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

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

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

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

[*]*Correspondence author*

Abstract

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

2. We evaluate the proposed methods using an assumption of 50m fixed radius point count and perform simulations comprising a broad range of sample sizes, true abundances and detectability values. 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. This hierarchical model avoids having to specify one detection probability parameter per species in the targeted community. Parameter estimation is done via Maximum Likelihood using data cloning.

3. We compared our methodology with standard N-mixture models, which we show here are severely biased and highly variable when the true abundances of species in the community are less than seven individuals per 100ha. For more common species, the number of point counts and replicates needed to reduce the bias of N-mixture model estimators estimation is high. The beta N-mixture model proposed here outperforms the traditional N-mixture model thus allowing the estimation of organisms at lower densities and control of the bias in the estimation.

4. We illustrate how our methodology can be used to determine the sample size required to estimate the abundance of organisms. We also give practical advice for researchers seeking to propose reliable sampling designs for single species' studies. When the interest is full communities, our model and estimation methodology can be seen as a practical solution to estimate organism densities from rapid inventories datasets. The statistical inferences with this model can also inform differences in ecology and behavior of species when they violate the assumption of a single distribution of detectabilities.

Keywords: Point Counts, Sample size estimation, Tropical Bird Species, Hierarchical models, Data cloning.

1 Introduction

One of the most common complications that ecologists face when estimating abundances of mobile organisms is that individuals and species differ in their detection probability. Such differences results in the under or overestimation of real abundance when the detection probability is ignored (MacKenzie *et al.*, 2002; Martin *et al.*, 2005; Royle & Dorazio, 2008). To date, quantitative ecologists have proposed many statistical methods to estimate the detection probabilities and correct the observed individual counts to estimate either density or abundance (Denes *et al.*, 2015).

N-mixture models are a family of hierarchical models 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). Because N is not known, it is considered to be a latent variable that is a product of some discrete distribution, such as the Poisson probability law. Inferences about the abundance of the species of interest therefore rely on estimating the detection probability and the underlying parameter of the distribution giving rise to N (Royle, 2004). Although their use has been widely advocated, very few examples exist of the use of N-mixture models for estimation of the abundance of neotropical bird species. Most of the density estimation for bird populations in the Neotropics comes from sampling of 100-ha plots using intensive field methods such as spot mapping or repeated mist-netting (Terborgh *et al.*, 1990; Thiollay, 1994; Robinson *et al.*, 2000; Blake, 2007). A probable reason for the sparse use of N-mixture models to estimate densities of neotropical birds is the well documented species abundance distributions of tropical organisms, which have a long tail to the right with very few abundant species and many rare ones (Hubbell, 2001). This is reflected in Parker III *et al.* (1996) database, in which they consider a bird to be common in the Neotropics if its population abundance is higher than 15 ind/100 ha. This high proportion of rare species in the overall community makes it difficult to obtain enough detections during field censuses for appropriate estimation of both abundance and detection probability for many, if not the majority of neotropical bird species.

To ensure independence of point counts used to estimate the abundance of neotropical bird species, ornithologists have suggested that points must be at least 200 m apart and the radius of the point count cannot be larger than 50 m (Ralph *et al.*, 1993, 1995; Matsuoka *et al.*, 2014). If the goal is to estimate the abundance of all species on a 100 ha plot (minimum area suggested to correctly describe a lowland local bird community in the neotropics; Terborgh *et al.*, 1990), considering the restrictions described above, the maximum number of points that fit in a 100 ha plot is 36 points of 0.78 ha each, based on a radius of 50m. Because of the excess of rarity in tropical birds, the majority of species will have abundances below 15 ind/100 ha. Assuming that individuals are homogeneously distributed across the plot, and hence their counts Poisson distributed (Pielou, 1969), the expected number of individuals in each point count is $\lambda \approx 0.12 \frac{individuals}{point\,count}$. This value of λ is well below the $\lambda = 2$ used to evaluate the performance of the N-mixture model (Royle, 2004). Below this value, it is not known how the model estimators perform, even though $\lambda < 2$ may be a very common scenario in neotropical bird communities. In addition, several neotropical species are known to be secretive and therefore have low detection probabilities, which imposes even stronger challenges for estimating their abundance. Our first objective in this study was to determine the minimum sample size required to reliably estimate the abundance of neotropical bird species using N-mixture models, given a desired level of precision (say, 10%). We believe that this objective will be particularly useful for population ecologists whose goal is to obtain a rough estimate of the density of a species without having to use one of the most field-intensive bird counting methods known as spot-mapping (*e.g.* Terborgh *et al.*, 1990).

A secondary goal of this paper is to develop a method to estimate the abundance of all of the species present in a community, in order to infer mechanisms driving species abundance distributions, which is an important issue in the current debates over the Unified theory (Hubbell, 2001; McGill *et al.*, 2007). While performing point counts, an observer can easily count all of the individual birds in the area for a particular amount of time irrespective of the identity of the species. Thus, the actual data will have information about all of the species present in the area and an approximation of their abundance. Because of their behavior, foraging strategy and evolutionary relationships, some, if not most of the species in the community can have very similar detection probabilities. Such similarities in detectabilities justify our approach of using the counts of the species in the targeted community to increase the information available on abundance corrected by detection probability. Thus, we expanded the N-mixture models to a scenario in which we used information from multiple species to estimate the parameters of a detection probability distribution of a set of species, and used such probability distribution to estimate the expected abundance per unit area of each of the species in the set.

1.1 The Model

In the following section, after summarizing the widely used N-mixture models, we develop a parsimonious, multi-species model extension that allows a more accurate estimation of the abundance of rare species. The essential contribution of our approach is the use of information from the counts of ecologically similar species to improve the estimation of both detectability and abundance.

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 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 a 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 (sum, in this case) 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.*,

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

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. Writing the likelihood in this way directly implies that in order to estimate the abundance of all the species present in a community, one would need to estimate $2 \times S$ parameters (S mean number of individuals λ_s plus S detection probabilities p_s). 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. These community-wide detection probabilities, for example, 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 $g(p_s;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_s^{\alpha-1}(1-p_s)^{\beta-1}$. The overall likelihood function now integrates over all the realizations of the community-wide detection probabilities P_s:

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

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

The difference of the former and the latter forms of the N-mixture model is that in the latter you need $S+2$ parameters to estimate the abundance of the full community instead of $2 \times S$. In large communities, this might be a significant decrease of parameters. The usefulness of specifying the likelihood is that in the case in which many species are rare, we can use the information on the abundant species to estimate the 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 detectabiities.

1.2 Maximum Likelihood Estimation

One drawback of the beta-N-mixture model is its computational complexity, which imposes a substantial numerical challenge for Maximum Likelihood (ML) estimation. To date, many numerical approximations for obtaining the Maximum Likelihood Estimates (MLEs) for hierarchical models have been proposed (de Valpine, 2012). Of these, the so-called "Data Cloning" methodology has proven to be a reliable approach to not only obtain the MLEs for these types of models, but also for hypothesis testing and model selection, as well as unequivocally measuring the estimability of parameters (Lele *et al.*, 2010; Ponciano *et al.*, 2009). 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 there is no need to find the exact or numerical solution to the likelihood function of the hierarchical model in order to find the MLE. Instead, one only needs to compute means and variances of certain posterior distributions.

Data Cloning proceeds by performing a typical Bayesian analysis on a dataset 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 by pure happenstance, one had observed k identical copies of the data set at hand. 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 $\underline{\lambda}$, α, and β, the variance covariance matrix of the posterior distribution converges to $\frac{1}{k}$ times the inverse of the observed Fisher's information matrix. In this way, 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 some influence over the posterior distribution, in data cloning the choice of the prior distribution does not determine the resulting estimates because these are the MLEs. In our case, the hierarchical model is of the form

$$\mathbf{Y} \sim \text{Binomial}\,(\underline{\mathbf{N}}, \mathbf{P}) = f(y|\underline{\mathbf{N}} = n, \mathbf{P} = p) \quad \text{(Observation model)},$$
$$\underline{\mathbf{N}} \sim \text{Pois}\,(\underline{\lambda}) = g(\underline{\mathbf{N}}; \underline{\lambda}) \quad \text{(Process model)},$$
$$\mathbf{P} \sim \text{Beta}\,(\alpha, \beta) = h(\mathbf{P}; \alpha, \beta) \quad \text{(Process model)}.$$

$\underline{\mathbf{N}}$ and $\mathbf{P}$ are unobserved quantities or latent variables which are products of a stochastic process given by the Poisson and Beta distributions respectively. Furthermore, the parameters left to be estimated (i.e., $\underline{\lambda}, \alpha, \beta$) are seen as random variables themselves that have a posterior distribution $\pi(\underline{\lambda}, \alpha, \beta|\mathbf{Y})$. A typical Bayesian approach would sample from the following posterior distribution:

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

where $\pi(\underline{\lambda}, \alpha, \beta)$ is the joint prior of the model parameters. This approach would yield many samples of the vector $(\underline{\lambda}, \alpha, \beta, \underline{\mathbf{N}}, \mathbf{P})$ and in order to sample from the marginal posterior $\pi(\underline{\lambda}, \alpha, \beta|\mathbf{Y})$ one only needs to look at the samples of the subset of parameters $(\underline{\lambda}, \alpha, \beta)$. The data cloning approach proceeds similarly, except one needs to sample from the following posterior distribution:

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

The notation $^{(k)}$ on the left side of this equation does not denote an exponent and is only there to denote 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 $\underline{\lambda}, \alpha, \beta$ are then simply obtained as the empirical average of the posterior distribution

$\pi(\underline{\lambda}, \alpha, \beta|\mathbf{Y})^{(k)}$ and the variance of the estimates are given by $\frac{1}{k}$ times the variance of this posterior distribution. For a detailed example that illustrates the calculations with posterior and cloned posterior distributions that are analytical and tractable, and where the MLEs can be easily computed, we refer the reader to Ponciano *et al.* (2012).

1.3 Scenarios In Which p_s Seem To Be Correlated Among Neotropical Bird Species

Several scenarios can arise in which p_s seem to be correlated among species. It is widely known that the probability of detecting diurnal species such as birds that defend territories by singing is highest at or right after dawn and decreases with time of day (Blake, 1992). Also, different types of forests have differences in structural characteristics that allow or hinder the detection of all of the individuals available. Thus, species sharing a microhabitat or even inhabiting a particular ecosystem should have similar detection probabilities. Operationally, this would amount to specifying a detection probability distribution that depends on variables such as time of the day, or forest structure indices or characteristics. Another natural phenomenon dictating the form of the detection probabilities is the ubiquitous foraging behavior and phenomenon in the Neotropics known as “mixed-species flocking” (Munn & Terborgh, 1979). These flocks are formed by individuals of different species that forage in groups, with each species segregating into their own forest micro-habitat yet moving together through the jungle. Thus the overall detection of the species in the flock appears as correlated because once you detect one species, you are likely to detect the rest of the species within the flock. Finally, several foraging behaviors and vocal activity patterns make species particularly easy or difficult to detect. Species that forage using a sit and wait behavior are usually much more difficult to detect than species that forage by gleaning on leaves actively searching for food, although there are some species whose high vocalization rates make them easier to detect irrespective of their foraging guild. Thus, sit and wait foragers may be either easy or very difficult to detect (e.g. *Monasa* vs. *Malacoptila*, puffbirds).

2 Methods

2.1 Sample Size Estimation for Neotropical Birds

To determine the minimum sample size required for accurate estimation of the abundance of neotropical species, we used a series of simulations in which we varied the number of points (r), visits to points (t), mean number of individuals in each point (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 for each of the 170 datasets and each of the parameter combinations. 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 used to simulate the data. Finally, we retained the mean bias for each combination of the model parameters. We could not retain the full distribution of the bias because of the large number of simulations performed (10,935,000). We considered an acceptable bias to be lower than 0.1, which 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 B.

2.2 The Beta N-mixture Model

Because 100-ha plots have become the standard for estimating abundances of neotropical birds (Terborgh *et al.*, 1990), we developed our example of the use of the beta N-mixture model using a sampling scenario in 100-ha plot. Assuming that a conservative distance between point counts required for the points to be independent is 200 m we selected the maximum number of points that fit in a square 100 ha plot given this requirement. For this example we used 36 point counts and 5 visits, which is reasonable enough to still be part of a rapid inventory but also has enough information to estimate the abundances of rare species. We simulated 1000 data sets that consisted of 15 species with the same λ values described in the previous simulation and three sets of parameters of the beta distribution. The sets of parameters where $\alpha = 10, 27, 30$ and $\beta = 30, 27, 10$, which account for scenarios of low, mid and high detection probability with the same variance ($\mathrm{E}[p] = 0.25, 0.5, 0.75$;$\mathrm{Var}[p] = 0.004$).

For each of the simulated data sets we estimated λ and p under the N-mixture model and λ, α and β under the beta N-mixture model. Then, we computed the bias in λ in the same way as presented above. We performed maximum likelihood estimation of the parameters under the N-mixture model by optimizing equation 1, with the optim function in R using the Nelder-Mead algorithm. To estimate the parameters under the Beta N-mixture model, we also used maximum Likelihood estimation but using Data Cloning (Lele *et al.*, 2007). We used the rjags (Plummer, 2014) interface for R to build the models and run the analysis with 2 chains, with 20000 iterations in each chain and retained the parameter values every 100 generations after a burn-in period of 1000 generations.

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. 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 N-mixture and Beta N-mixture models and compared their fit using AIC following the protocol presented in Ponciano *et al.* (2009). 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 ecoregion (Olson *et al.*, 2001). Because they are in the same ecoregion, 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. Maximum likelihood estimation for the N-mixture and beta-mixture models where performed in the same way as described in the previous section. The R code and jags models used for model selection using AIC are presented in the Appendix B

3 Results

3.1 Sample Size Estimation for Neotropical Birds

We found that the required minimum sample size needed to accurately estimate the abundance of neotropical bird species 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 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 The Beta N-mixture Model

The Beta n-mixture model performed better than the regular N-mixture model for the data simulated. In both cases, as λ increased, the median of the distribution of $\hat{\lambda}$ converged to the true value of λ (Figure 2). Such results allow us to conclude that using the regular N-mixture model, the minimum abundance that the model is able to estimate is around 10 individuals/100 ha, with a sample size of 36 point counts replicated each five times (Figure 2). The use of multi-species information to estimate the abundance of single species had mainly two results: 1) it decreased the variance in the distribution of $\hat{\lambda}$ and 2) it improved the ability of the model to estimate lower densities with similar sample sizes. Using the beta N-mixture model, the minimum λ that the model is able to estimate is around 4 individuals/100 ha and possibly 3 individuals/100 ha.

Even though the $\hat{\lambda}$ distribution has high variance, many outliers and, it tends to overestimate the density of rare species (Figure 2, A3, A2), it maintains the structure of the species abundance distribution (Figure 3). This is especially true when the mean detection probability is at least 0.5 (Figure 3). This is not true for the N-mixture model, which predicts many more mid-range species and overestimates by far the abundance of abundant species (Figure 3).

The beta N-mixture model also performs well in estimating the distribution of the community's detection probability (Figure 4). The distribution of E[p] for the simulations is almost centered in the true value of p. There is a slight underestimation of p when $p = 0.25$ (Figure 4). However, the model overestimates Var[p], but it estimates the variance to be similar across the different types of simulations, which reflects the reality with which the data were simulated (Figure 4).

3.3 Example Using Real Data

We present the estimates of $\hat{\lambda}$ for both models in Table 1, as well as $\hat{p}$ for each species estimated using the N-mixture model. For the beta N-mixture case, the model estimates that the mean detection probability of an insectivorous bird in the upper Magdalena Valley was of $E[p] = 0.2$ with $Var[p] = 0.009$ ($\alpha = 3.15$,$\beta = 12.7$). Assuming model 1 to be the N-mixture and model 2 to be the beta N-mixture, following Ponciano *et al.* (2009), the difference in AIC between the model was extremely high giving very strong support in favor of the beta model ($\Delta AIC = 14725.04$).

4 Discussion

We found that N-mixture models require high sample sizes in both the number of point counts and the number of replicates of each point to estimate accurately the abundance of tropical birds. An interesting result is that the models are unable to accurately estimate rare species with less than 7 ind/ha and a low detection probability (< 0.5), at least with samples of 50 points and 20 replicates. Our model, uses the information collected for a set of species to estimate the abundance of rare species. The beta model is particularly useful for rare species that are detected few times. In this case, even with high sample sizes, the N-mixture model estimates a very high λ and very low p (our study and Sólymos & Lele, 2016) and in some cases resulting in non-identifiablity of λ and p (Sólymos & Lele, 2016). This problem is solved with the beta model, in which the detection probability of a rare species can be thought of as coming from the same probability distribution than its ecologically similar species. In this way, each species has its own detection probability but is linked through an underlying process that can vary from one set of species to another or from one habitat to another.

N-mixture models can be used if the sampled area is increased to raise λ to around 2 individuals per unit area. Because this simulation was performed for tropical forests, we simulated point counts with a 50-m radius. This distance has been proposed to meet the assumption that the detection probability is homogeneous across the whole sampling area and to increase the detection probability of species within it . Other methods have relaxed this assumption (e.g. distance sampling Buckland *et al.*, 1993), and recent studies have even suggested methods to perform estimation in a multi-species fashion (Dorazio & Royle, 2005; Dorazio *et al.*, 2015; Sollmann *et al.*, 2016; Yamaura *et al.*, 2011). Our objective, however, was to evaluate the N-mixture model for fixed-radius point counts as applied to neotropical forests, a method that is very commonly used (*e.g.* Blake, 1992). One solution to get accurate estimates by increasing the area is to discourage the use of point counts in favor of fixed width line transects. Nonetheless, increasing the area would require a decrease in the number of sampling units because the objective is to sample a bird community in a 100-ha plot. This might not be the most favorable solution since our simulations suggest that the increase in sampling units decreases the bias faster than the increase in replication of the sampling units (Yamaura *et al.*, 2016; Figure 1; Figure A1). Alternatively, novel statistical methods allow accurate estimation of abundance using point counts with no replication (Sólymos *et al.*, 2012; Sólymos & Lele, 2016). In this case, the replication for detectability estimation is replaced with covariances in the detection and abundance process. Even though abundance and detection covariates are commonly accounted for in most point count studies, at small scales (*e.g.* 100 ha plots), abundance covariates might be difficult identify because of habitat homogeneity. Such approaches might be more useful in large scale studies investigating the variation in abundance of species across the landscape.

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 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 can be two fold: increase the sample size to a very high number of points and many replicates (>50 point counts and >20 replicates) or to keep the sampling design of a 36 points and 5 replicates and use our proposed model of the multi-species sampling. We are aware that in the case of estimating the abundance of rare species, the maximum acceptable bias that we selected of 10% might be too conservative. In such cases, the acceptable bias can be increased to 100 or 200% with little risk. Even though we selected the 10% bias across abundances for simplicity, we present depict full results in appendix A so that researchers are able to make decisions about the sample size required with the desirable amount of bias in the estimates.

We show that by using the information of other more abundant species, the model is able to predict correctly the abundance of rare species with $\lambda = 4$ and better approximate the abundance of species with $\lambda < 4$. By restricting the detection probability of target species to arise from the same distribution than the one of other species, the beta N-mixture model allocates into estimation of λ the same amount of information with considerably less parameters (Figure A2). While in the N-mixture model, allowing p to vary freely for every species can result in strong estimability problems between p and λ (i.e. models with high p and λ have similar likelihood to models with low p and high λ), our beta model appears to be much more reliable and accurate (mean detection probability of 0.2 and 2.5% and 97.5% quantiles of 0.04 and 0.41 respectively). Such correction, allows the estimate of λ, even for rare species, to be much closer to the true value (Figure 2; Figure A3 and A4). Even when the mean detection probability is low ($p \sim 0.25$), the beta model tends to over estimate the abundance of the entire set of species, but maintains the structure of the species abundance distribution (Figure 3 and Figure A3).

Our model is different from other approaches to multi-species abundance and occurrence estimation (Dorazio & Royle, 2005; Yamaura *et al.*, 2011; Dorazio *et al.*, 2015; Sollmann *et al.*, 2016) because we do not assume that detection probabilities of species are unrelated quantities. The assumption of a common detection probability allow us to make inferences about the abundance of rare species that are usually discarded when estimating the composition of communities. Yamaura *et al.* (2011) made similar assumptions in which species respond as a community to changes in environmental covariates. However, we assume that the detections of species are the product of a stochastic process, instead of deterministically predicting the detection probability for each species as a function of some other covariates. An important consideration of our approach is that the grouping of species used to estimate the distribution of detection probabilities has to be carefully justified and informed by their ecology and vocal behavior. In our (field) experience makes little sense to assume that species that are extremely different in their ecologies have detection probabilities drawn from the same probability distribution.

Martin *et al.* (2011) and Dorazio *et al.* (2013) used a similar approach to ours, but for single-species abundance estimation. In their models, they assumed correlated behavior among the individuals of the same species and variation across sites, adding an additional layer of hierarchy to the traditional N-mixture models (Royle, 2004). In their model, the binomial distribution is substituted by a beta binomial that assumes that the probability of detecting one individual is slightly different from another, but the result of the same stochastic process. Similarly, in this study we assumed that species have their own detection probability which for species with similar ecologies is drawn from the same probability distribution. However, following the sampling structure of the point counts, our model averages out all the possible realizations of the beta process after taking into account the poisson sampling layer, and thus the likelihood function becomes an intractable integral. After integration, the likelihood matches a multivariate distribution (Sibuya *et al.*, 1964) whose covariance structure arises naturally from the sampling scheme. Fortunately, the Data Cloning algorithm allows us to make Maximum likelihood inferences without having to solve this integral.

Because our model is essentially identical to any N-mixture model, it has the advantage that it can be adapted to any underlying distribution of abundances. Similarly, 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). 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. For example, one can assume that the detection probability distribution is a function of variables such as the ecological guild a bird belongs to or to the microhabitat used for foraging and nesting. 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.

The estimates of the abundance of the understory insectivores of the upper Magdalena Valley show that the difference between the N-mixture and beta-mixture models relies on the estimation of the abundance of rare species. For example, for species with less than five detections, the N-mixture model estimates the abundance to be extremely high (Table 1). Instead, by assuming the detection probability is correlated with the other species in the set, our approach lowers the estimation of the abundance to values closer to densities reported for the same species or similar in other regions (e.g., Karr *et al.*, 1990; Parker III *et al.*, 1996). 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. 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.

Overall, our study can be used as a baseline to determine the number of point counts required to estimate the density of neotropical bird species using N-mixture models. We showed that for many species in neotropical communities, the sample size needed to correctly estimate their density is high and thus we advocate for more field intensive methods such as spot mapping. Probably, in the neotropics, the needed spatial resolution will be larger areas than the standard 100-ha plot because of the very large number of species with fewer than 2 territories/100ha. Distance sampling could also be used as an alternative method, but it has been previously shown that it requires a high number of detections to appropriately estimate the abundance of species (Matsuoka *et al.*, 2014). Such high number of detections might be impossible to achieve, particularly for rare species. We conclude that our method might be a good alternative when sample sizes are low, but information for many other species is available. We expect that for large communities implementing our model using data cloning would be a computer intensive task. Also, in large neotropical communities, the assumption of detectabilities arising from the same probability distribution might not hold for many species. Here, we showed an example in which we estimated the abundance of 26 species of insectivorous birds in the Magdalena Valley, allowing us to demonstrate that for a community of this size, the maximum likelihood estimation with this size of community is feasible. Such approach represents a significant practical improvement for neotropical bird studies. We also note that the larger the community, the easier it will be for the model to estimate the abundance of rare species (Sollmann *et al.*, 2016), but there is a tradeoff with the computational power needed for ML estimation.

Estimation of animal abundances is the ecologists' starting point to confront novel theoretical models and hypotheses with evidence in nature, and this scientists' field has long understood the importance of such task (*e.g.*, Seber, 1986). It is in that sense that we hope that our work is seen as a practical, and easy to use extension of the N-mixture models. Our work shows that the approach that we propose can serve as a platform to design community ecology studies that require, as a starting point, the joint estimation of abundances while taking into account differences in detection probabilities among species.

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 provided useful comments for the development of the model and improvement of the manuscript.

References

Blake, J.G. (1992) Temporal variation in point counts of birds in a lowland wet forest in costa rica. *The Condor*, pp. 265–275.

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.

Buckland, S.T., Anderson, D.R., Burnham, K.P. & Laake, J.L. (1993) *Distance Sampling: estimating abundance of biological populations*. Chapman and Hall, London.

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., Connor, E.F. & Askins, R.A. (2015) Estimating the effects of habitat and biological interactions in an avian community. *PloS one*, **10**, e0135987.

Dorazio, R.M., Martin, J. & Edwards, H.H. (2013) Estimating abundance while accounting for rarity, correlated behavior, and other sources of variation in counts. *Ecology*, **94**, 1472–1478.

Dorazio, R.M. & Royle, J.A. (2005) Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, **100**, 389–398.

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

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.

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.

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.

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.

Matsuoka, S.M., Mahon, C.L., Handel, C.M., Sólymos, P., Bayne, E.M., Fontaine, P.C. & Ralph, C.J. (2014) Reviving common standards in point-count surveys for broad inference across studies. *The Condor*, **116**, 599–608.

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.

Munn, C.A. & Terborgh, J.W. (1979) Multi-species territoriality in neotropical foraging flocks. *The Condor*, **81**, 338–347.

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

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.

Pielou, E.C. (1969) *An introduction to mathematical ecology*. Wiley-Interscience, New York.

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.

Ponciano, J.M., Taper, M.L., Dennis, B. & Lele, S.R. (2009) Hierarchical models in ecology: confidence intervals, hypothesis testing, and model selection using data cloning. *Ecology*, **90**, 356–362.

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

Ralph, C.J., Geupel, G.R., Pyle, P., Martin, T.E. & DeSante, D.F. (1993) Handbook of field methods for monitoring landbirds. *USDA Forest Service/UNL Faculty Publications*, p. 105.

Ralph, C.J., Sauer, J.R. *et al.* (1995) Monitoring bird populations by point counts. *USDA Forest Service General Technical Report PSW-GTR-149*, pp. 161–175.

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.

Seber, G.A. (1986) A review of estimating animal abundance. *Biometrics*, **42**, 267–292.

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

Sollmann, R., Gardner, B., Williams, K.A., Gilbert, A.T. & Veit, R.R. (2016) A hierarchical distance sampling model to estimate abundance and covariate associations of species and communities. *Methods in Ecology and Evolution*, **doi: 10.1111/2041-210X.12518**.

Sólymos, P., Lele, S. & Bayne, E. (2012) Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. *Environmetrics*, **23**, 197–205.

Sólymos, P. & Lele, S.R. (2016) Revisiting resource selection probability functions and single-visit methods: clarification and extensions. *Methods in Ecology and Evolution*, **7**, 196–205.

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.

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.

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.

628 # 6 Tables

		N-mixture		Beta		
Species	Detections	p	λ	λ	lower	upper
Atalotriccus pilaris	83	0.315	119.7	119.8	67.9	171.7
Basileuterus rufifrons	104	0.219	215.4	214.9	104.3	325.5
Campylorhynchus griseus	7	0.311	10.2	10.5	0	22.2
Cantorchilus leucotis	3	0.0004	3832.3	32.5	0	193.0
Cnemotriccus fuscatus	31	0.174	80.9	78.7	8.5	149.0
Contopus cinereus	2	0.004	211.4	14.4	0	69.9
Cymbilaimus lineatus	4	0.0005	3663.8	41.1	0	181.2
Dromococcyx phasianellus	1	0.0005	905.9	6.3	0	39.9
Elaenia flavogaster	67	0.126	241.3	231.4	62.1	400.6
Euscarthmus meloryphus	26	0.265	44.6	44.3	14.9	73.8
Formicivora grisea	172	0.280	279.3	279.5	168.7	390.3
Hemitriccus margaritaceiventer	106	0.408	118.1	118.4	81.8	155.0
Henicorhina leucosticta	28	0.124	102.3	95.5	0	201.3
Hylophilus flavipes	144	0.064	1023.2	829.2	0	2086.4
Leptopogon amaurocephalus	23	0.194	53.8	53.0	9.5	96.6
Myrmeciza longipes	64	0.257	113.4	113.0	55.8	170.2
Myrmotherula pacifica	1	0.001	905.9	5.9	0	32.1
Pheugopedius fasciatoventris	83	0.230	164.0	163.1	77.2	249.0
Poecilotriccus sylvia	69	0.239	131.0	130.0	52.6	207.3
Ramphocaenus melanurus	5	0.206	11.0	11.0	0	29.0
Synallaxis albescens	1	0.0005	905.9	6.4	0	48.7
Thamnophilus atrinucha	93	0.255	165.9	165.5	92.4	238.6
Thamnophilus doliatus	192	0.246	354.5	353.8	186.3	521.2
Todirostrum cinereum	51	0.255	91.1	90.3	44.2	136.4
Tolmomyias sulphurescens	80	0.216	168.4	166.5	75.4	257.7
Troglodytes aedon	26	0.322	36.7	37.2	13.5	60.8

Table 1: Estimates for understory insectivorous birds in the dry forest of the Magdalena Valley Colombia. Estimates are in individuals/100 ha

629 # 7 Figures

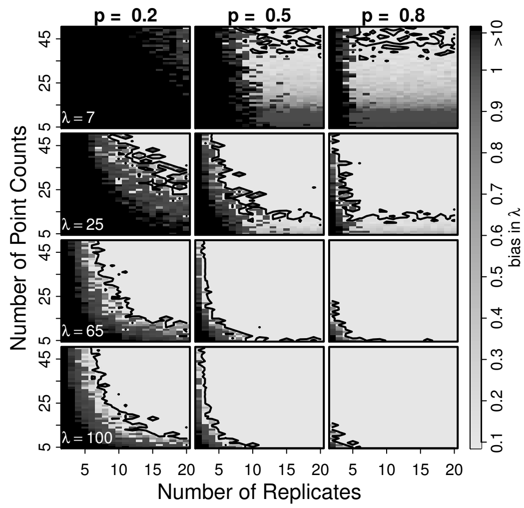

Figure 1: Mean bias in mean number of individuals per 100 ha λ for a range of point counts, number of replicates, and true parameter values to for mid low and high abundances and detection probabilities ($\lambda = 7, 25, 65, 100$ and $p = 0.2, 0.5, 0.8$). The grayscale in each panel represent the bias 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

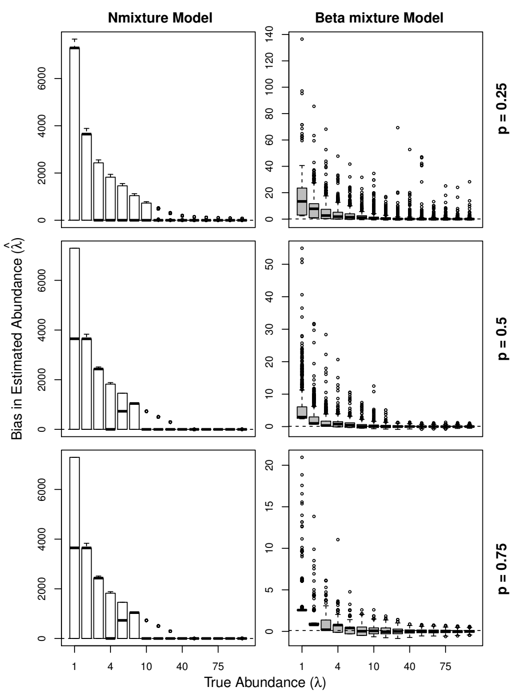

Figure 2: Bias in the estimated value of λ for both the N-mixture and beta N-mixture model for 1000 simulations of data under three different scenarios of low, mid and high detection probabilities and 36 point counts replicated each five times.

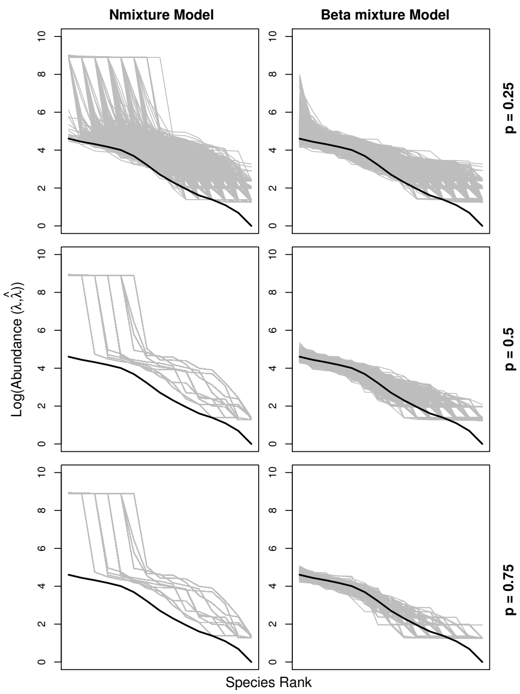

Figure 3: True (black) and each of the estimated species abundance (gray) distributions from the 1000 simulations of data under three different scenarios of low, mid and high detection probabilities.

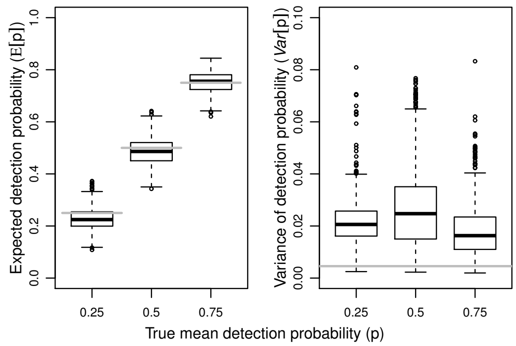

Figure 4: Distribution of Expected (E[p]) and variance (Var[p]) in detection probability across the 1000 simulations performed under scenarios of low, mid and high E[p] and the same variance Var[p].

A Supplementary Figures

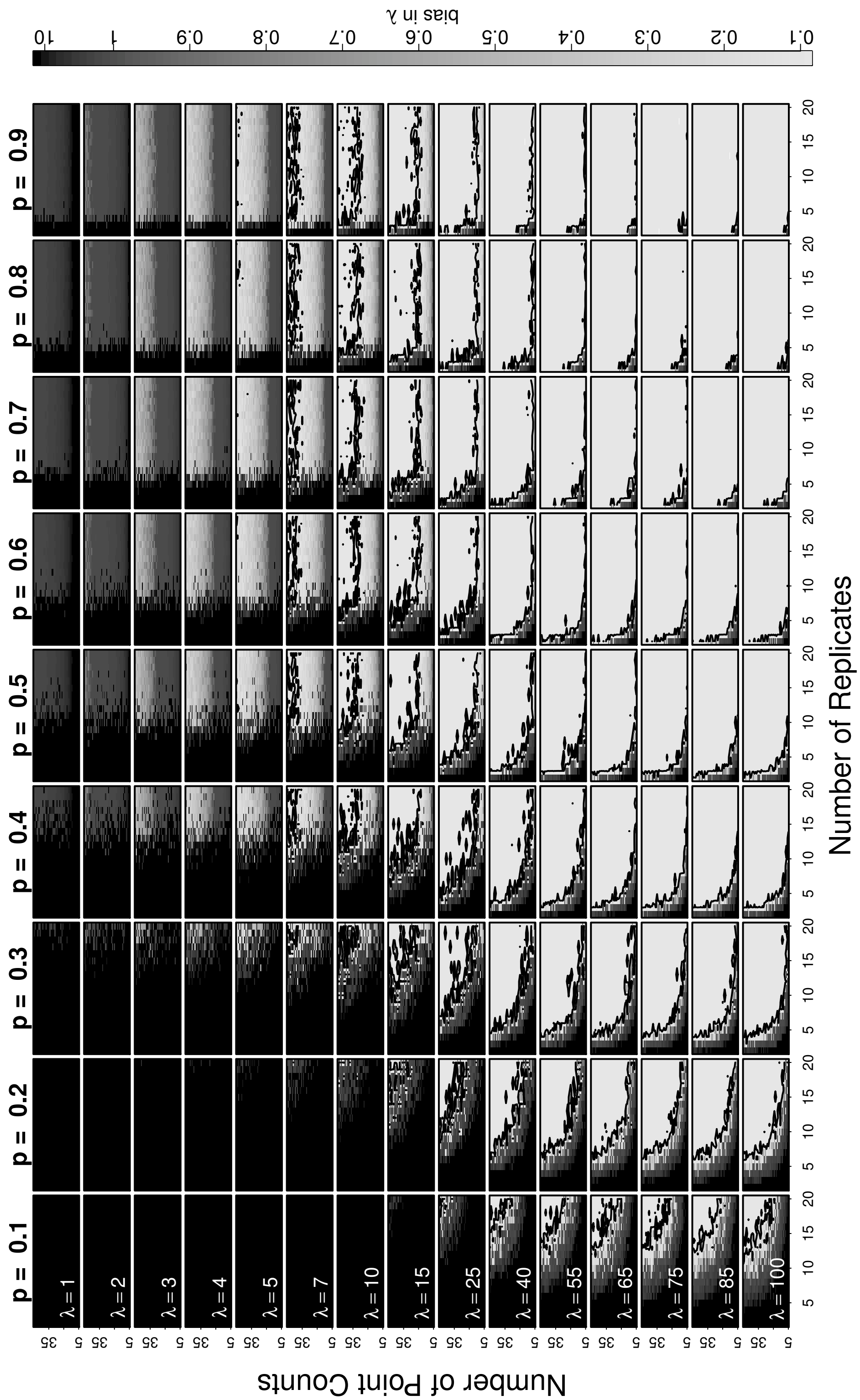

Figure A1: Mean bias in mean number of individuals per 100 ha λ for range of point counts, number of replicates, and true parameter values to for low, mid and high abundances and detection probabilities ($\lambda = 7, 25, 65, 100$ and $p = 0.2, 0.5, 0.8$). The grayscale in each panel represent the bias 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.

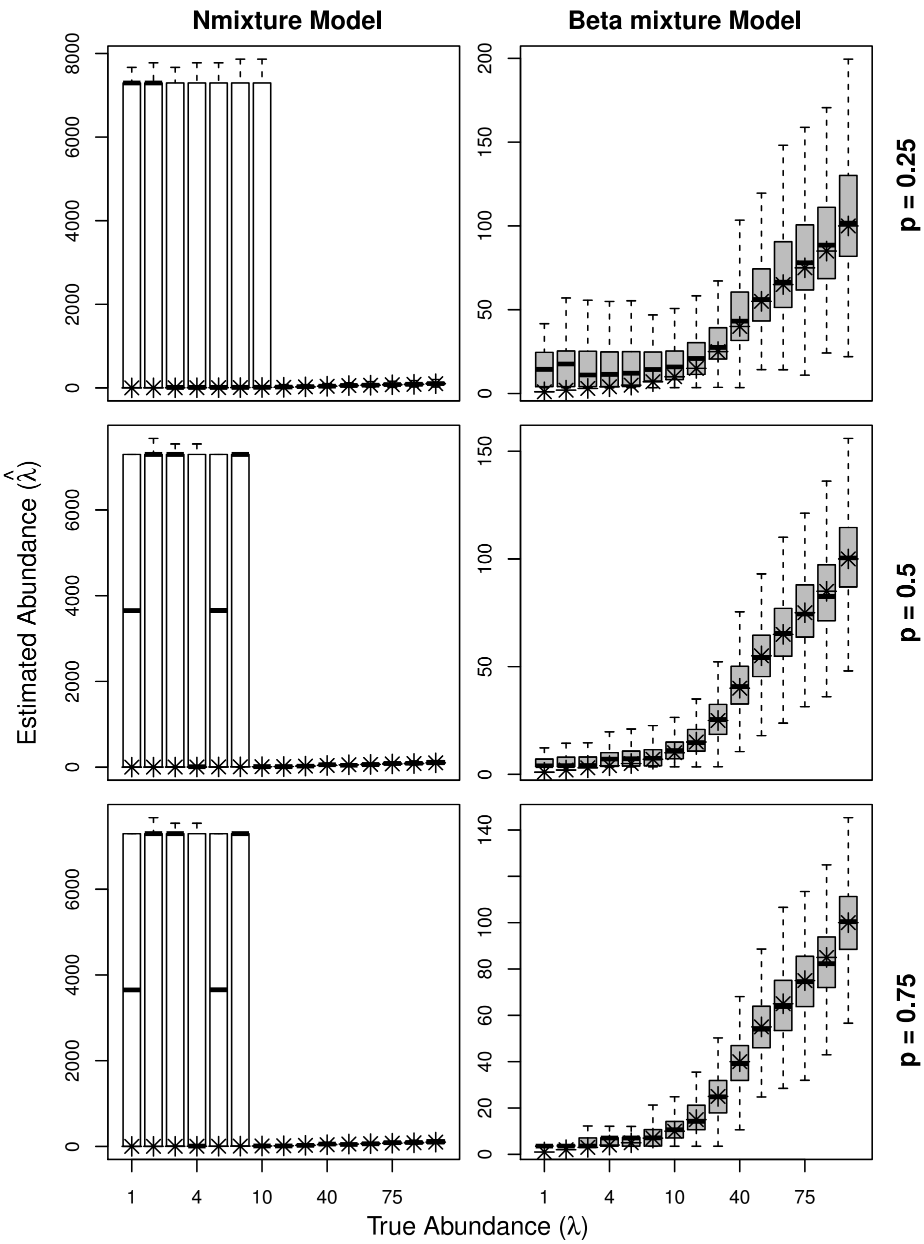

Figure A2: Histogram of estimated detection probabilities based on the N-mixture model estimates of 26 understory insectivorous birds of the dry forest of the Magdalena Valley Colombia. The distribution of the detection probabilites estimated by the beta model is also shown (black dotted line) based on the parameters estimated as $\alpha = 3.15$ and $\beta = 12.7$

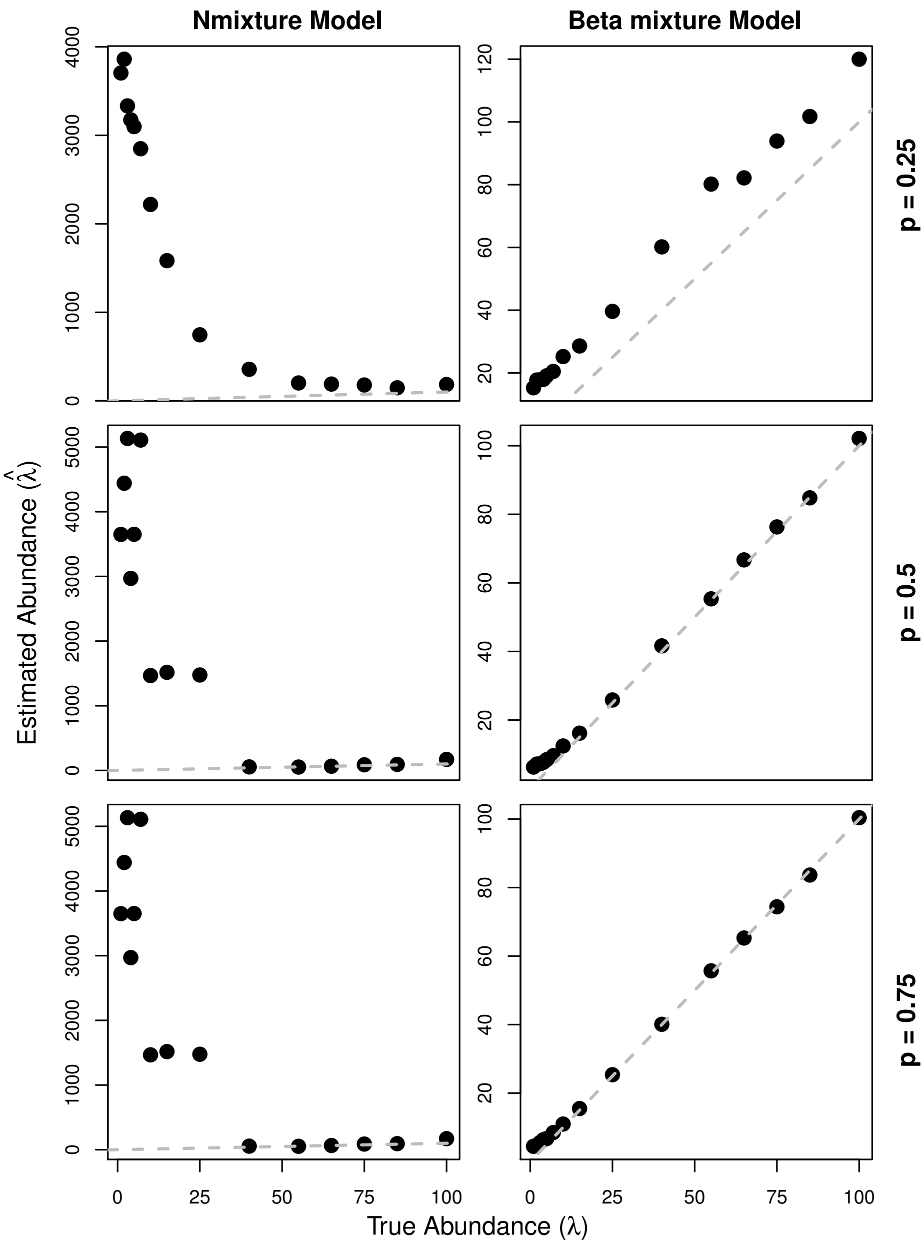

Figure A3: Barplot showing the distribution of $\hat{\lambda}$ using N-mixture and beta N-mixture models, showing the location of the true value of λ. The outliers for the N-mixture and beta N-mixture models have been omitted for clarity.

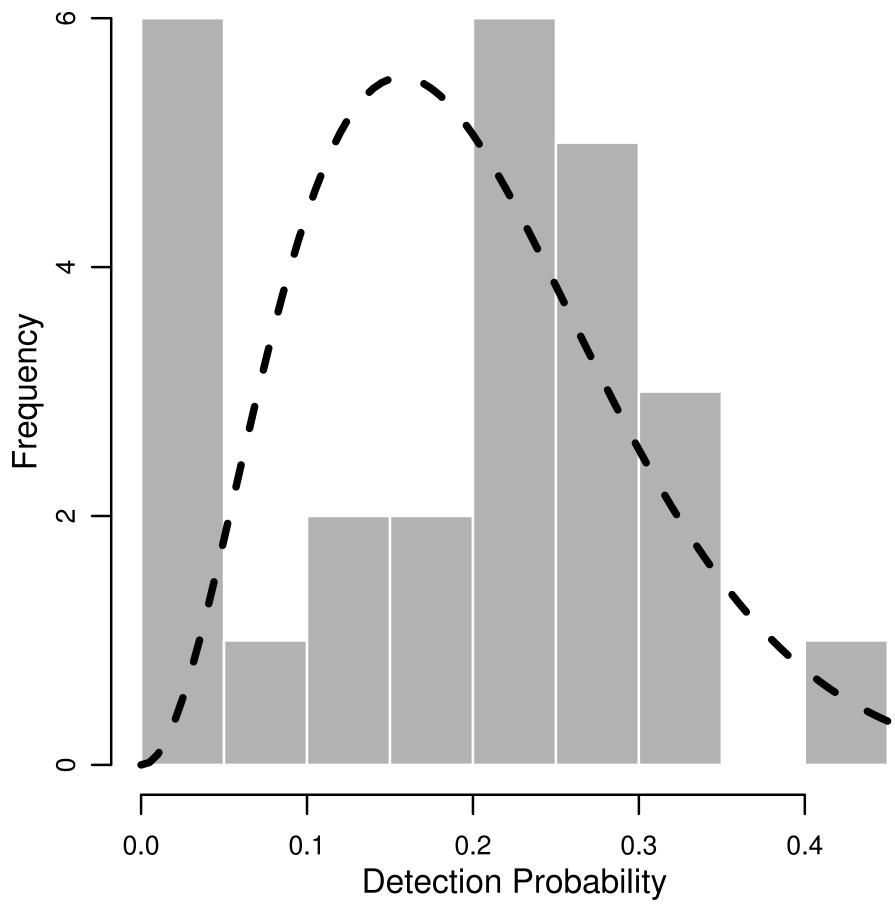

Figure A4: Relationship between the mean value of $\hat{\lambda}$ from the 1000 simulations and the true value of λ. For reference, we show the one-to-one relationship line (gray dotted line).

B R Code

Appendix B contains the source codes necessary for estimating abundance using the Beta N-mixture model. It is based on bugs specification of the model, R functions for abundance estimation using N-mixture model and the R code necessary to reproduce the example using real data. The data have been saved in a separate file named UIFcounts.RData.