lambda$, for $p = 0.5$ was $\hat{\lambda} = 4 + 0.9\lambda$ and for $p = 0.75$ was $\hat{\lambda} = 3.3 + 0.95\lambda$. The bias decreased (approximately) as a function of the true value of λ according to the equation $bias(\lambda) = -0.45(\frac{1}{\lambda} + 7.5)$ for

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$p = 0.25$, and $bias(\lambda) = -0.26(\frac{1}{\lambda} + 5.6)$ for $p = 0.5$ and $bias(\lambda) = -0.2(\frac{1}{\lambda} + 5)$ for $p = 0.75$.

Assuming that a 10% bias in the estimation is acceptable, the minimum λ that the model is able to estimate is 13 - 17 individuals/100 ha irrespective of the detection probability. It is noted however, that a bias of 100% in the low abundance end has little impact over the ecological interpretation of the estimates. Thus, if one sets bias in the abundance estimates to 100% (left hand side in the bias functions above) the model is able to predict the density of species with 3 - 5 individuals/100 ha.

The beta N-mixture model also performs well in estimating the distribution of the community's detection probability (Figure A3). The distribution of $\overline{p}$ for the simulations is almost centered in the true value of p. There is a slight overestimation of p when $p = 0.25$ (Figure A3). The model tends to underestimate $\widehat{\text{Var}[p]}$, but estimates it to be similar across the different types of simulations (Figure A3).

3.2.2 Comparison to other community abundance models

The beta N-mixture model performed better than the Normal model in estimating the abundance and detection probability of rare species. While the posterior means and modes of the Normal model were biased towards species with abundances lower than 4 individuals/100 ha, Maximum Likelihood Estimates of the Beta model were not (Figure 3). Furthermore, we show that the posterior means tended to be more biased than the posterior mode in estimating λ (Figure 3). The opposite seems to be true for the detection probabilities p. Both, the posterior mode and mean underestimated p for rare species (Figure 4).

3.3 Example Using Real Data

We present the estimates of $\hat{\lambda}$ for both models in Table 1. The estimates of the abundances resulted very similar for both Beta and Normal N-mixture models. The

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

confidence intervals of the Beta N-mixture and Normal N-mixture overlapped for every species (Table 1). The differences in the estimates are slightly higher for rare species when estimated using the Normal N-mixture model. The Beta model estimated $\overline{p} = 0.26(0.2, 0.3)$ and $\tau = 13.5(11.9, 15)$. The normal model estimated $\mu = -1.22(-1.5, -1)$ and $\sigma^2 = 0.2(0.01, 0.6)$. The latter result translates in mean detection probability across species of $\hat{p} = 0.23(0.18, 0.27)$.

4 Discussion

Our results can be discussed around three main findings. The first one is that most tropical species are too rare to estimate with single species N-mixture models and a typical sample size in tropical studies. Single species N-mixture models require a high number of spatial and temporal replicates to accurately estimate the abundance of tropical organisms (Figure 1, see also Yamaura, 2013). The second one is that we found that the MLEs of a wide range of abundances computed using the beta N-mixture model have good statistical properties. Among these properties is a low relative bias of the quantities we estimate (the detectabilities and the mean abundances). Our approach leads to unbiased estimates of the abundance of extremely rare species with 1-3 individuals/100 ha (Figure 3, Figure A2). Third, we show that the MLEs of the Beta N-mixture model parameters have lower bias than the estimates provided by Yamaura *et al.* (2016)'s Bayesian fitting of the Normal N-mixture model (Figures 3,4).

N-mixture models have been proven to be useful in scenarios where species are abundant (e.g. Royle, 2004; Joseph *et al.*, 2009). If the objective of the study is to estimate the abundance of a single species correcting for its detection probability, then our simulations are a guide to the sampling effort required. Published databases (e.g. Parker III *et al.*, 1996; Karr *et al.*, 1990), include estimates of abundance of

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

many neotropical species, which could provide general guidelines to researchers in the field about the approximate λ they are dealing with and thus the approximate sample sizes needed to correctly estimate the abundance using N-mixture models.

For rare species, the solution is to use the community abundance models. Our study and Yamaura *et al.* (2016) provide two examples of how to apply the estimation of the abundance to a set of species. Our approach has the additional advantage that it provides estimates with low bias even for species with low abundance and detection probabilities. For example, for communities with $\overline{p} = 0.25$, the mean bias for species with one individual/100 ha is around 700% (Figure A2). This number sounds extreme but it only increases the abundance from one to seven individuals/100ha having little effect over the ecological inferences drawn from the model. Furthermore, estimating the parameters of the Beta N-mixture model using a larger set of species in the community seems to correct this bias. For example, our simulation under a more complex model, shows that the Beta N-mixture model has almost no bias in estimating the abundance of species close to 1 individual/100 ha (Figure 3). The bias correction demonstrate that the larger the community is, the less biased the estimates are likely to be. The latter is particularly convenient for tropical communities that are likely to have high species richness increasing the amount of information available to estimate the parameters of the entire community.

In comparison to other community abundance models, and specifically to the one in Yamaura *et al.* (2016), the Beta N-mixture model has lower bias in both $\hat{\lambda}$ and $\overline{p}$. It is unknown however, why the bias of rare species arises, since an exponential transformation of a normal distribution predicts a high number of rare species. The same scenario arises with $\overline{p}$ since the logit transformation of the normal distribution is more flexible than the beta distribution (Hafley & Schreuder, 1977). One explanation is that the extra level of hierarchy required by performing the transformations of the normal distribution has an influence over the estimates. Another possibility is that

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the prior distribution selected to perform the bayesian estimation affects the location of the posterior means and modes. This result only highlights the fact that prior elicitation can be the most difficult step in a bayesian analysis of a hierarchical model, given that the parameters for which priors need to be specified are by definition un-observable. Because of these difficulties, and as (Dorazio, 2016) and (Taper & Ponciano, 2016) state, in a Bayesian analysis of a hierarchical models, it is important to carefully validate the inference of these computer-intensive techniques by means of extensive simulations that combine frequentist ideas to test the properties of posterior modes or means of a Bayesian analysis.

One little-explored issue of the estimation of abundances using complex hierarchical models fitted *via* a bayesian approach, is assessing if and when prior distributions affect the estimates of the model parameters. As Lele & Dennis (2009) mention, different un-informative priors can produce different posterior distributions that alter the inferences drawn from the model. In particular, the use of different priors in the estimation of the probability of the detection parameter in a binomial distribution has been shown to have strong effects on the posterior distribution (Tuyl *et al.*, 2008). The latter result is of particular interest for community abundance estimation since the counts used to estimate abundance in community models are assumed to be binomially distributed. It is important to recognize that strong effects from the priors might not occur in cases where the data is so extensive and complete that the information contained in the samples widely overshadows the information provided by the priors. However, without extensive simulations it is difficult to known if such is the case. To carry Maximum Likelihood estimation *via* Data Cloning (Lele *et al.*, 2010) one essentially tricks a bayesian algorithm into computing the Maximum Likelihood estimates but notably, this procedure can be started with any prior distribution for the model parameters (as long as their support makes biological and mathematical sense) and always converge to the same estimates (Lele *et al.*, 2007). Also, the data

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cloning approach has the advantage that one can easily assess parameter identifiability for hierarchical models and determine when the model has too many hierarchy levels. Here, we demonstrated that all the beta N-mixture model parameters are identifiable using Lele *et al.* (2010)'s approach. Finally, we speculate that Normal N-mixture model, when specified with one less hierarchy, could very well lead to unbiased estimates similar to ours. If so, then this would imply that the information in the data "swamps" the diffuse priors specified by these authors by default.

Because our model is essentially identical to any N-mixture model, it can be adapted to any underlying distribution of abundances. For example, the Poisson distribution used to model the mean number of individuals can be replaced by any other distribution that relaxes the homogeneity assumption (e.g. Negative Binomial or Zero Inflated Poisson). In addition, ecological inferences can be made by incorporating covariates of the abundance process in the model as previously suggested with N-mixture models (Joseph *et al.*, 2009; Yamaura *et al.*, 2011, 2012). The detection process can also depend on variables influencing the overall detectability of species by making the parameters of the beta distribution a function of the covariates (Dorazio *et al.*, 2013). One can assume that the detection probability distribution is a function of variables such as the functional groups or to the microhabitat used for foraging and other species' intrinsic characteristics that might be evolutionarily constrained (Yamaura *et al.*, 2011, 2012; Ruiz-Gutiérrez *et al.*, 2010). Model selection comparing models with and without abundance and detection covariates can be useful for inferring ecological mechanisms underlying the abundance of species (Joseph *et al.*, 2009). In the beta N-mixture model, the assumption of the correlated behavior can be tested by comparing it to a regular N-mixture model, and because the main difference is in the assumptions underlying detection probability, it allows us to make inferences about ecological similarity among species in the same guild, habitat or functional group. We note however, that our simulations shown above were performed using a

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

uniform distribution for p_i. Such model clearly violates the assumption of correlated detection probabilities, but the flexibility of the beta and logit-normal distributions allow to estimate with high confidence the parameters underlying the species' counts.

The estimates of the abundance of the understory insectivores of the upper Magdalena Valley show little difference between the beta N-mixture and and Normal N-mixture models relies on the estimation of the abundance of rare species (Table 1). It is worth noting that the abundance of more common species with higher numbers of detections in our dataset might be a little bit higher than in other published data sets (Karr *et al.*, 1990). There are three possible reasons for this. First, when the mean detection probability of the species is low, our simulations showed that the beta-mixture model overestimated the true abundance of species (Figure A3). The second reason is more ecological: the data presented here comes from the dry forests of the Magdalena valley. Even though this ecosystem is a less species rich than wet forest ecosystems, the biomass of the community does not change (Gomez et al. unpublished data). This means that the populations of most species tend might be higher than in wet forests from which most of the abundance data for neotropical birds have been collected (Terborgh *et al.*, 1990; Thiollay, 1994; Robinson *et al.*, 2000; Blake, 2007). Third, it is also possible that rare species do not have to sing much to defend their territories because they have few neighbors. Common species, on the other hand, face a constant threat of territorial intrusion and may have to sing more. The categorical abundance estimates from Parker III *et al.* (1996) compared to the estimates using both Beta and Normal N-mixture models are similar. In particular, Table 1 shows how most of the species that are categorized as common (C) and fairly common (F) by Parker III *et al.* (1996), the models estimate abundances to be larger than 30 individuals/100 ha. In our opinion, the most exciting result is the appropriate estimation of extremely rare species (*e.g. Dromococcyx phasianellus*) which the models accurately estimate them as rare with only 1 or 2 detections in the

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

entire data set. In these species are for which the single species N-mixture models have particular problems estimating because of the lack of information.

Our simulations have pushed the limits of community abundance models by simulating species with lower abundance than any other simulation (see Yamaura *et al.*, 2016). We hope that our results encourage tropical ecologists to use community abundance hierarchical models as a means to adequately estimate the abundance of full communities. In the recent North American Ornithological congress (August 2016), two of us (JPG and SKR) participated in a wide, round table discussion where it was evident that tropical ornithologists are currently facing strong publishing challenges because so far, abundance estimating techniques have not explicitly targeted estimation in a setting like the tropics: with very low abundances and sparse counts. Unlike temperate forests, where these methodologies have been widely used, in the tropics the species number is typically very large, but the counts per species very low. Our results, although worked out using birds as a study system, suggest that it is possible to have a reasonable estimates of the density of all of the species in the community for this particular scenario and different taxonomic groups (*e.g.* mammals, insects, plants, fungi, bacteria). Unbiased estimation of abundances using these hierarchical models will hopefully enable building more accurate species abundance distributions, which in turn can be extremely useful for understanding the mechanisms governing biodiversity patterns (McGill *et al.*, 2007)

5 Acknowledgements

We would like to thank the farm owners Cesar Garcia, Hacienda los Limones and Constanza Mendoza for allowing us to perform bird counts in their properties. Gordon Burleigh, Bette Loiselle, David Steadman and Philip Shirk, and two anonymous reviewers provided useful comments for the development of the model and improve-

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

ment of the manuscript. This work was supported by the National Institutes of Health Grant 1R01GM117617-01 to JKB (PI) and JMP (Co-PI).

References

Barnagaud, J.Y., Barbaro, L., Papaïx, J., Deconchat, M. & Brockerhoff, E.G. (2014) Habitat filtering by landscape and local forest composition in native and exotic new zealand birds. *Ecology*, **95**, 78–87.

Blake, J.G. (2007) Neotropical forest bird communities: a comparison of species richness and composition at local and regional scales. *The Condor*, **109**, 237–255.

Chandler, R.B., King, D.I., Raudales, R., Trubey, R., Chandler, C. & Arce Chávez, V.J. (2013) A small-scale land-sparing approach to conserving biological diversity in tropical agricultural landscapes. *Conservation Biology*, **27**, 785–795.

Cressie, N., Calder, C.A., Clark, J.S., Hoef, J.M.V. & Wikle, C.K. (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, **19**, 553–570.

de Valpine, P. (2012) Frequentist analysis of hierarchical models for population dynamics and demographic data. *Journal of Ornithology*, **152**, 393–408.

Denes, F.V., Silveira, L.F. & Beissinger, S.R. (2015) Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation. *Methods in Ecology and Evolution*, **6**, 543–556.

Dorazio, R.M. (2016) Bayesian data analysis in population ecology: motivations, methods, and benefits. *Population Ecology*, **58**, 31–44.

Dorazio, R.M., Martin, J. & Edwards, H.H. (2013) Estimating abundance while ac-

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

counting for rarity, correlated behavior, and other sources of variation in counts. *Ecology*, **94**, 1472–1478.

Hafley, W. & Schreuder, H. (1977) Statistical distributions for fitting diameter and height data in even-aged stands. *Canadian Journal of Forest Research*, **7**, 481–487.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography*, volume 32. Princeton University Press, Princeton, NY.

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.

Joseph, L.N., Elkin, C., Martin, T.G. & Possingham, H.P. (2009) Modeling abundance using n-mixture models: the importance of considering ecological mechanisms. *Ecological Applications*, **19**, 631–642.

Karr, J.R., Robinson, S.K., Blake, J.G., Bierregaard Jr, R.O. & Gentry, A. (1990) Birds of four neotropical forests. A.H. Gentry, ed., *Four neotropical rainforests*, pp. 237–269. Yale University Press New Haven, Connecticut.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D. *et al.* (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology letters*, **7**, 601–613.

Lele, S.R. & Dennis, B. (2009) Bayesian methods for hierarchical models: are ecologists making a faustian bargain. *Ecological Applications*, **19**, 581–584.

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

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

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

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature*, **405**, 243–253.

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

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J. & Possingham, H.P. (2005) Zero tolerance ecology: improving ecological inference by modeling the source of zero observations. *Ecology letters*, **8**, 1235–1246.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I. & White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology letters*, **10**, 995–1015.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C. *et al.* (2001) Terrestrial ecoregions of the world: A new map of life on earth a new global

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, **51**, 933–938.

Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**, 289–295.

Parker III, T., Stotz, D. & Fitzpatrick, J. (1996) Ecological and distributional databases for neotropical birds. D. Stotz, J. Fitzpatrick, T. Parker III & D. Moskovits, eds., *Neotrpical birds: ecology and conservation*. University of Chicago Press, Chicago.

Plummer, M. (2014) *rjags: Bayesian graphical models using MCMC*. R package version 3-13.

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

R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robinson, W.D., Brawn, J.D. & Robinson, S.K. (2000) Forest bird community structure in central panama: influence of spatial scale and biogeography. *Ecological Monographs*, **70**, 209–235.

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

Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Academic Press, San Diego, CA.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Ruiz-Gutiérrez, V., Zipkin, E.F. & Dhondt, A.A. (2010) Occupancy dynamics in a tropical bird community: unexpectedly high forest use by birds classified as non-forest species. *Journal of Applied Ecology*, **47**, 621–630.

Sauer, J.R. & Link, W.A. (2002) Hierarchical modeling of population stability and species group attributes from survey data. *Ecology*, **83**, 1743–1751.

Sibuya, M., Yoshimura, I. & Shimizu, R. (1964) Negative multinomial distribution. *Annals of the Institute of Statistical Mathematics*, **16**, 409–426.

Taper, M.L. & Ponciano, J.M. (2016) Evidential statistics as a statistical modern synthesis to support 21st century science. *Population Ecology*, **58**, 9–29.

Terborgh, J., Robinson, S.K., Parker III, T.A., Munn, C.A. & Pierpont, N. (1990) Structure and organization of an amazonian forest bird community. *Ecological Monographs*, **60**, 213–238.

Thiollay, J.M. (1994) Structure, density and rarity in an amazonian rainforest bird community. *Journal of Tropical Ecology*, **10**, 449–481.

Tuyl, F., Gerlach, R. & Mengersen, K. (2008) A comparison of bayes–laplace, jeffreys, and other priors: The case of zero events. *The American Statistician*, **62**, 40–44.

Yamaura, Y. (2013) Confronting imperfect detection: behavior of binomial mixture models under varying circumstances of visits, sampling sites, detectability, and abundance, in small-sample situations. *Ornithological Science*, **12**, 73 – 78.

Yamaura, Y., Andrew Royle, J., Kuboi, K., Tada, T., Ikeno, S. & Makino, S. (2011) Modelling community dynamics based on species-level abundance models from detection/nondetection data. *Journal of applied ecology*, **48**, 67–75.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Yamaura, Y., Kéry, M. & Royle, J.A. (2016) Study of biological communities subject to imperfect detection: bias and precision of community n-mixture abundance models in small-sample situations. *Ecological Research*, **31**, 289–305.

Yamaura, Y., Royle, J.A., Shimada, N., Asanuma, S., Sato, T., Taki, H. & Makino, S. (2012) Biodiversity of man-made open habitats in an underused country: a class of multispecies abundance models for count data. *Biodiversity and Conservation*, **21**, 1365–1380.

Zipkin, E.F., DeWan, A. & Andrew Royle, J. (2009) Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. *Journal of Applied Ecology*, **46**, 815–822.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

6 Tables

			Yamaura model			Beta model		
Species	**Det**	**Parker**	**97.5%**	**Mean**	**2.5%**	**97.5%**	**MLE**	**2.5%**
Atalotriccus pilaris	83	F	97.3	145.2	206.1	71.3	122.8	174.3
Basileuterus rufifrons	104	C	146.4	208.6	300.9	111.2	204.3	297.3
Campylorhynchus griseus	7	C	5.0	14.5	30.1	0.0	11.2	22.5
Cantorchilus leucotis	3	C	2.9	10.3	24.1	0.0	8.2	19.5
Cnemotriccus fuscatus	31	F	39.3	67.0	110.9	24.3	67.2	110.2
Contopus cinereus	2	F/P	1.7	7.8	19.8	0.0	5.2	13.4
Cymbilaimus lineatus	4	F	4.1	12.9	28.8	0.0	11.3	25.0
Dromococcyx phasianellus	1	U	0.8	5.5	15.8	0.0	2.5	7.7
Elaenia flavogaster	67	C	107.9	162.8	260.6	85.7	192.3	298.8
Euscarthmus meloryphus	26	C	28.1	49.8	81.0	17.3	44.3	71.3
Formicivora grisea	172	C	225.4	315.0	433.1	172.6	279.0	385.4
Hemitriccus margaritaceiventer	106	C	104.2	161.6	231.4	83.6	124.4	165.1
Henicorhina leucosticta	28	F	37.7	65.8	113.6	20.9	70.9	121.0
Hylophilus flavipes	144	C	236.1	344.8	580.2	134.1	445.8	757.5
Leptopogon amaurocephalus	23	F	27.0	49.1	83.4	15.1	47.1	79.2
Myrmeciza longipes	64	C	81.2	121.6	178.9	60.1	111.6	163.1
Myrmotherula pacifica	1	F	0.8	5.5	15.4	0.0	2.5	7.5
Pheugopedius fasciatoventris	83	F	114.0	164.2	237.2	85.9	157.3	228.7
Poecilotriccus sylvia	69	F	89.2	135.3	201.7	61.9	125.4	189.0
Ramphocaenus melanurus	5	F/P	3.8	12.3	27.3	0.0	9.7	20.9
Synallaxis albescens	1	C	0.8	5.6	15.6	0.0	2.5	7.5
Thamnophilus atrinucha	93	C	124.1	177.1	251.6	91.9	162.7	233.6
Thamnophilus doliatus	192	C	269.2	369.7	516.5	211.2	345.7	480.2
Todirostrum cinereum	51	C	63.2	97.6	144.3	46.9	89.5	132.2
Tolmomyias sulphurescens	80	F	110.8	162.1	240.4	80.8	157.1	233.3
Troglodytes aedon	26	C	25.6	45.8	74.3	15.7	38.5	61.3

Table 1: Estimates for understory insectivorous birds in the dry forest of the Magdalena Valley Colombia. Estimates are in individuals/100 ha. Det shows the number of detections of each species in the data set. Parker refers to the abundance category in the Parker III *et al.* (1996) database. U= Uncommon, C = Common, F= Fairly Common, F/P = Fairly common but with patchy distribution.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

7 Figures

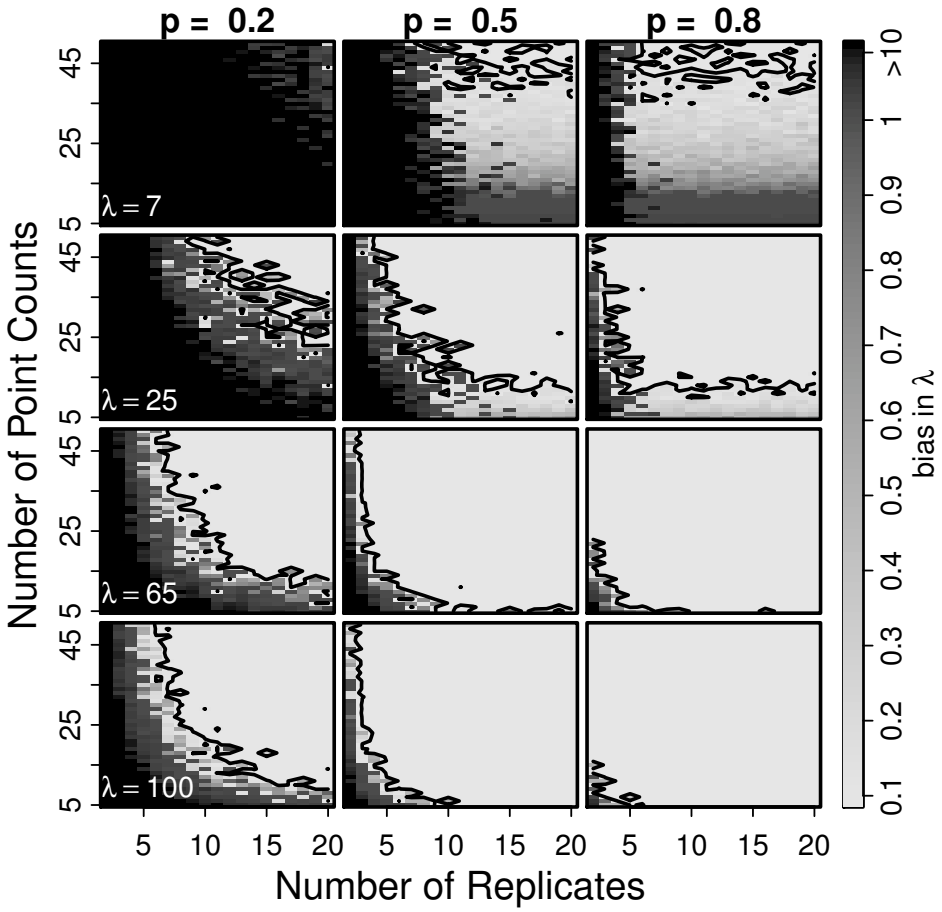

Figure 1: Mean bias in mean number of individuals per 100 ha λ for a range of point counts, number of replicates, and true parameter values to for mid low and high abundances and detection probabilities ($\lambda = 7, 25, 65, 100$ and $p = 0.2, 0.5, 0.8$). The grayscale in each panel represent the bias from low (light gray) to high (black). The color scale is presented in the right. We selected a threshold for acceptable bias in estimation of abundance of 0.1 which isocline is presented as a black line in each of the panels. The results for the entire set of simulations are presented in a similar figure in appendix A

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

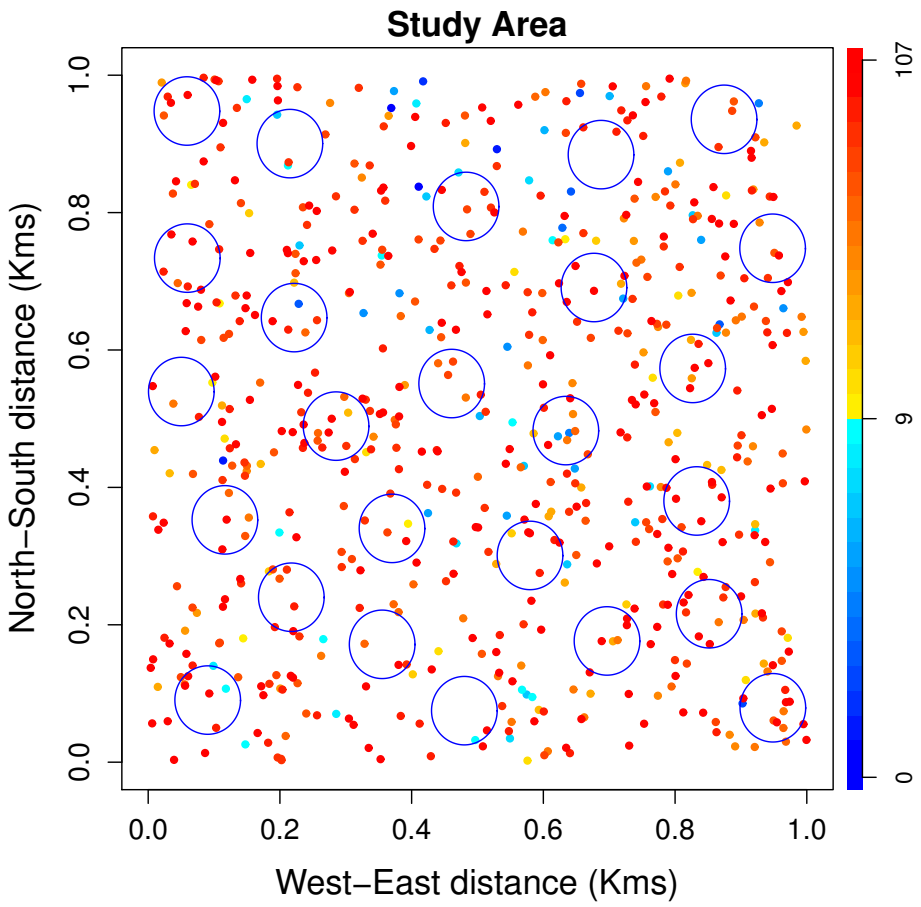

Figure 2: Graphic representation of the sampling design used to simulate the 500 count datasets of a community consisting of 27 species. We assumed the plot 20 be 100 ha ($1\,km^2$) and circular sampling point to be of 0.78 ha ($\sim 0.008\,km^2$). We show the true abundances in the plot represented by colors in the scale bar

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

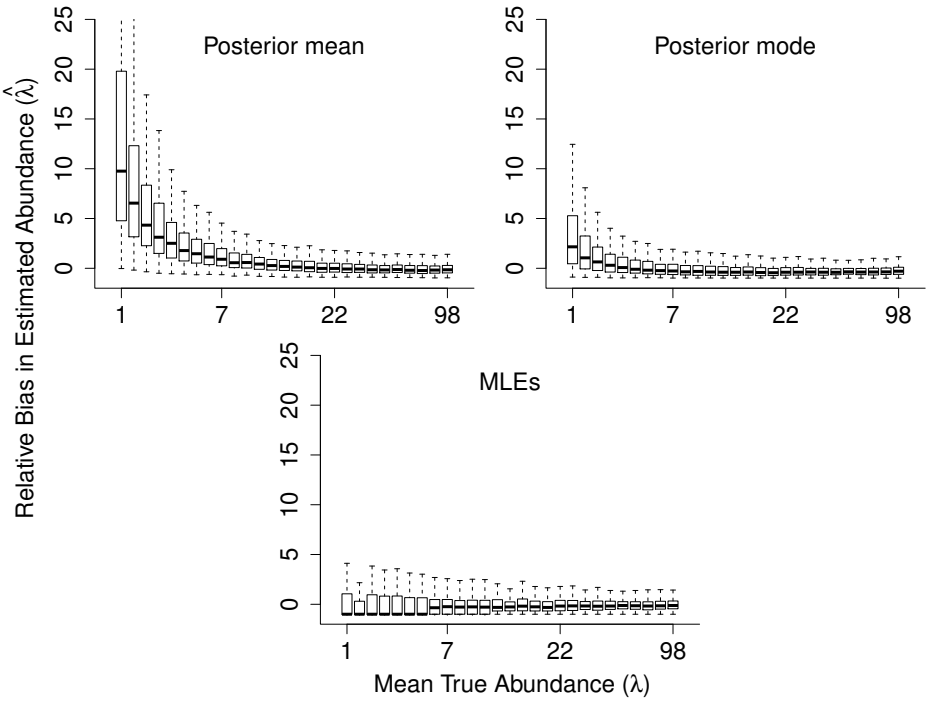

Figure 3: Relative bias in the estimated value of λ ((Estimate-True)/True)) for both the Beta and Normal N-mixture model for 500 simulations of count data, for a community consisting of 27 species. We show the boxplots of the 500 posterior means and modes for the Normal model and the 500 Maximum Likelihood Estimates (MLEs) for the Beta model based on the same simulated data sets. The mean true abundances for each of the 27 species varied from 1 to 98 individuals/100 ha. Because there are 27 true abundances in the community the figure shows one boxplot for each species in the community.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

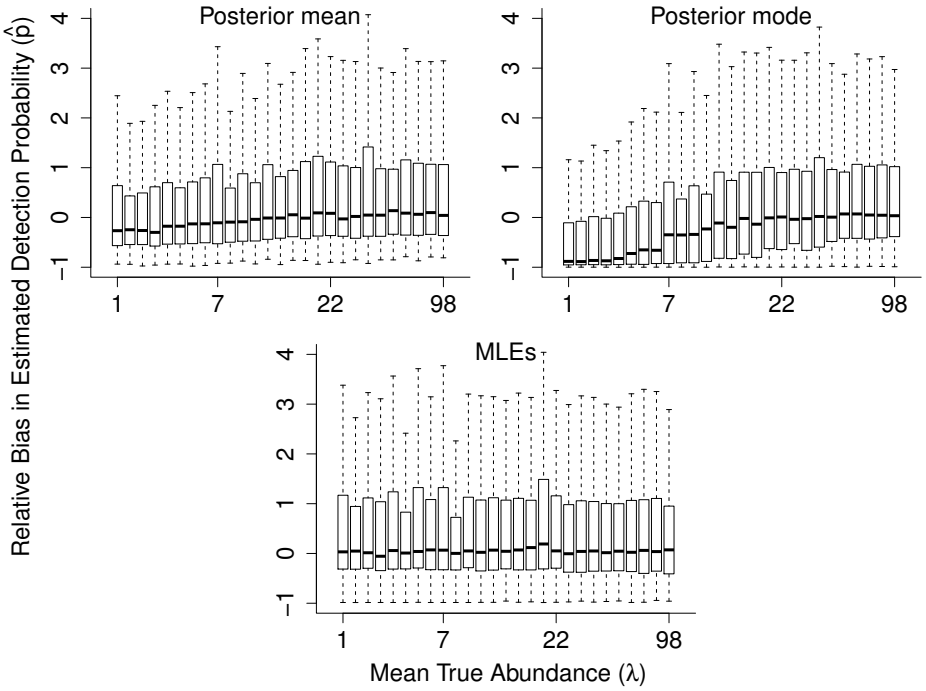

Figure 4: Relative bias in the estimated value of p ((Estimate-True)/True) as a function of the true abundance for both the Beta and Normal N-mixture model for 500 simulations of count data, for a community consisting of 27 species. We show the boxplots of the 500 posterior means and modes for the Normal model and the 500 Maximum Likelihood Estimates (MLEs) for the Beta model based on the same simulated data sets. The mean true abundances for each of the 27 species varies from about 1 to 98 individuals/100 ha. Because there are 27 true abundances in the community the figure shows one boxplot for each species in the community.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

A Supplementary Figures

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

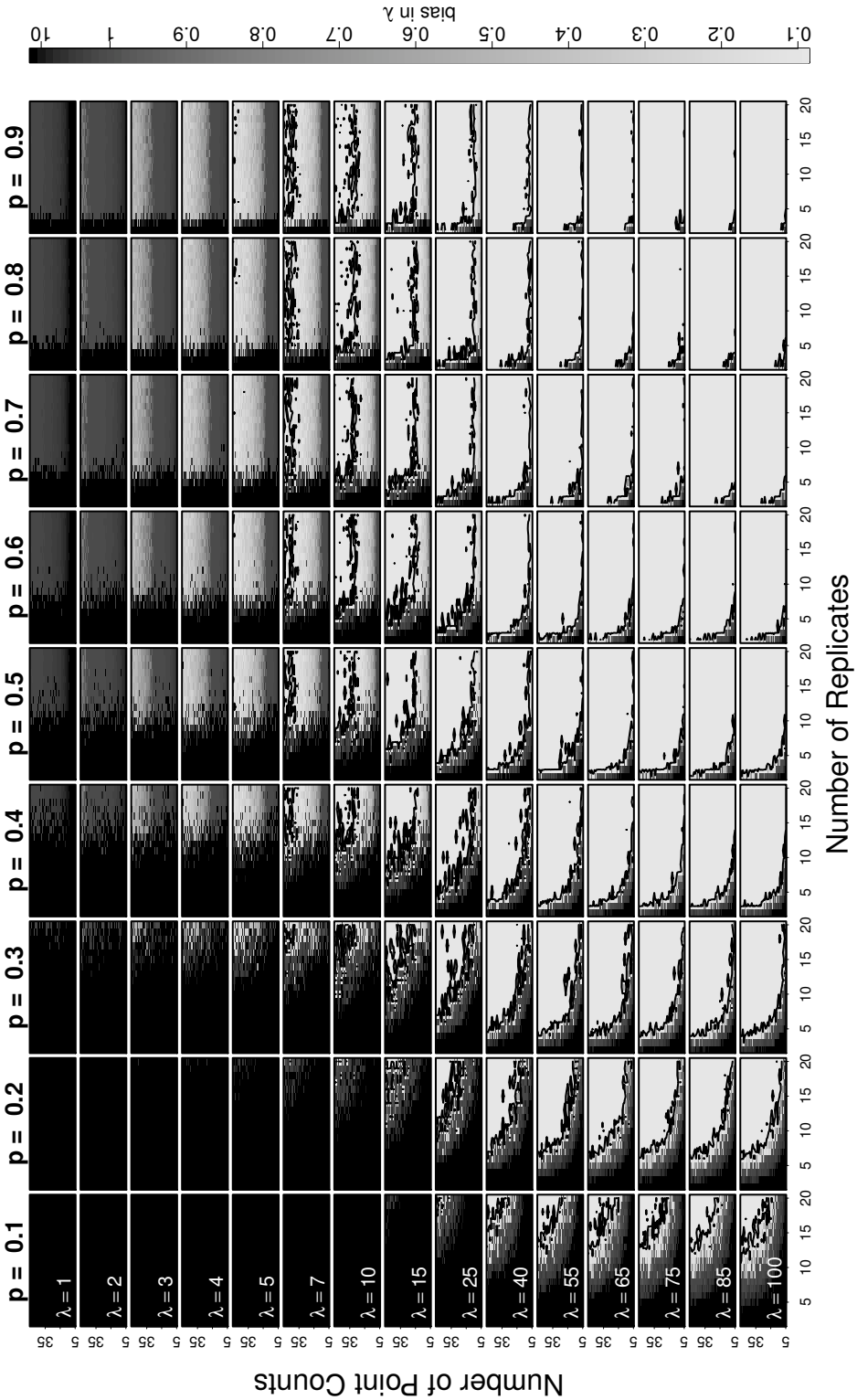

Figure A1: Mean bias in mean number of individuals per 100 ha λ for range of point counts, number of replicates, and true parameter values to for low, mid and high abundances and detection probabilities ($\lambda = 7, 25, 65, 100$ and $p = 0.2, 0.5, 0.8$). The grayscale in each panel represent the bias from low (light gray) to high (black). The color scale is presented in the right. We selected a threshold for acceptable bias in estimation of abundance of 0.1, which is the isocline presented as a black line in each of the panels.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

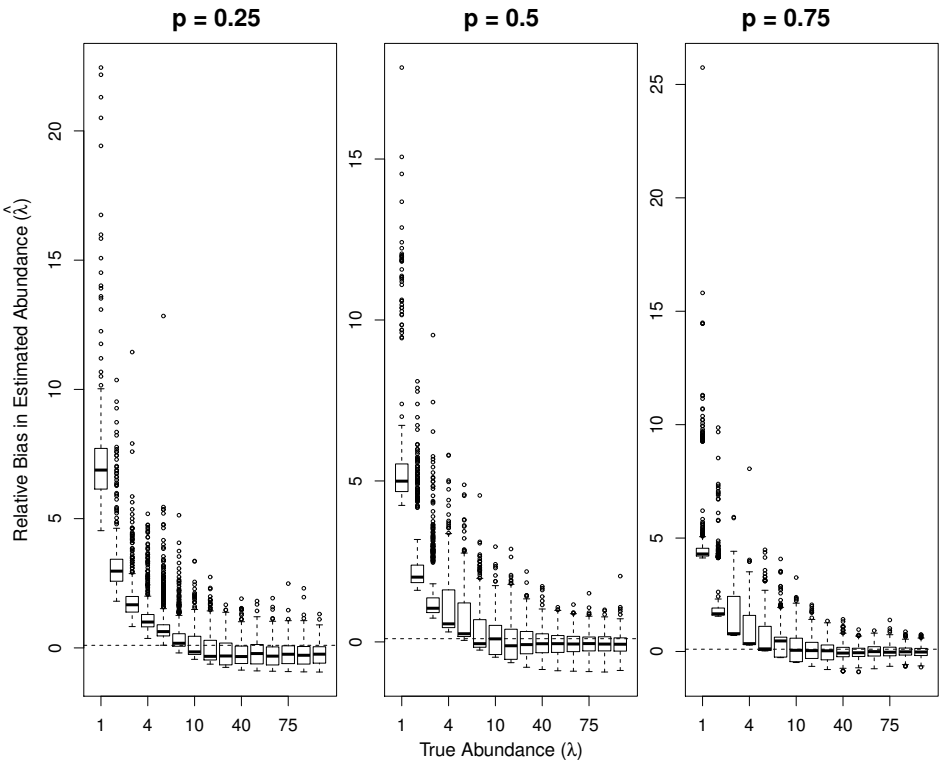

Figure A2: Boxplot showing the distribution of $\hat{\lambda}$ using Beta N-mixture model, showing the location of the true value of λ.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

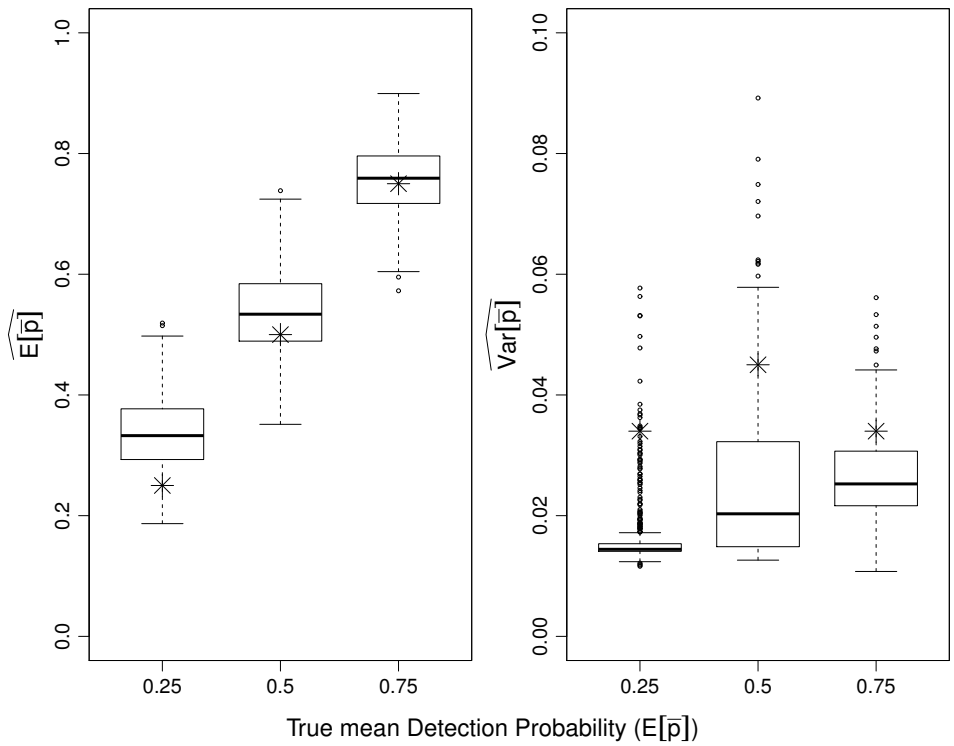

Figure A3: Boxplots showing the distribution of $\widehat{E[\overline{p}]}$ and $\widehat{Var[\overline{p}]}$ as a function of the true mean detection probability $E[\overline{p}]$ with which data was simulated.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

B Bird sampling in the dry forests of the Magdalena Valley

Each point count was replicated three times from January 2013 to July 2014. From this data set, we selected the understory insectivore species that forage over foliage (Karr *et al.*, 1990; Parker III *et al.*, 1996) to meet the requirement of the Beta N-mixture model of correlated detection probabilities among species. In total, we estimated the abundance of 26 species using both the Beta and Normal N-mixture models. We are aware that it is likely that the closed population assumption for this data set does not necessarily hold, but it is unlikely that populations of species have changed drastically from one year to another during these years. The point counts were performed in three different forest patches in the upper Magdalena valley in Central Colombia. To maximize the sample size for abundance estimation, we lumped the point counts into a single data set, such that the inferences of species abundances are made for the entire region instead of the particular patch. The three forest patches were separated by less than 150 km and were located within the Magdalena valley dry forest. Because they are in the same habitat type, the structural variables of the forest are similar and thus it is unlikely that the detection probabilities vary among patches as well as the abundance of species, allowing us to lump the data together.

bioRxiv preprint doi: https://doi.org/10.1101/073577; this version posted November 16, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

C R Code

Appendix B contains the source codes necessary for estimating abundance using the Beta and Normal N-mixture models. It is based on bugs specification of the model, R functions for abundance estimation using N-mixture model are also provided in the code. The data to the three steps of the Beta N-mixture validation are separated in different .RData files. The data sets for the 1500 simulations with hi, mid and low $\overline{p}$ are saved in the bias.RData. The 500 data sets simulated under the complicated model used to compare the Beta and Normal N-mixture model along with the λ and p used in each simulation are saved under the comparison.RData. The real count data from the point counts performed in central Colombia are saved in the file real.RData. The entire code is saved in the Gomez_et_al_code.R from which all of the analysis of this paper can be easily replicated. The only step fro which we did not save the simulated data was the bias estimation of the single species N-mixture model because of the large amount of simulations performed. Using the code and function provided however, the reader should be able to reproduce the simulations and the bias estimation.