

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

Juan Pablo Gomez<sup>1,2,3,\*</sup>, Scott K. Robinson<sup>2</sup>, Jason K. Blackburn<sup>3,4</sup>,  
and José Miguel Ponciano<sup>1,\*</sup>

<sup>1</sup>Department of Biology, University of Florida, Gainesville, Florida

<sup>2</sup>Florida Museum of Natural History, Gainesville, Florida

<sup>3</sup>Spatial Epidemiology and Ecology Research Laboratory, Department  
of Geography, University of Florida, Gainesville Florida

<sup>4</sup>Emerging Pathogens Institute, University of Florida, Gainesville,  
Florida

\* *Correspondence author*

## Abstract

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

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

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

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

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

## 34 1 Introduction

35 Unbiased abundance and occupancy estimates are of paramount value for making in-  
36 ferences about ecological processes and making sound conservation decisions (Hubbell,  
37 2001; Leibold *et al.*, 2004; Margules & Pressey, 2000). To date, quantitative ecologists  
38 have proposed several statistical methods to estimate species' detection probabilities  
39 and use these to correct the occupancy or abundance estimates (Denes *et al.*, 2015).  
40 Our study was motivated by the attempt to use these novel models to estimate the  
41 abundance of rare species in tropical communities. In these communities, it is well-  
42 known that abundance distributions are typically characterized by long right tails  
43 with few abundant species and many rare ones (see Hubbell, 2001). Such high pro-  
44 portion of rare species in the overall community makes it very difficult to obtain  
45 enough detections during field censuses for appropriate estimation of both abundance  
46 and detection probability for many, if not the majority of tropical species. When  
47 we extensively tested via simulations these recent methodologies, we found persistent  
48 bias in estimates of low abundances that corresponded to abundance ranges previ-  
49 ously not dealt with in temperate forest studies yet common in neotropical studies  
50 (see also Yamaura, 2013; Yamaura *et al.*, 2016). As an answer to this problem, in  
51 this study we present an alternative, community-based abundance estimation ap-  
52 proach that markedly improves these estimates. Our method is widely applicable in  
53 communities with similarly abundance patterns.

54 The N-mixture models aim to tackling the problem of the bias in abundance  
55 estimation induced by species differences in detection probabilities (MacKenzie *et al.*,  
56 2002; Martin *et al.*, 2005; Royle & Dorazio, 2008). It uses spatially and temporally  
57 replicated counts in which, the counts of species  $y$  are binomially distributed with  $N$   
58 being the total number of individuals available for detection and  $p$  the probability of  
59 detecting an individual of that species (Royle, 2004). The model is hierarchical be-  
60 cause the abundance  $N$  is assumed to be a latent, random process adopting a discrete

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

75         In recent multi-species abundance models, both abundance and detection prob-  
76 abilities are assumed to be normally distributed random effects governed by the com-  
77 munity’s “hyper parameters” (Iknayan *et al.*, 2014). For that reason they have been  
78 named community abundance models, because they focus in describing the charac-  
79 teristics of the entire community from spatially and temporally replicated counts or  
80 detections (Yamaura *et al.*, 2012, 2011, 2016). The main assumption behind the com-  
81 munity abundance models is that groups of species in the community might share  
82 characteristics that make their abundance and detection probability to be correlated  
83 (Yamaura *et al.*, 2011, 2012, 2016; Sauer & Link, 2002; Barnagaud *et al.*, 2014; Ruiz-  
84 Gutiérrez *et al.*, 2010). These type of abundance community models have been useful  
85 for estimating diversity properties of the species assemblages while accounting for  
86 imperfect detection (Yamaura *et al.*, 2011, 2012).

87         While the assumption of normally distributed logit transformed random effects

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

99 In this study, we: (1) increase the simulation scenarios presented in Yamaura  
100 (2013) to provide a full baseline for the sampling design for ecologists that want to  
101 estimate the abundance of tropical organisms using N-mixture models, (2) propose  
102 and alternative multi-species abundance model that uses a beta distribution for the  
103 random effects of detection probability instead of a normal distribution and (3) pro-  
104 pose a maximum likelihood approach for multi-species abundance estimation using  
105 data cloning (4) compare our alternative multi-species abundance model to one of the  
106 previously proposed ones.

## 107 **1.1 The Model**

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

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

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

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

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

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

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

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

133 where  $y_{sij}$  is a three dimensional array of dimensions  $r \times t \times S$ , and both  $\underline{\lambda} =$   
 134  $\{\lambda_1, \dots, \lambda_S\}$  and  $\underline{p} = \{p_1, \dots, p_S\}$  are vectors of length  $S$ . In what follows, we will  
 135 refer to the  $n_{si}$ 's as the latent, realized abundance and to the mean abundances,  
 136 the  $\lambda_s$ 's simply as the ‘‘abundances’’. To avoid the proliferation of parameters one  
 137 could assume that all the  $p_s$  come from a single probability model that describes the  
 138 community-wide distribution of detection probabilities (Yamaura *et al.*, 2011, 2012,  
 139 2016; Sauer & Link, 2002; Barnagaud *et al.*, 2014; Ruiz-Gutiérrez *et al.*, 2010). These  
 140 community-wide detection probabilities can be modeled with a beta distribution in  
 141 which we let  $P_s \sim \text{Beta}(\alpha, \beta)$ . The probability density function of the random detec-  
 142 tion probabilities is then  $g(p_s; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_s^{\alpha-1} (1-p_s)^{\beta-1}$ .

143 Following (Dorazio *et al.*, 2013), we parameterize the Beta distribution as  
 144  $\text{Beta}(\alpha = \tau\bar{p}, \beta = \tau(1 - \bar{p}))$  such that the parameters are related to biological pro-  
 145 cesses. Here,  $\bar{p}$  is the mean detection probability among species in the community  
 146 and  $\tau$  is a measurement of the similarity in detection probabilities (Dorazio *et al.*,  
 147 2013).

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

$$\begin{aligned} \mathcal{L}(y_{sij}; \underline{\lambda}, \bar{p}, \tau) &= \int_0^1 \prod_{s=1}^S \prod_{i=1}^r \sum_{N_{si}=\max(\mathbf{y}_{si})}^{\infty} \prod_{j=1}^t \binom{N_{si}}{y_{sij}} p_s^{y_{sij}} (1-p_s)^{N_{si}-y_{sij}} \frac{e^{-\lambda_s} \lambda_s^{N_{si}}}{N_{si}!} \\ &\quad \times \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_s^{\alpha-1} (1-p_s)^{\beta-1} dp_s. \end{aligned} \tag{3}$$

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

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

## 159 **1.2 Maximum Likelihood Estimation**

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

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

178 that consists of  $k$  copies of the originally observed data set. In other words, to imple-  
179 ment this method, one has to write the likelihood function of the data as if, one had  
180 observed  $k$  identical copies of the data set. Then, Lele *et al.* (2007, 2010) show that as  
181  $k$  grows large, the mean of the resulting posterior distribution converges to the MLE.  
182 In addition, for continuous parameters as  $\underline{\lambda}$ ,  $\bar{p}$ , and  $\tau$ , the variance covariance matrix  
183 of the posterior distribution converges to  $\frac{1}{k}$  times the inverse of the observed Fisher's  
184 information matrix. Thus, the variance estimated by the posterior distribution can  
185 be used to calculate Wald-type confidence intervals of the parameters (Lele *et al.*,  
186 2007, 2010). The advantage of data cloning over traditional Bayesian algorithms is  
187 that while in Bayesian algorithms the prior distribution might have influence over the  
188 posterior distribution, in data cloning the choice of the prior distribution does not  
189 determine the resulting estimates. In our case, the hierarchical model is

$$\mathbf{Y} \sim \text{Binomial}(\underline{\mathbf{N}}, \mathbf{P}) = f(y|\underline{\mathbf{N}} = n, \mathbf{P} = p) \quad (\text{Observation model}),$$

$$\underline{\mathbf{N}} \sim \text{Pois}(\underline{\lambda}) = g(\underline{\mathbf{N}}; \underline{\lambda}) \quad (\text{Process model}),$$

$$\mathbf{P} \sim \text{Beta}(\bar{p}\tau, (1 - \bar{p})\tau) = h(\mathbf{P}; \bar{p}, \tau) \quad (\text{Process model}).$$

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

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

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

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

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

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

## 205 **2 Methods**

### 206 **2.1 Estimation for Single Species**

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

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

## 222 **2.2 Assessing the Beta N-mixture Model performance**

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

### 239 **2.2.1 Bias benchmark assessment**

240 To evaluate the bias of the Beta N-mixture model, we simulated species counts in a  
241 100 ha quadrant sampled using 25, 50 meter circular plots visited three times each.

242 We assumed that the community was composed by 15 species with mean number of  
243 individuals/100 ha of  $\lambda = 1, 2, 3, 4, 5, 7, 10, 15, 25, 40, 55, 65, 75, 85, 100$ . In the latter  
244 vector each value of  $\lambda$  represents the abundance of a single species. In each simulation  
245 we drew  $N_{ij}$  individuals in each plot from a Poisson distribution with mean  $\lambda_i$ . We  
246 then simulated the detection process using a Binomial distribution with parameters  
247  $N_{ij}$  and  $p_i$ . We varied mean detection probability by assuming  $\bar{p} = 0.25, 0.5, 0.75$  and  
248  $\tau = 4.5$  ( $E[p] = 0.25, 0.5, 0.75$ ;  $\text{Var}[p] = 0.03, 0.04, 0.03$ ). Even though the variance  
249 seems small, the 2.5% and 97.5% quantiles of the three distributions range over a  
250 large portion of the  $[0,1]$  interval (quantiles 2.5 and 97.5: low =  $(0.01, 0.68)$ ; mid =  
251  $(0.1, 0.89)$ ; high =  $(0.31, 0.98)$ ). For each type of community we simulated 500 data  
252 sets, and estimated  $\lambda_i$ ,  $\bar{p}$  and  $\tau$  using data cloning. To determine the number of clones  
253 required to accurately get to the Maximum Likelihood Estimates of  $\lambda_i$ ,  $\bar{p}$  and  $\tau$  we  
254 used one randomly generated data set and estimated the parameters cloning the data  
255 sequentially from 1 to 64 times (Lele *et al.*, 2010). This allowed us to determine an  
256 adequate number of clones to get convergence of the  $k$ -th posterior mean to the  
257 MLEs. We used rjags v. 4.2.0 (Plummer, 2014) with two Markov chains allowing  
258 each chain to run for 14000 generations sampling every 10 generations and discarded  
259 the first 4000 iterations. For each type of community we estimated the relative bias  
260  $((\text{Estimated}-\text{True})/\text{True})$  in  $\lambda_i$ ,  $\bar{p}$  and  $\tau$ .

### 261 **2.2.2 Comparison to other community abundance models**

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

268 detection probabilities are the inner hierarchies. The Normal N-mixture model goes  
269 one more level and assumes that the parameters  $\lambda$  governing the realized abundances  
270  $N$  also come from a stochastic process governed itself by hyper-parameters. The  
271 second difference between our model and the Normal N-mixture model is the dis-  
272 tributional assumption giving rise to detection probabilities. In our model  $p_s$  are  
273 assumed to be  $p_s \sim Beta(\tau\bar{p}, \tau(1 - \bar{p}))$  and in the Normal model,  $p_s = \frac{1}{1+e^{-(r_i)}}$  where  
274  $r_i \sim Norm(\mu, \sigma^2)$  which gives a Johnson's SB distribution between 0 and 1. Besides  
275 these two model differences, Yamaura *et al.* (2016) use a Bayesian approach to fit  
276 their hierarchical model, while we use the method of Maximum Likelihood estima-  
277 tion. Much discussion exists regarding the merits of each inferential approach for  
278 hierarchical models in Ecology (see for instance Lele & Dennis, 2009; Cressie *et al.*,  
279 2009). Here we limit ourselves to compare the results from Yamaura *et al.* (2016)'s  
280 estimation approach, taken as the benchmark of a known method in the literature,  
281 to our approach. Note that using data cloning, one could compute the Maximum  
282 likelihood estimates of the model parameters in Yamaura *et al.* (2016)'s model, and  
283 their associated properties but doing so is outside the scope of our study.

284 We simulated 500 data sets under a spatially explicit model and for each data  
285 set we fitted the Normal N-mixture model then compared the posterior mean and  
286 mode estimates with the MLEs for the model proposed here (see Figure 2). For each  
287 simulation, we randomly drew 30  $\lambda_i$  from a gamma distribution with parameters  $\alpha =$   
288 0.65,  $\beta = 0.033$  and excluded  $\lambda_i$  values smaller than 1 individuals/100 ha, resulting in  
289 a community of 27 species. The gamma distribution used is the best fit of an observed  
290 species abundance distribution of a neotropical bird assemblage that was gathered  
291 using field intensive methods (Robinson *et al.*, 2000). Following, we randomly drew  
292 from a poisson distribution with mean  $\lambda_i$  the number of individuals of the  $i$ -th species  
293 ( $N_i$ ) present in a 100 hectares plot. We located each individual randomly across the  
294 plot and following, we randomly placed 25 circular plots with a radius of 50 meters

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

302 For each of the simulated data sets we estimated  $\lambda_i$ ,  $\bar{p}$  and  $\tau$  under the Beta N-  
303 mixture model using maximum Likelihood estimation with Data Cloning (Lele *et al.*,  
304 2007). We used rjags (Plummer, 2014) to build the model and run the analysis with  
305 2 chains, with 15000 iterations in each chain and retained the parameter values every  
306 10 generations after a burn-in period of 4000 generations. After initial parameter  
307 estimation, we sampled the posterior distribution given the estimated parameters to  
308 obtain the realized values of  $p_i$  given the data. For the Normal N-mixture model  
309 we performed bayesian parameter estimation using rjags and ran the analysis using  
310 2 chains, with 50000 iterations and retained parameters values every 20 generations  
311 after a burn-in of 10000 generations. In the latter case, we retained the mean and  
312 mode of  $\lambda_i$ ,  $p_i$  for comparison with the beta N-mixture model.

### 313 **2.3 Example Using Real Data**

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

## 319 **3 Results**

### 320 **3.1 Estimation for Single Species**

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

### 331 **3.2 Assessing the Beta N-mixture Model performance**

#### 332 **3.2.1 Bias Benchmark assessment**

333 We found that the parameters of the Beta N-mixture model are fully identifiable since  
334 the relative magnitude of the first eigenvalue of the parameter variance-covariance ma-  
335 trix decreased very closely at a rate of  $1/k$  (*eigenvalue* =  $-0.066 + 1.019(1/k)$ ;  $r^2 =$   
336  $0.98$ ). This result also identified that 20 clones were sufficiently large to guarantee  
337 convergence to the MLEs. The Beta model tends to slightly overestimate the abun-  
338 dance of rare species and underestimate the abundance of abundant species but this  
339 tendency decreases with increasing detection probability (Figure A2). This is ev-  
340 idenced by the slopes estimated by the relationship between estimated and true  $\lambda$ .  
341 The relationship for  $p = 0.25$  resulted was  $\hat{\lambda} = 5.8 + 0.7\lambda$ , for  $p = 0.5$  was  $\hat{\lambda} = 4 + 0.9\lambda$   
342 and for  $p = 0.75$  was  $\hat{\lambda} = 3.3 + 0.95\lambda$ . The bias decreased (approximately) as a func-  
343 tion of the true value of  $\lambda$  according to the equation  $bias(\lambda) = -0.45(\frac{1}{\lambda} + 7.5)$  for

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

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

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

### 357 **3.2.2 Comparison to other community abundance models**

358 The beta N-mixture model performed better than the Normal model in estimating the  
359 abundance and detection probability of rare species. While the posterior means and  
360 modes of the Normal model were biased towards species with abundances lower than  
361 4 individuals/100 ha, Maximum Likelihood Estimates of the Beta model were not  
362 (Figure 3). Furthermore, we show that the posterior means tended to be more biased  
363 than the posterior mode in estimating  $\lambda$  (Figure 3). The opposite seems to be true  
364 for the detection probabilities  $p$ . Both, the posterior mode and mean underestimated  
365  $p$  for rare species (Figure 4).

### 366 **3.3 Example Using Real Data**

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

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

## 375 4 Discussion

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

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

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

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

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

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

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

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

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

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

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

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

504 Our simulations have pushed the limits of community abundance models by  
505 simulating species with lower abundance than any other simulation (see Yamaura  
506 *et al.*, 2016). We hope that our results encourage tropical ecologists to use commu-  
507 nity abundance hierarchical models as a means to adequately estimate the abundance  
508 of full communities. In the recent North American Ornithological congress (August  
509 2016), two of us (JPG and SKR) participated in a wide, round table discussion where  
510 it was evident that tropical ornithologists are currently facing strong publishing chal-  
511 lenges because so far, abundance estimating techniques have not explicitly targeted  
512 estimation in a setting like the tropics: with very low abundances and sparse counts.  
513 Unlike temperate forests, where these methodologies have been widely used, in the  
514 tropics the species number is typically very large, but the counts per species very  
515 low. Our results, although worked out using birds as a study system, suggest that it  
516 is possible to have a reasonable estimates of the density of all of the species in the  
517 community for this particular scenario and different taxonomic groups (*e.g.* mam-  
518 mals, insects, plants, fungi, bacteria). Unbiased estimation of abundances using these  
519 hierarchical models will hopefully enable building more accurate species abundance  
520 distributions, which in turn can be extremely useful for understanding the mecha-  
521 nisms governing biodiversity patterns (McGill *et al.*, 2007)

## 522 **5 Acknowledgements**

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

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

## 529 References

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

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

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

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

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

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

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

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

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

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

553 Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography*, vol-  
554 ume 32. Princeton University Press, Princeton, NY.

555 Iknayan, K.J., Tingley, M.W., Furnas, B.J. & Beissinger, S.R. (2014) Detecting diver-  
556 sity: emerging methods to estimate species diversity. *Trends in ecology & evolution*,  
557 **29**, 97–106.

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

561 Karr, J.R., Robinson, S.K., Blake, J.G., Bierregaard Jr, R.O. & Gentry, A. (1990)  
562 Birds of four neotropical forests. A.H. Gentry, ed., *Four neotropical rainforests*, pp.  
563 237–269. Yale University Press New Haven, Connecticut.

564 Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes,  
565 M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D. *et al.* (2004) The metacom-  
566 munity concept: a framework for multi-scale community ecology. *Ecology letters*,  
567 **7**, 601–613.

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

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

- 573 Lele, S.R., Nadeem, K. & Schmuland, B. (2010) Estimability and likelihood inference  
574 for generalized linear mixed models using data cloning. *Journal of the American*  
575 *Statistical Association*, **105**, 1617–1625.
- 576 MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Lang-  
577 timm, C.A. (2002) Estimating site occupancy rates when detection probabilities are  
578 less than one. *Ecology*, **83**, 2248–2255.
- 579 Margules, C.R. & Pressey, R.L. (2000) Systematic conservation planning. *Nature*,  
580 **405**, 243–253.
- 581 Martin, J., Royle, J.A., Mackenzie, D.I., Edwards, H.H., Kery, M. & Gardner, B.  
582 (2011) Accounting for non-independent detection when estimating abundance of  
583 organisms with a bayesian approach. *Methods in Ecology and Evolution*, **2**, 595–  
584 601.
- 585 Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy,  
586 S.J., Tyre, A.J. & Possingham, H.P. (2005) Zero tolerance ecology: improving  
587 ecological inference by modeling the source of zero observations. *Ecology letters*, **8**,  
588 1235–1246.
- 589 McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K.,  
590 Dornelas, M., Enquist, B.J., Green, J.L., He, F., Hurlbert, A.H., Magurran, A.E.,  
591 Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I. & White,  
592 E.P. (2007) Species abundance distributions: moving beyond single prediction the-  
593 ories to integration within an ecological framework. *Ecology letters*, **10**, 995–1015.
- 594 Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.,  
595 Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C. *et al.*  
596 (2001) Terrestrial ecoregions of the world: A new map of life on earth a new global

- 597 map of terrestrial ecoregions provides an innovative tool for conserving biodiversity.  
598 *BioScience*, **51**, 933–938.
- 599 Ovaskainen, O. & Soininen, J. (2011) Making more out of sparse data: hierarchical  
600 modeling of species communities. *Ecology*, **92**, 289–295.
- 601 Parker III, T., Stotz, D. & Fitzpatrick, J. (1996) Ecological and distributional  
602 databases for neotropical birds. D. Stotz, J. Fitzpatrick, T. Parker III &  
603 D. Moskovits, eds., *Neotropical birds: ecology and conservation*. University of  
604 Chicago Press, Chicago.
- 605 Plummer, M. (2014) *rjags: Bayesian graphical models using MCMC*. R package  
606 version 3-13.
- 607 Ponciano, J.M., Burleigh, J.G., Braun, E.L. & Taper, M.L. (2012) Assessing param-  
608 eter identifiability in phylogenetic models using data cloning. *Systematic biology*,  
609 **61**, 955–972.
- 610 R Core Team (2013) *R: A Language and Environment for Statistical Computing*. R  
611 Foundation for Statistical Computing, Vienna, Austria.
- 612 Robinson, W.D., Brawn, J.D. & Robinson, S.K. (2000) Forest bird community struc-  
613 ture in central panama: influence of spatial scale and biogeography. *Ecological*  
614 *Monographs*, **70**, 209–235.
- 615 Royle, J.A. (2004) N-mixture models for estimating population size from spatially  
616 replicated counts. *Biometrics*, **60**, 108–115.
- 617 Royle, J.A. & Dorazio, R.M. (2008) *Hierarchical modeling and inference in ecology:*  
618 *the analysis of data from populations, metapopulations and communities*. Academic  
619 Press, San Diego, CA.

- 620 Ruiz-Gutiérrez, V., Zipkin, E.F. & Dhondt, A.A. (2010) Occupancy dynamics in a  
621 tropical bird community: unexpectedly high forest use by birds classified as non-  
622 forest species. *Journal of Applied Ecology*, **47**, 621–630.
- 623 Sauer, J.R. & Link, W.A. (2002) Hierarchical modeling of population stability and  
624 species group attributes from survey data. *Ecology*, **83**, 1743–1751.
- 625 Sibuya, M., Yoshimura, I. & Shimizu, R. (1964) Negative multinomial distribution.  
626 *Annals of the Institute of Statistical Mathematics*, **16**, 409–426.
- 627 Taper, M.L. & Ponciano, J.M. (2016) Evidential statistics as a statistical modern  
628 synthesis to support 21st century science. *Population Ecology*, **58**, 9–29.
- 629 Terborgh, J., Robinson, S.K., Parker III, T.A., Munn, C.A. & Pierpont, N. (1990)  
630 Structure and organization of an amazonian forest bird community. *Ecological*  
631 *Monographs*, **60**, 213–238.
- 632 Thiollay, J.M. (1994) Structure, density and rarity in an amazonian rainforest bird  
633 community. *Journal of Tropical Ecology*, **10**, 449–481.
- 634 Tuyl, F., Gerlach, R. & Mengersen, K. (2008) A comparison of bayes–laplace, jeffreys,  
635 and other priors: The case of zero events. *The American Statistician*, **62**, 40–44.
- 636 Yamaura, Y. (2013) Confronting imperfect detection: behavior of binomial mixture  
637 models under varying circumstances of visits, sampling sites, detectability, and  
638 abundance, in small-sample situations. *Ornithological Science*, **12**, 73 – 78.
- 639 Yamaura, Y., Andrew Royle, J., Kuboi, K., Tada, T., Ikeno, S. & Makino, S. (2011)  
640 Modelling community dynamics based on species-level abundance models from de-  
641 tection/nondetection data. *Journal of applied ecology*, **48**, 67–75.

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

645 Yamaura, Y., Royle, J.A., Shimada, N., Asanuma, S., Sato, T., Taki, H. & Makino,  
646 S. (2012) Biodiversity of man-made open habitats in an underused country: a class  
647 of multispecies abundance models for count data. *Biodiversity and Conservation*,  
648 **21**, 1365–1380.

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

652 **6 Tables**

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

653 **7 Figures**

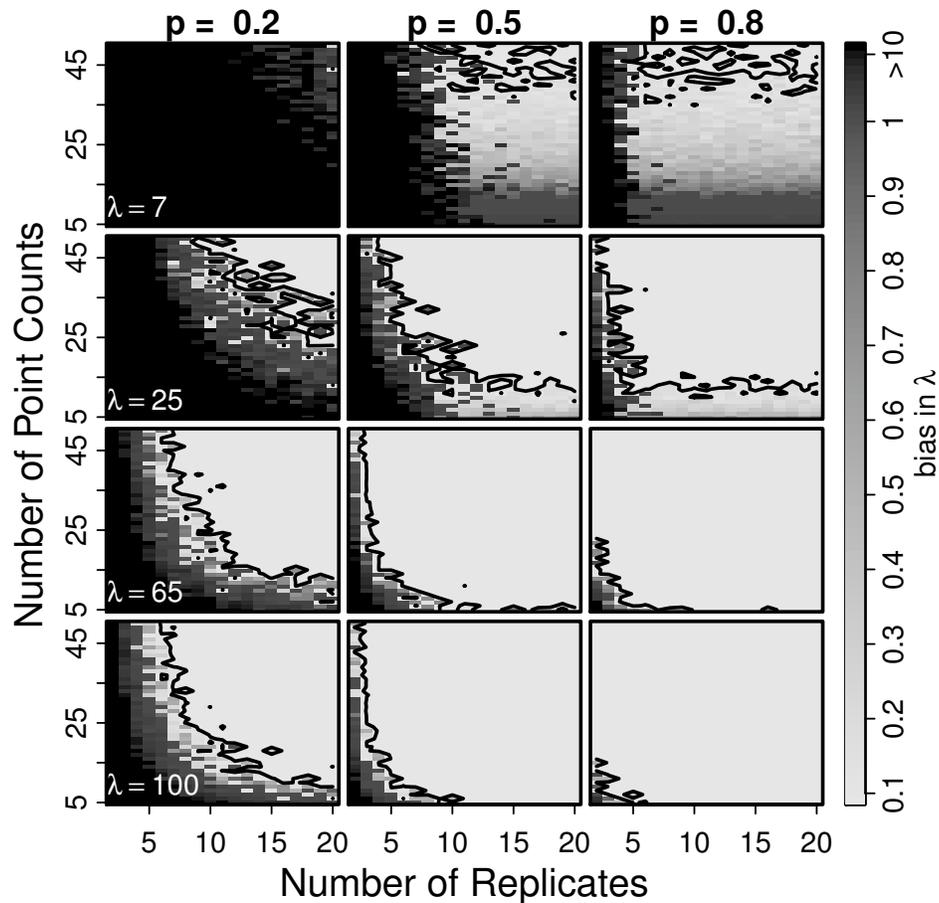


Figure 1: Mean bias in mean number of individuals per 100 ha  $\lambda$  for a range of point counts, number of replicates, and true parameter values for mid low and high abundances and detection probabilities ( $\lambda = 7, 25, 65, 100$  and  $p = 0.2, 0.5, 0.8$ ). The grayscale in each panel represents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

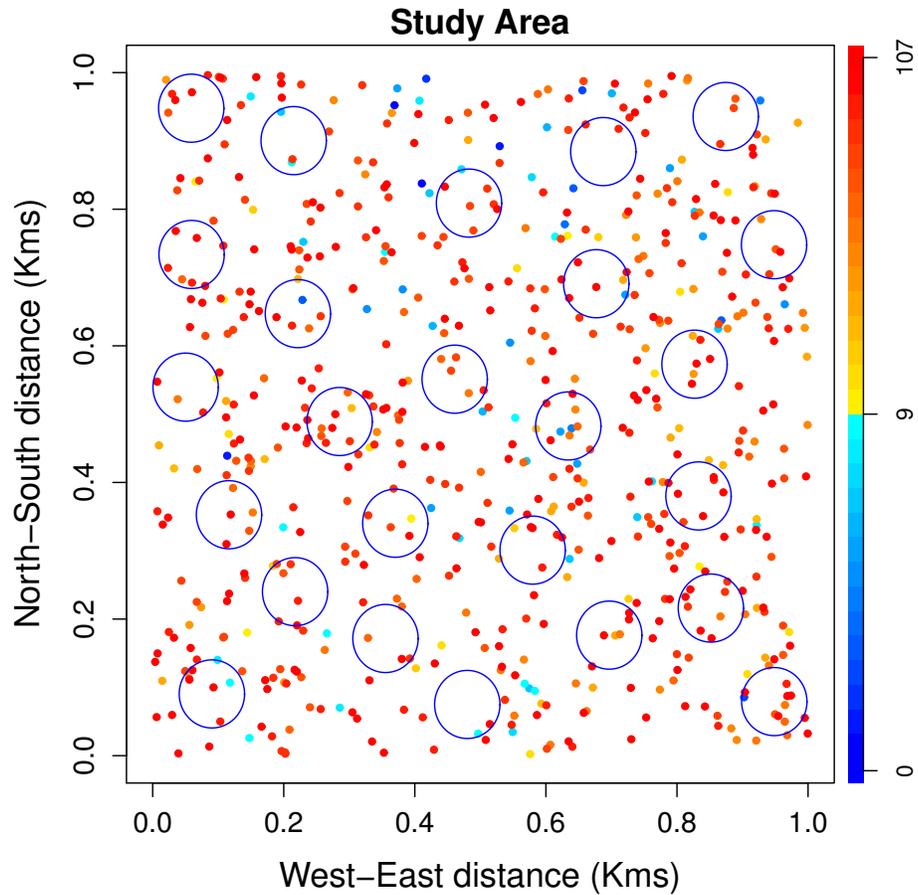


Figure 2: Graphic representation of the sampling design used to simulate the 500 count datasets of a community consisting of 27 species. We assumed the plot 20 be 100 ha ( $1 \text{ km}^2$ ) and circular sampling point to be of 0.78 ha ( $\sim 0.008 \text{ km}^2$ ). We show the true abundances in the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

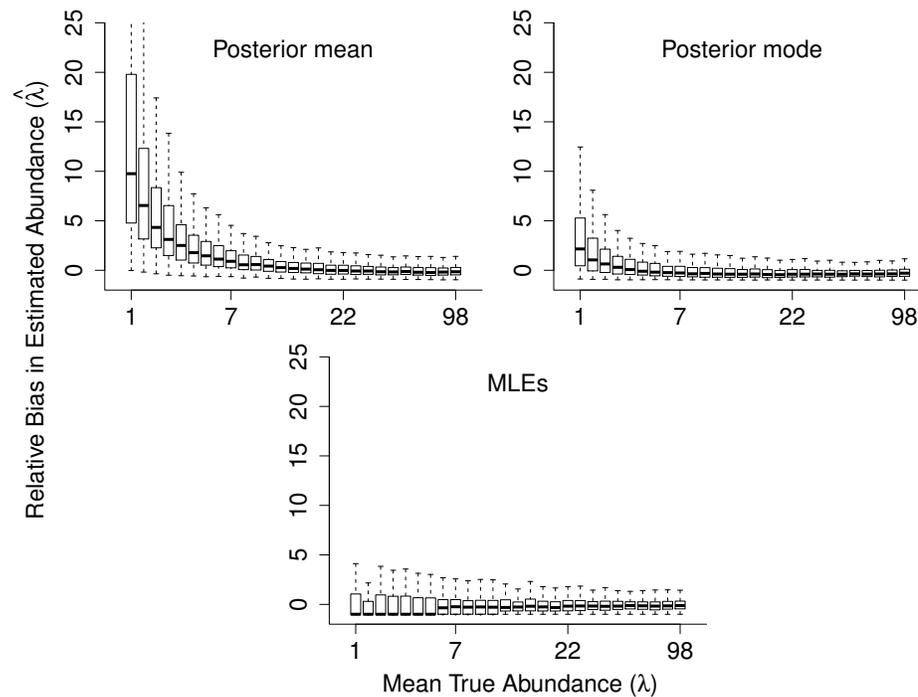


Figure 3: Relative bias in the estimated value of  $\lambda$  ( $(\text{Estimate}-\text{True})/\text{True}$ ) for both the Beta and Normal N-mixture model for 500 simulations of count data, for a community consisting of 27 species. We show the boxplots of the 500 posterior means and modes for the Normal 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.

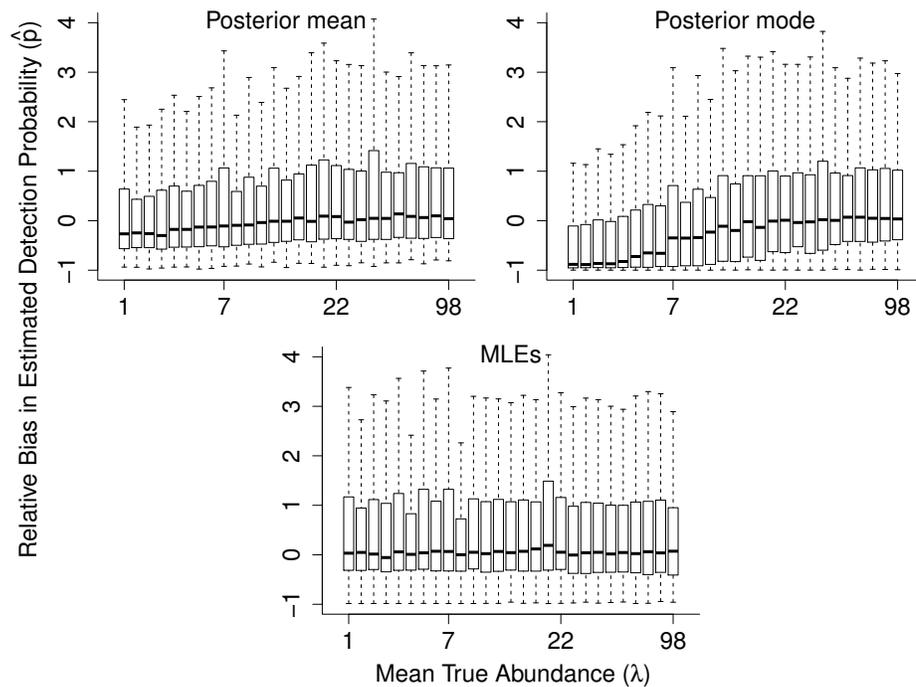


Figure 4: Relative bias in the estimated value of  $p$  ( $(\text{Estimate}-\text{True})/\text{True}$ ) as a function of the true abundance for both the Beta and Normal N-mixture model for 500 simulations of count data, for a community consisting of 27 species. We show the boxplots of the 500 posterior 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.

<sup>654</sup> **A** **Supplementary Figures**

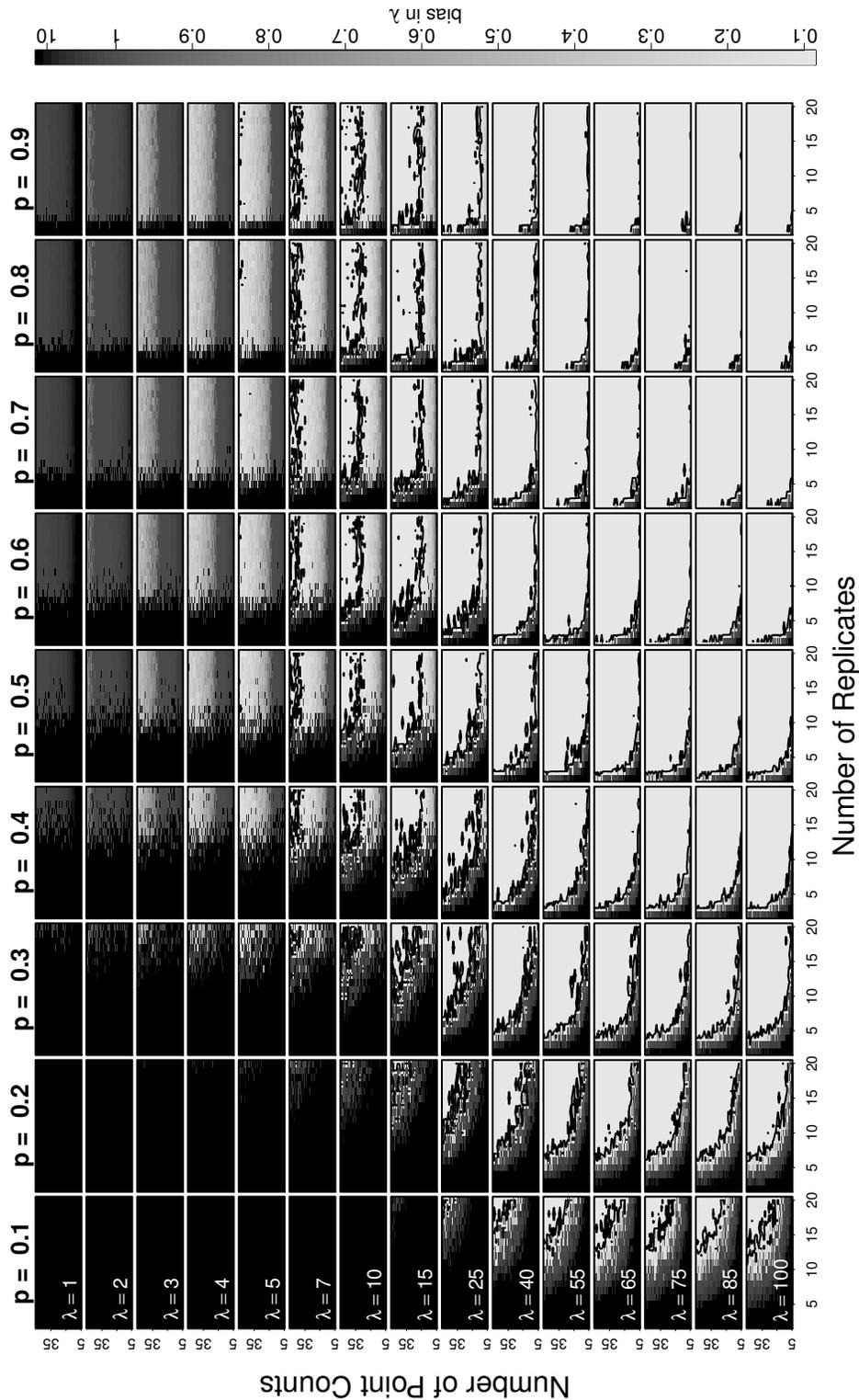


Figure A1: Mean bias in mean number of individuals per 100 ha  $\lambda$  for range of point counts, number of replicates, and true parameter values for low, mid and high abundances and detection probabilities ( $\lambda = 7, 25, 65, 100$  and  $p = 0.2, 0.5, 0.8$ ). The grayscale in each panel represent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

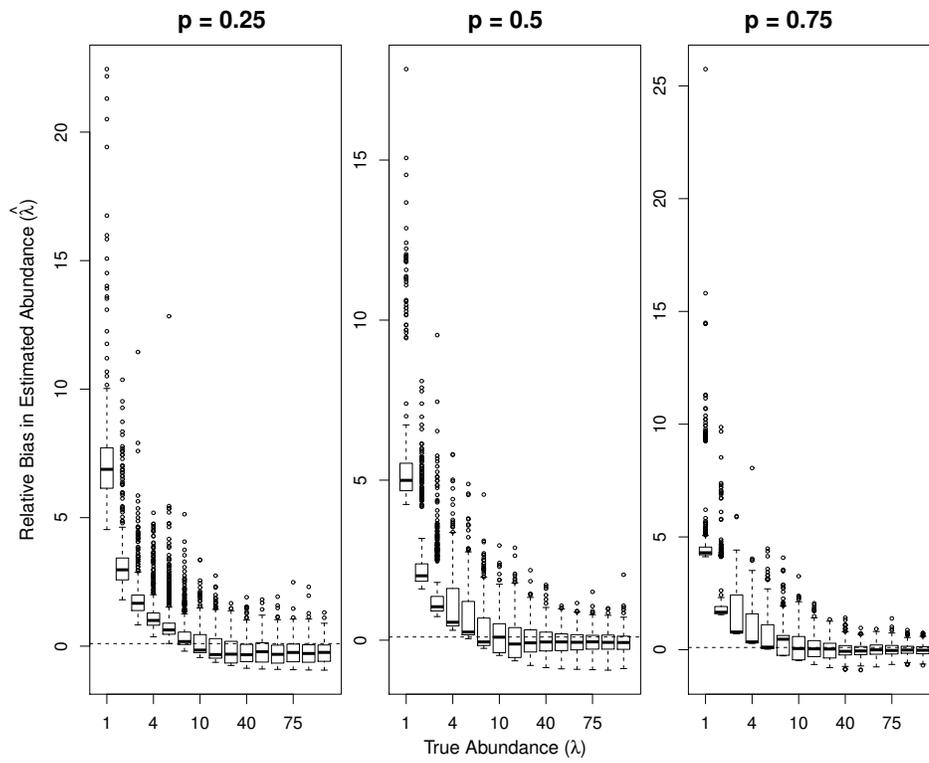


Figure A2: Boxplot showing the distribution of  $\hat{\lambda}$  using Beta N-mixture model, showing the location of the true value of  $\lambda$ .

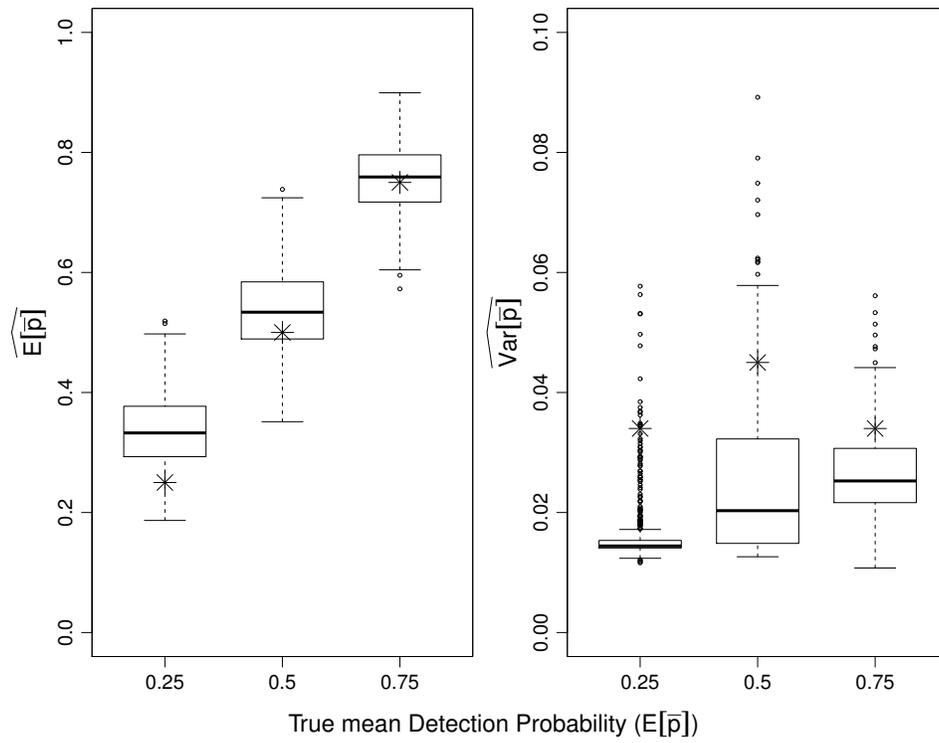


Figure A3: Boxplots showing the distribution of  $\widehat{E}[\bar{p}]$  and  $\widehat{Var}[\bar{p}]$  as a function of the true mean detection probability  $E[\bar{p}]$  with which data was simulated.

## 655 **B Bird sampling in the dry forests of the Mag-** 656 **dalena Valley**

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

## 673 C R Code

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