




## RESEARCH ARTICLE

# Sensitivity of binomial N-mixture models to overdispersion: The importance of assessing model fit

Jonas Knape<sup>1</sup>  | Debora Arlt<sup>1</sup> | Frédéric Barraquand<sup>2</sup>  | Åke Berg<sup>1</sup> | Mathieu Chevalier<sup>1</sup> | Tomas Pärt<sup>1</sup> | Alejandro Ruete<sup>1,3</sup>  | Michał Żmihorski<sup>1</sup>

<sup>1</sup>Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

<sup>2</sup>University of Bordeaux, Talence, France

<sup>3</sup>Greensway AB, Uppsala, Sweden

## Correspondence

Jonas Knape, Department of Ecology, Swedish University of Agricultural Sciences, Box 7044, S-75007, Uppsala, Sweden.  
Email: jonas.knape@slu.se

## Funding information

Swedish Research Council VR, Grant/Award Number: 621-2012-4076; FORMAS, Grant/Award Number: 942-2015-287; Swedish EPA; LabEx COTE, Grant/Award Number: ANR-10-LABX-45

Handling Editor: Nick Isaac

## Abstract

1. Binomial N-mixture models are commonly applied to analyse population survey data. By estimating detection probabilities, N-mixture models aim at extracting information about abundances in terms of absolute and not just relative numbers. This separation of detection probability and abundance relies on parametric assumptions about the distribution of individuals among sites and of detections of individuals among repeat visits to sites. Current methods for checking assumptions are limited, and their computational complexity has hindered evaluations of their performance.
2. We use simulations and a case study to assess the sensitivity of binomial N-mixture models to overdispersion in abundance and in detection, develop computationally efficient graphical goodness of fit checks to detect it, and evaluate the ability of the checks to identify overdispersion.
3. The simulations show that if the parametric assumptions are not exact the bias in estimated abundances can be severe: underestimation if there is overdispersion in abundance relative to the fitted model and overestimation if there is overdispersion in detection. Our goodness-of-fit checks performed well in detecting lack of fit when the abundance distribution was overdispersed, but struggled to detect lack of fit when detections were overdispersed. We show that the inability to detect lack of fit due to overdispersed detection is caused by a fundamental similarity between N-mixture models with beta-binomial detections and N-mixture models with negative binomial abundances.
4. The strong biases that can occur in the binomial N-mixture model when the distribution of individuals among sites, or the detection model, is mis-specified implies that checking goodness of fit is essential for sound inference about abundance. To check the assumptions we provide computationally efficient goodness of fit checks that are available in an R-package `nmixgof`. However, even when a binomial N-mixture model appears to fit the data well, estimates are not robust in the presence of overdispersion. We show that problems can occur even when estimated detection probabilities are high, and that previously reported problems with negative binomial models cannot always be diagnosed by checking the sensitivity of abundance estimates to numerical cutoff values used in likelihood computations.

**KEYWORDS**

abundance, beta-binomial, binomial, N-mixture model, negative binomial, overdispersion, Poisson, population survey

## 1 | INTRODUCTION

Count surveys are often conducted as parts of population monitoring programmes and ecological studies to follow changes in abundance of organisms in the wild. N-mixture models (Royle, 2004; Royle & Dorazio, 2006) are increasingly being applied to data from count surveys to correct for imperfect detection and yield estimates of absolute abundances instead of just relative abundances. These models are intuitively appealing because they can be applied to data from surveys with simple as well as more complex field protocols and allow simultaneous inclusion of explanatory variables for both abundance and detection processes.

Binomial N-mixture models (Royle, 2004) (hereafter referred to simply as “N-mixture models”) are a particular class of models that rely only on repeated counts from a large number of sites to estimate absolute abundance while accounting for imperfect detection using binomial detection models. The simplicity of collecting data under the protocol of N-mixture models has led to their adoption for a range of different purposes, including to evaluate the effectiveness of conservation actions (Romano et al., 2017), to gain knowledge about absolute abundance and population dynamics (Studds et al., 2017), to predict population responses to differing conservation scenarios (Ladin, D’Amico, Baetens, Roth, & Shriver, 2016), and to forecast shifts in species distribution (Hunter, Nibbelink, & Cooper, 2017).

Because N-mixture models rely on multiple assumptions, it is vital for reliable inference to investigate how sensitive estimates are to deviations from assumptions, and to devise methods for checking any assumptions that the models are sensitive to. Specifically, overdispersion is a common feature of population count data (ver Hoef & Boveng, 2007; Linden & Mäntyniemi, 2011), and the focus of this paper.

Previous studies have found that abundance can be severely overestimated under N-mixture models when detection probabilities vary randomly among visits and that such overdispersion may be dealt with using a beta-binomial detection model (Martin et al., 2011). Similarly, abundance may be greatly overestimated when population sizes vary among repeat visits to the same sites (Duarte, Adams, & Peterson, 2018) or when animals are double counted (Link, Schofield, Barker, & Sauer, 2018). Several studies have addressed overdispersion in abundance among sites relative to a Poisson distribution, which has a restrictive variance-mean scaling, by assuming a zero inflated Poisson, negative binomial or Poisson log-normal abundance mixtures. Results have shown that abundance estimates vary depending on which abundance mixture is being used (Kéry, Royle, & Schmid, 2005; Joseph, Elkin, Martin, & Possingham, 2009), and that estimates from models with a negative binomial abundance mixture often are preferred by model selection criteria but behave

poorly, yielding infinite maximum likelihood estimates of abundance (Dennis, Morgan, & Ridout, 2015; Haines, 2016; Kéry, 2018).

Our main aims in this paper are threefold. First, we examine how sensitive N-mixture models are to overdispersion in detection among visits relative to the standard binomial detection model, or in abundance relative to the standard Poisson distribution. Second, we develop computationally efficient tools to assess to what extent data are overdispersed relative to a fitted N-mixture model. They are based on graphical checks of randomized quantile residuals (Dunn & Smyth, 1996; Warton, Lyons, Stoklosa, & Ives, 2016) and provide an answer to recent calls for further development of methods for assessing model fit of N-mixture models (Kéry & Royle, 2016; Knappe & Korner-Nievergelt, 2016). Third, we investigate the feasibility of separating between overdispersion in abundance and detection.

We study the effects of overdispersion on abundance estimates and the performance of goodness-of-fit checks in a case study of a wetland bird and in two simulation scenarios with overdispersion in the abundance distribution and in the detection model respectively. The goodness-of-fit checks are available in an R-package `nmixgof`.

## 2 | MATERIALS AND METHODS

In this section, we first introduce the basics of the N-mixture model. In Section 2.2, we then develop graphical methods for assessing the fit of N-mixture models. In Section 2.3, we demonstrate the use of the goodness-of-fit checks in a case study on wetland birds in Sweden. Finally, in Section 2.4, we investigate the sensitivity of N-mixture models to overdispersion in the abundance and detection models and the ability of the goodness-of-fit checks to detect violation of the distributional assumptions.

### 2.1 | N-mixture models

N-mixture models are a suite of models for abundance data obtained from repeat count surveys at multiple sites (Royle, 2004). They model the data as arising from an abundance process describing the spatial variation in the number of individuals among sites and a detection process describing how many of the individuals present at each site are found at each visit. Data come from a set of  $R$  different sites and for the abundance process it is assumed that the numbers of individuals at sites,  $N_i$ , are distributed according to some discrete statistical distribution with probability function  $g$ ,

$$N_i \sim g(N_i; \lambda_i, \theta),$$

where the draw for each site is independent,  $\lambda_i$  is describing the expected abundance in site  $i$  and  $\theta$  is an optional parameter for

overdispersion in the abundance distribution. To avoid overparameterization, restrictions on  $\lambda_i$  are necessary and are achieved by using a shared parameter across all sites (i.e., a constant  $\lambda$ ) or by letting  $\lambda_i$  be a function of site specific covariates.

In most applications,  $g$  is the probability function of either a Poisson, a zero-inflated Poisson (ZIP), or a negative binomial distribution. We will focus on these three mixtures in this paper. For the ZIP mixture, we use the parameterization

$$\begin{aligned} N_i &= 0 && \text{with probability } \psi \\ N_i &\sim \text{Poisson}(\lambda_i) && \text{with probability } 1 - \psi \end{aligned}$$

where  $\psi$  is the probability of an excess zero. For the negative binomial mixture, we use the parameterisation

$$N_i \sim \text{NegBin}(\lambda_i, \theta)$$

such that the variance of  $N_i$  is  $V(N_i) = \lambda_i + \theta \lambda_i^2$  (Supporting Information Appendix S1).

For each site  $i$ , observations come in the form of  $T$  counts,  $y_{i1}, \dots, y_{iT}$ , and for the detection model it is assumed that the counts are independent binomial draws with population size as index (Royle, 2004),

$$y_{it} \sim \text{Bin}(N_i, p_{it}),$$

where  $p_{it}$  is the detection probability associated with finding an individual that is present at site  $i$  at visit  $t$ . As for  $\lambda_i$ , restrictions on  $p_{it}$  are required to avoid overparameterization and are introduced, for example, by letting them be a function of site or visit specific covariates or by assuming identical detection probabilities across sites and visits.

The design idea underlying this model is that counts are conducted during repeat visits to each site during a period of time for which the local abundance is closed, so that at each visit all individuals are present but only a fraction is detected.

Sometimes additional variation in detection is allowed for by letting

$$y_{it} \sim \text{Bin}(N_i, p'_{it})$$

where the  $p'_{it}$  are varying stochastically according to a probability distribution. We restrict our attention to a beta distribution where  $p'_{it}$  vary independently among sites and visits,

$$p'_{it} \sim \text{Beta}\left(p_{it} \frac{1 - \delta^2}{\delta^2}, (1 - p_{it}) \frac{1 - \delta^2}{\delta^2}\right),$$

resulting in a beta-binomial detection model. The specific parameterization in the above equation ensures that  $p_{it}$  is the mean detection probability and that the standard deviation of  $p'_{it}$  scales linearly with  $\delta$  and is equal to  $\delta \sqrt{p_{it}(1 - p_{it})}$ , with  $0 \leq \delta \leq 1$ . The beta-binomial distribution acts as overdispersion for the detection probability, and the strength of the overdispersion is determined by the parameter  $\delta$ .

## 2.2 | Checking for over-dispersion and goodness of fit

Checking the fit and assumptions of hierarchical models is difficult in general because distributional and independence assumptions occur

at multiple levels in the hierarchy, and through conditioning on unobserved stochastic variables. Current common practice for assessing goodness of fit of N-mixture models is to use parametric bootstrapping in combination with some goodness of fit statistic, often sums of squares or a Freeman–Tukey statistic (Kéry & Royle, 2016). This approach is computationally intensive since in each bootstrap sample the model under investigation needs to be fitted to simulated data a large number of times. In this section, we suggest three types of residuals to check the goodness of fit of N-mixture models. The benefit of these over the bootstrap procedure is that (1) they are faster to compute, and (2) residuals can be used to graphically check a range of assumptions such as overdispersion via quantile–quantile plots (qq plots), residual plots against fitted values to check homoscedasticity, and plots of residuals against covariates to check functional assumptions (Warton, Stoklosa, Guillera-Aroita, MacKenzie, & Welsh, 2017). Additionally, two measures of overdispersion relative to a fitted model, based on Pearson residuals, are described in Supporting Information Appendix S1. They provide simple scalar indicators of the level of overdispersion.

### 2.2.1 | Randomized-quantile residuals

We will define three types of randomized-quantile (rq), or Dunn–Smyth, residuals (Dunn & Smyth, 1996). Rq residuals have recently gained popularity in ecological analyses (Warton et al., 2016) due to their convenient property that they are standard normally distributed under the correct model. For sparse count data this means that plots of e.g., residuals against fitted values behave similarly to such plots for ordinary linear models which is not the case for standard residuals for count data. That the residuals are indeed normally distributed is easy to check, for example, using qq plots (Warton et al., 2016).

The normality of rq residuals is achieved by randomization. For a random count variable  $z$  with cumulative distribution function (CDF)  $F$  (which may depend on model parameters), they are defined by

$$\begin{aligned} r_{rq} &= \Phi^{-1}(u) \\ u &\sim \text{Unif}(F(z - 1), F(z)) \end{aligned} \quad (1)$$

where  $\Phi^{-1}$  is the inverse of the standard normal CDF and  $u$  is a value randomly generated from a uniform distribution. To compute rq residuals the function  $F$  needs to be computed and below we define three variants of rq residuals using CDFs corresponding to different aspects of the data and potentially picking up different aspects of model fit. The second type of residuals defined below (site-sum residuals) correspond to the “omnibus” residuals for occupancy models defined by Warton et al. (2017) while the other two have no direct correspondence to their occupancy model residuals.

#### Marginal rq residuals

For the first type of rq residuals, we take  $F$  to be the marginal distribution of the counts (i.e., the distribution of the counts when all possible latent abundances have been summed over). For the N-mixture model without heterogeneity in  $p_{it}$  and with a Poisson,

ZIP or negative binomial mixture distribution, the marginal distribution of each observation comes from the same type of distribution as that used for the abundance mixture. If, for example, the abundance mixture is given by  $N_i \sim \text{ZIP}(\lambda_i, \psi)$ , the marginal distribution of each  $y_{it}$  is  $\text{ZIP}(p_{it}\lambda_i, \psi)$ . In these cases the randomized-quantile residuals can be computed using the definition above (Equation 1).

For beta-binomial detection models with Poisson and ZIP abundance mixtures the marginal distributions are, respectively, beta-Poisson and zero inflated beta-Poisson distributions (Leask & Haines, 2014). For the beta-binomial model with negative binomial abundance mixture, the marginal distribution can be computed through numeric summation over  $N$  or by using the methods of Haines (2016).

A property of the marginal rq residuals computed from an  $N$ -mixture model is that residuals from the same site are not independent. Hence they should not be used directly in qq plots which assume independent variables. However, sets of residuals containing only one residual from each site are independent. Separate qq plots can therefore be drawn for, e.g., the first observations from each site, the second observations, and so on. Since there is one marginal rq residual per observation, they can be plotted against visit specific detection covariates as well as against site specific covariates.

### Site-sum rq residuals

The second type of residual is defined from the marginal distribution of the sum of the counts within each site  $y_{Sj} = \sum_t y_{it}$ . The marginal CDF for the site sums can be computed numerically using

$$F(y_{Sj}) \approx \sum_{N=y_{Sj}}^K F_{\text{BinSum}}(y_{Sj}; N, p_{i1}, \dots, p_{iT}) P_i(N)$$

where  $F_{\text{BinSum}}$  is the CDF of a sum of independent binomial variables, all with the same index  $N$  but potentially different probabilities  $p_{it}$ ,  $P_i(N)$  is the probability that the abundance at site  $i$  is equal to  $N$  given by the abundance distribution, and  $K$  is an upper bound to truncate the infinite sum over  $N$ . If the  $p_{it}$  are all the same  $F_{\text{BinSum}}$  is simply the cumulative probability function of a binomial distribution with index  $TN$  but if the  $p_{it}$  are not all identical then  $F_{\text{BinSum}}$  is more complex. In the general case it can be computed by brute force as a numeric sum:

$$F_{\text{BinSum}}(y_{Sj}; N, p_{i1}, \dots, p_{iT}) = \sum_{k_1 + \dots + k_T \leq y_{Sj}} P_{\text{Bin}}(k_1; N, p_{i1}) \times \dots \times P_{\text{Bin}}(k_T; N, p_{iT})$$

where  $P_{\text{Bin}}$  is the probability function of the binomial distribution. The same computation may be used for beta-binomial detection models by replacing  $P_{\text{Bin}}$  with  $P_{\text{BetaBin}}$ .

The idea of aggregating counts across sites is to make residuals independent and potentially to increase their informativeness in cases where counts are sparse. Since there is one site-sum residual per site, they can be used in plots against site-specific covariates.

### Observation rq residuals

We also explore a third type of, admittedly ad hoc, residuals that we refer to as observation residuals. The idea is to compute residuals from the observation model only by conditioning on the abundances, with the intent of more specifically checking the detection part of the model. Since abundances are not directly available from a fitted model, we use a random sample of abundances from the empirical Bayes distribution (the distribution of the abundances given the data and under the parameters obtained by maximum likelihood) for the conditioning. That is, residuals were computed using the binomial or beta-binomial CDF with  $N_i$  equal to a draw from the empirical Bayes distribution. The random draw introduces additional stochasticity to the residuals, which is likely to reduce power to detect lack of fit to some degree.

## 2.3 | Case study: Northern shoveler

To illustrate the effects of overdispersion and the performance of the residuals in assessing model fit, we analyse data from a wetland survey conducted in May and June of 2016 at 50 wetland sites across southern Sweden. Most sites (90%) were visited 10 times during a three week period, split between five visits by each of two observers. Remaining sites had fewer visits. The number of individuals for each of 70 bird species associated with wetlands was recorded on each visit. Here, we use counts for Northern shoveler *Anas clypeata* (Supporting Information Figure S1), a dabbling duck moderately common in lakes and wetlands in southern Sweden. We fit six  $N$ -mixture models to the data using combinations of Poisson (P), ZIP and negative binomial (NB) abundance mixtures and binomial (B) and beta-binomial (BB) detection. Hereafter the models will sometimes be referred to using abbreviations such as BB-ZIP with prefix denoting the detection model and suffix denoting the abundance distribution. All models included two covariates for abundance, the log transformed area of water at the wetland representing its size and the latitude of the wetland, and two covariates for detection, the date of the survey and the percentage of reed cover at the wetland as a proxy for visibility. All covariates were introduced as linear functions on the log (abundance) and logit scale (detection) and were standardized to mean 0 and standard deviation 1 prior to analyses. We used maximum likelihood estimation, and fitted models with binomial detection using the R-package *unmarked* (Fiske & Chandler, 2011) and models with beta-binomial detection using custom code (Supporting Information).

The  $N$ -mixture model as implemented in *unmarked* approximates the likelihood by truncating an infinite sum over all possible values of  $N$ . The upper bound,  $K$ , needs to be set when fitting the model, but estimates can be unstable to changes in this bound, possibly due to maximum likelihood estimates of abundance being infinite (Dennis et al., 2015). We used a numeric upper bound  $K = 400$  for abundance in the calculation of the likelihoods but also fitted the same models a second time using  $K = 1,000$  to check if the estimates were stable to this numeric cutoff. Closed form expressions for the

likelihoods, avoiding the need for a numeric bound, are available for N-mixture models with a binomial detection distribution (Dennis et al., 2015; Haines, 2016), but have yet to be implemented in general software for N-mixture models.

### 2.3.1 | Results of case study

Estimates under the Poisson and ZIP abundance mixtures were not sensitive to the numerical cutoff  $K$  while both models with an NB mixture were sensitive to the cutoff. The estimates obtained for the NB mixtures are thus not maximum likelihood estimates, and estimates of abundance will increase and those of detection decrease as  $K$  is increased. We will refer to them as truncated estimates. Models with binomial and beta-binomial detection give similar estimates under the same abundance mixture, but the estimates differ among abundance mixtures (Figure 1, Supporting Information Figure S2).

Qq plots of site-sum randomized quantile residuals show that models with Poisson or ZIP mixtures provide poor fits to the data since the quantiles deviate clearly from the identity line (Figure 2), while the truncated estimates of the NB mixtures appear adequate (Figure 2). The qq plots for the Poisson mixtures indicate that the largest residuals are larger and the smallest smaller than would be expected under Poisson mixtures while the qq plots for the ZIP mixtures show some improvement in terms of explaining the smallest (zero) observations, but are still at loss in explaining larger counts. Similar patterns are seen for the marginal rq residuals (Supporting Information Figure S3). AIC values indicate a poor fit of the Poisson and ZIP mixtures relative to the truncated NB mixture estimates (Table 1). AIC in addition suggests a poor fit of the truncated B-NB model relative to the truncated BB-NB model, which is not picked up by the qq plots of site-sum residuals. Qq plots of observation residuals however do suggest lack of fit of the truncated B-NB model (Figure 3). Qq plots of observation residuals for the truncated BB-NB model show no obvious lack of fit (Figure 3).

These results leave us in a quandary. The NB mixtures give unstable estimates and cannot be used for inferences about abundance, and the poor fit of the Poisson and ZIP mixtures suggest that we cannot use estimates from these models for reliable inference either. To check if the reason for the poor fit of the Poisson and ZIP mixtures might be due to incorrect functional covariate relationships we plot rq residuals against each of the covariates for the BB-ZIP model, which has the best fit among the models with stable estimates (Supporting Information Figure S4). Since there is no clear pattern in the residuals as a function of covariates for this model there appears to be no simple correction to improve its fit. The conclusion from this case study therefore has to be that we are not able to find an adequately fitting N-mixture model that provides reasonable estimates for the data at hand.

## 2.4 | Simulations

To investigate biases in abundance estimates due to overdispersion, and to evaluate the ability of our goodness-of-fit checks to identify

problems, we ran two simulation scenarios: one with overdispersion in the abundance distribution relative to the Poisson and one with overdispersion in detection relative to the binomial distribution such that detection probabilities vary independently among sites and visits.

### 2.4.1 | Scenario 1: Overdispersed abundance

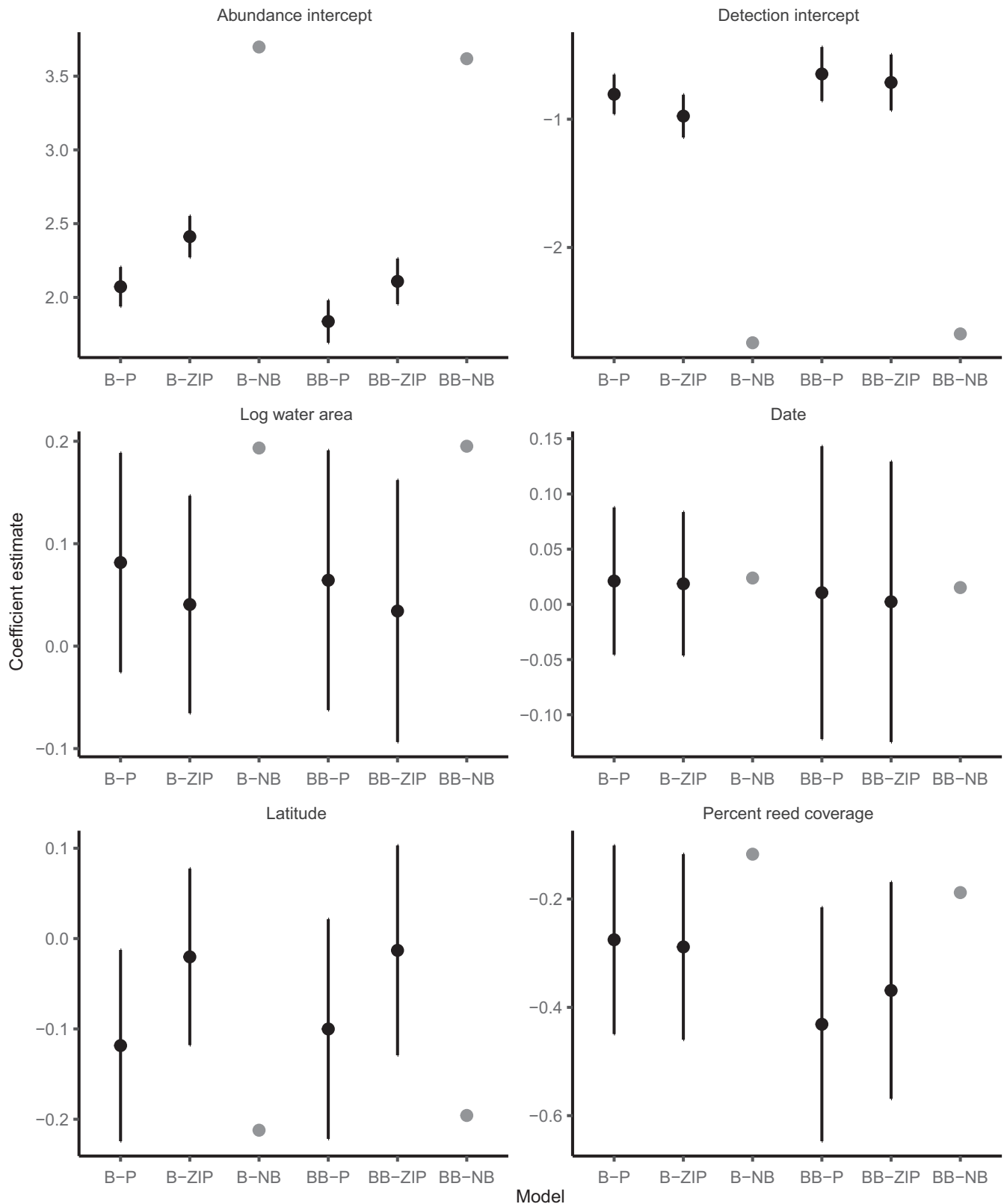
We simulated data over 200 sites, each visited five times, using a binomial detection model with  $p_{it}$  set to 0.25 for all visits and sites and with a constant expected abundance across all sites  $\lambda_i = 10$ . To investigate effects of overdispersion we used a negative binomial abundance distribution and varied the overdispersion coefficient  $\theta$  from 0 to 2 in steps of 0.25. Thus, data were generated using a B-NB model. For each value of  $\theta$ , 500 datasets were generated. For each simulated dataset, we fit a B-P, B-ZIP, B-NB (which in this simulation is the correct model), and a BB-P N-mixture model, each with a single intercept for detection and abundance but no covariates. The models with binomial detection (B-P, B-ZIP, and B-NB) were fitted using `unmarked` while the BB-P model was fitted using custom R-code. For computational reasons, we did not fit BB-ZIP and BB-NB models to the simulated data.

In addition, we fitted a second set of models that were identical to the ones described above except for the addition of a single covariate for abundance. The covariate was generated from a standard normal distribution and was used in the fitted models but was unrelated to the simulated data. These three models with covariates were fitted in order to investigate if overdispersion might lead to finding spurious effects of covariates (Richards, 2008).

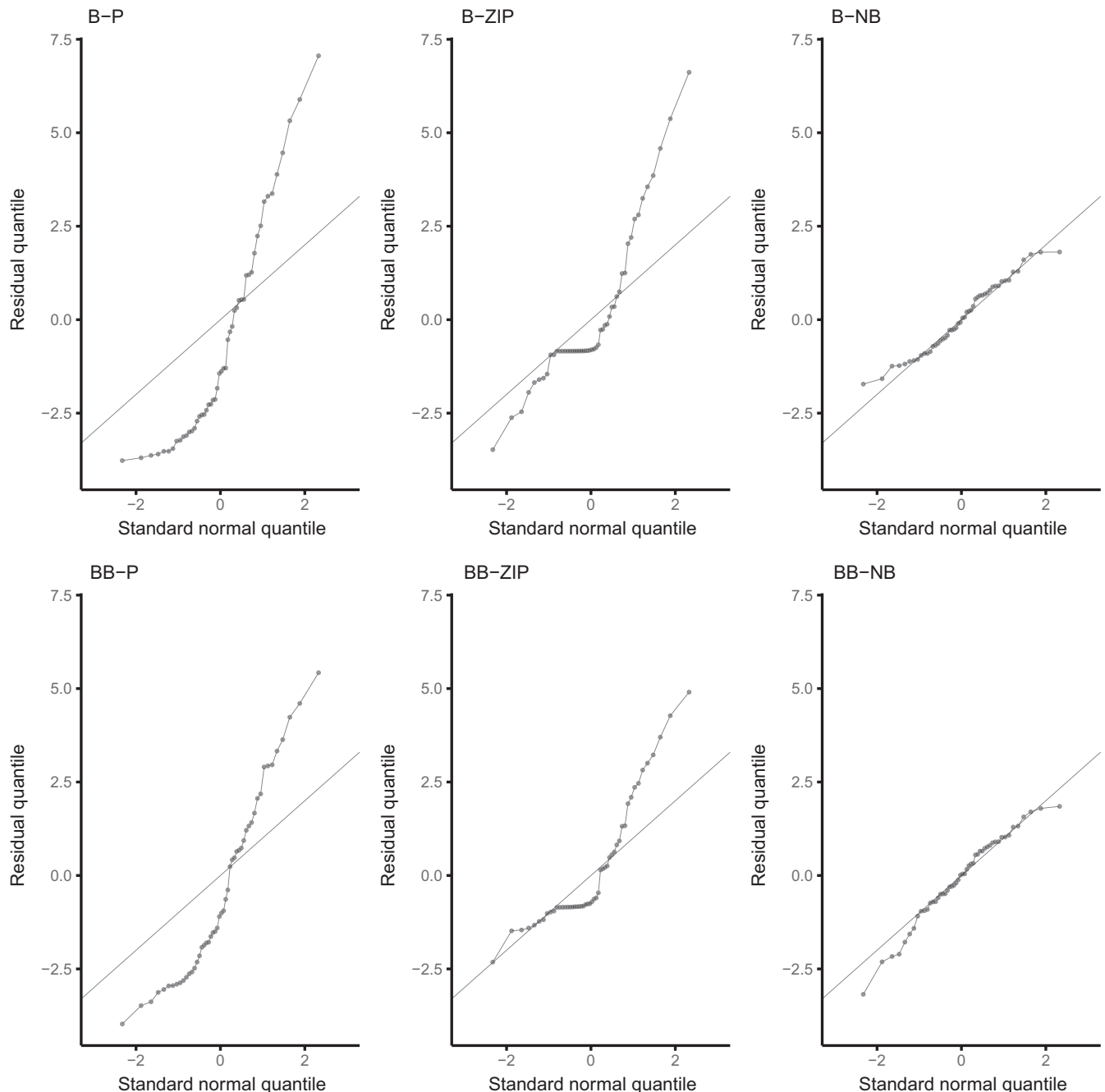
We used a numeric cutoff value  $K = 200$  for the calculation of the likelihood during model fitting. To check for stability of estimates with respect to  $K$  we additionally fitted each model using a  $K$  value of 400 and classified estimates as stable if the abundance intercept between the two  $K$  values differed by less than 0.01.

For all the fitted models, we retrieved parameter estimates and AIC values. As a rough estimate of power to detect non-normality from the qq plots of the randomized quantile residuals we computed the  $p$ -value from a Shapiro–Wilks test for the site-sum and observation residuals (this was not done for the marginal residuals because they are not independent among visits). We do not recommend this procedure in applications but used it here to obtain a crude but objective measure of the ability to detect lack of fit from the residuals. In applications, we suggest using graphical checks via qq plots and plots of residuals against fitted values and covariates because such checks provide more information about the nature of the lack of fit than a  $p$ -value does. To get an understanding of the variability that would be expected in a qq plot if the model was correct it can also be useful to compare qq plots for the fitted model to a few qq plots for data generated from a standard normal distribution with the same sample size (Loy, Follett, & Hofmann, 2016).

A simulation similar to scenario 1 but with covariates affecting abundance and detection is in Supporting Information Appendix S1.



**FIGURE 1** Estimates and 95% confidence intervals for intercepts and covariates coefficients for abundance (left panels) and detection (right panels) of the models fitted to Northern shoveler data. Prefix B and BB refers to, respectively binomial and beta-binomial detection models. Suffix P, ZIP and NB refers to Poisson, zero-inflated Poisson, and negative binomial abundance mixtures. Estimates under the NB mixtures are unstable and not maximum likelihood estimates. Truncated point estimates are given in grey for  $K = 400$  for those models, but confidence intervals are omitted



**FIGURE 2** QQ plots of site-sum randomized-quantile residuals against standard normal residuals for fits of models to the Northern shoveler data. Under a good fit residuals should be close to the identity line (grey). Prefix B and BB refers to, respectively binomial and beta-binomial detection models. Suffix P, ZIP and NB refers to Poisson, zero-inflated Poisson, and negative binomial abundance mixtures

## Results

Nearly all model fits converged and were stable with respect to  $K$  in this scenario (Figure 4a). Fitting the true B-NB model provided the least bias, nearly nominal confidence interval coverage for the covariate effect, and rejected the normality test for the rq residuals with probability at approximately the nominal 10% level (Figure 4).

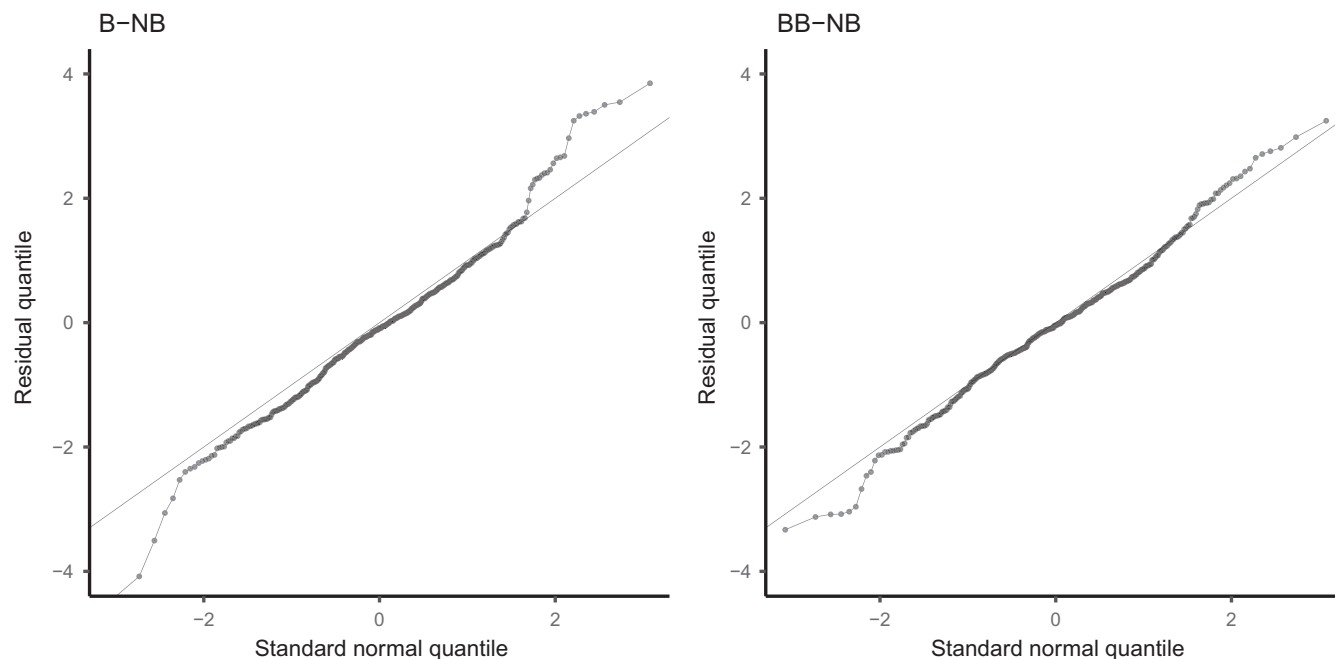
The B-P, B-ZIP and BB-P models strongly underestimated abundance for high levels of overdispersion with a relative bias of  $-50\%$  or below when overdispersion values were larger than 1 (Figure 4c), i.e., underestimation by more than a half. The BB-P model had the

strongest bias. These levels of bias are of similar magnitude to estimates not adjusted for detection, which had a relative bias of around  $-60\%$ . Overdispersion also led to poor confidence interval coverage for the spurious covariate effect, except when fitting the correct model (Figure 4d).

Lack of fit relative to the true B-NB model was readily identified by AIC in the simulations (Figure 4b).

Normality tests of the site-sum rq residuals rejected incorrect models at high rates (Figure 4g), but observation rq residuals had considerably lower power (Figure 4h).





**FIGURE 3** QQ plots of observation randomized quantile residuals against standard normal residuals for fits of binomial and beta-binomial NB models to the Northern shoveler data. Under a good fit residuals should be close to the identity line (grey). B and BB refers to, respectively binomial and beta-binomial detection models, while NB refers to the negative binomial abundance mixture

**TABLE 1** AIC values for fits to Northern shoveler data

Model	AIC
B-P	2,026.3
B-ZIP	1,915.6
B-NB	1,601.6
BB-P	1,789.5
BB-ZIP	1,719.8
BB-NB	1,568.3

Results for the additional simulation with covariates are similar (Supporting Information Figure S7).

## 2.4.2 | Scenario 2: Overdispersed detection

In the second scenario, we explored the effects of overdispersion in detection relative to the binomial distribution. The setup was similar to the setup in scenario 1, except that we used a Poisson abundance mixture and a beta-binomial detection model to simulate data (i.e., a BB-P model). We varied  $\delta$ , the amount of overdispersion in detection, from 0 to  $1/\sqrt{5}$ . The upper bound was chosen so that the distribution of the detection probability has an interior mode for all values of  $\delta$  except for  $\delta = 1/\sqrt{5}$  where the mode is at 0. We fitted the same models as in scenario 1.

A simulation similar to scenario 2 but with covariates affecting abundance and detection is in Supporting Information Appendix S1.

## Results

Most model fits in scenario 2 converged and were stable with respect to  $K$ , except under the B-NB model that failed for almost all simulated datasets when  $\delta > 0.2$  (Figure 5a). Properties of the model fits like bias, coverage, etc. were computed only from fits that converged and were stable with respect to  $K$ .

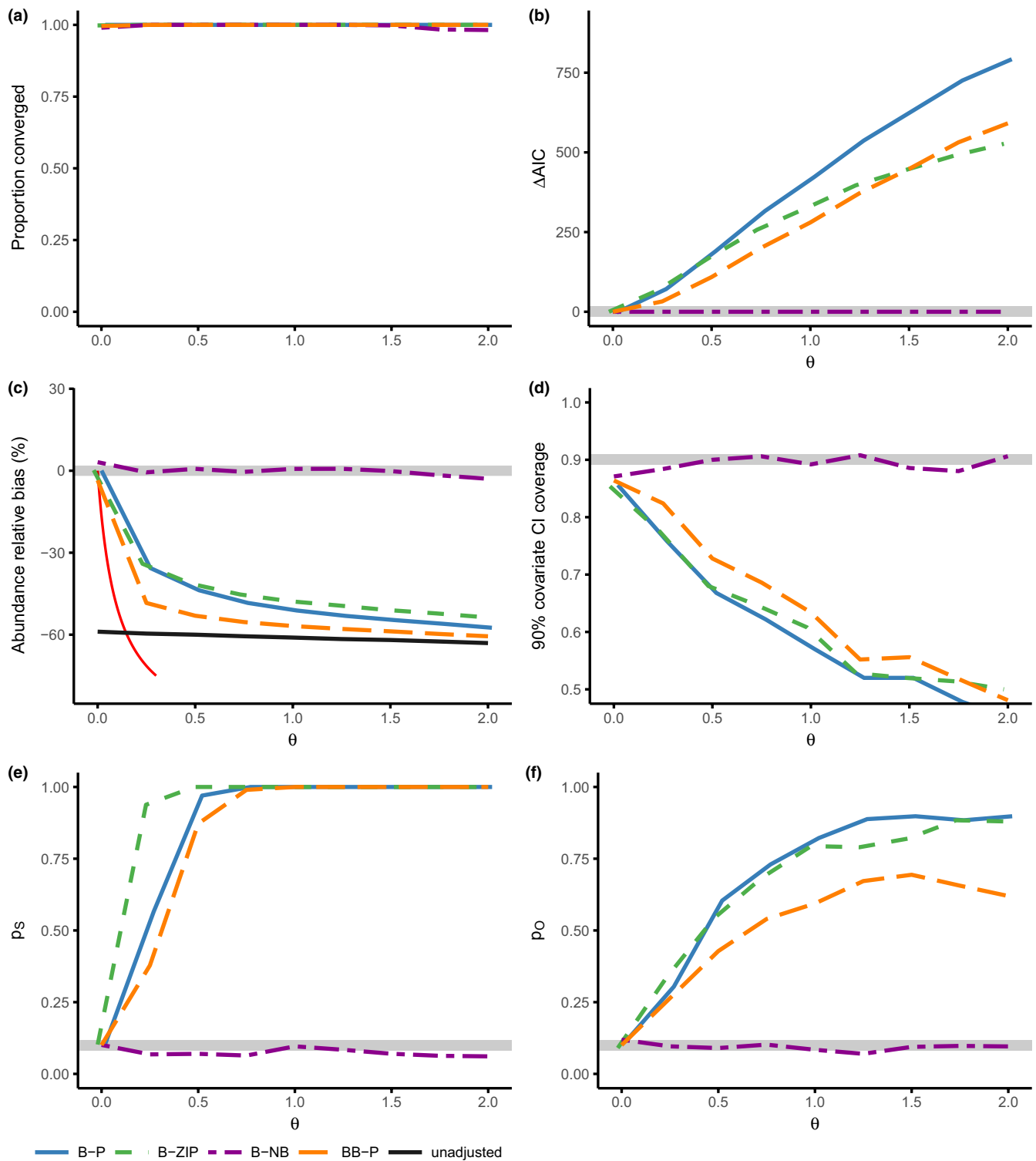
The B-NB model, when it converged, strongly overestimated abundance even for small amounts of overdispersion in detection (overestimating abundance by over 100% for overdispersion in detection above 0.15), while the B-P and B-ZIP models strongly overestimated abundance when the overdispersion in detection was larger (with over 100% for overdispersion above 0.25; Figure 5c). The correct beta-binomial Poisson model (BB-P) provided unbiased estimates. Confidence intervals for the spurious covariate had acceptable coverage for moderate overdispersion in detection but declined as overdispersion increased, except under the correct model (Figure 5d).

Normality tests of rq residuals failed to detect lack of fit for small to moderate overdispersion in detection. For large overdispersion in detection, the test of the observation rq residuals did often detect lack of fit and had better power than the test of the site-sum rq residuals (Figure 5e,f).

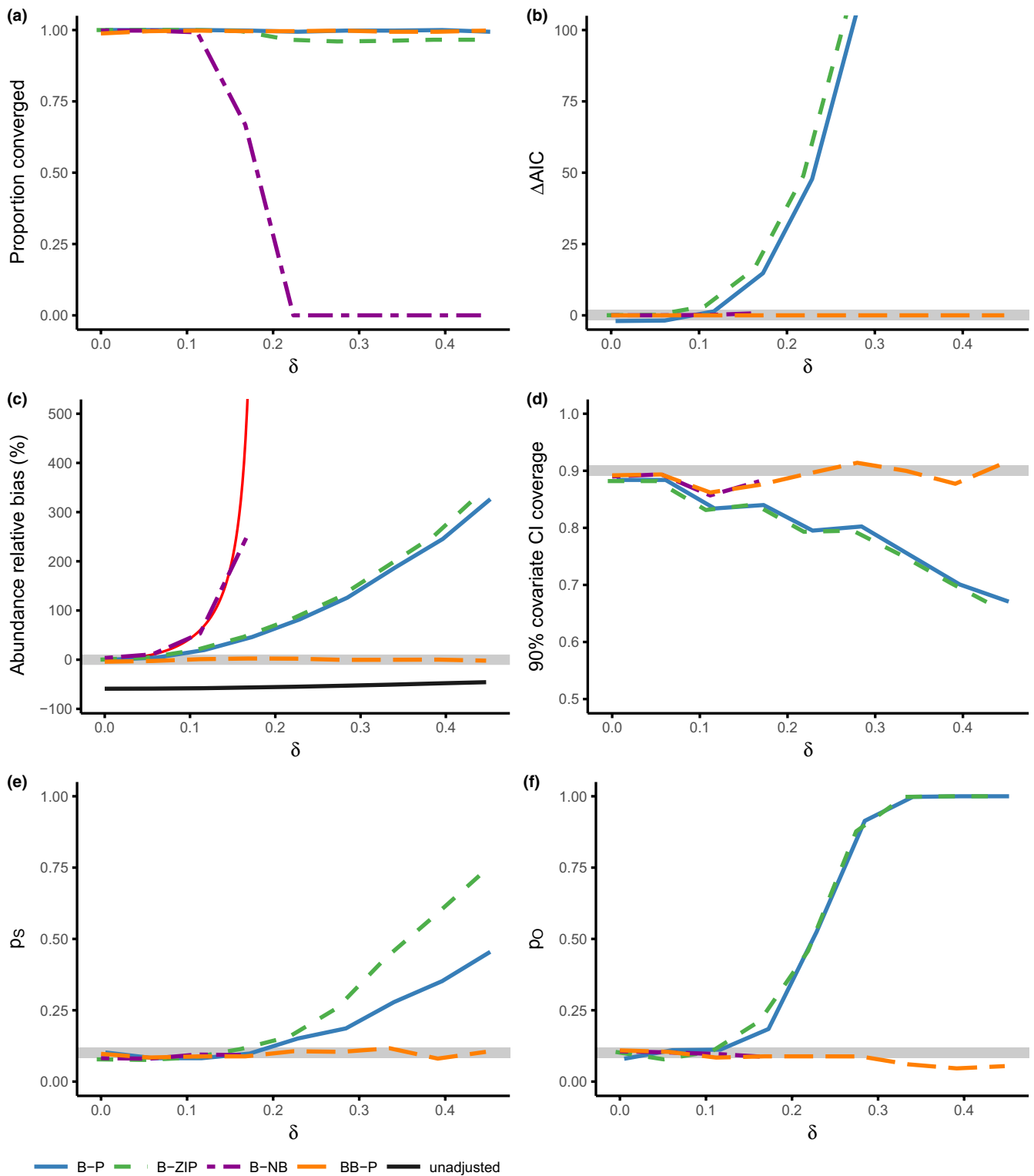
AIC had better performance in determining relative lack of fit of the B-P and B-ZIP model in relation to the true BB-P model, but was unable to distinguish between the B-NB model and the true model (Figure 5b).

Results for the simulation including covariates are similar but with less strong biases at similar levels of overdispersion (Supporting Information Figure S8).





**FIGURE 4** Results for binomial Poisson (B-P, blue), binomial ZIP (B-ZIP, green), binomial NB (B-NB, magenta), and beta-binomial Poisson (BB-P, orange) models fitted to data simulated from a negative binomial mixture with binomial detection (scenario 1) as a function of the overdispersion  $\theta$ . Grey lines give the reference level in each panel. (a) Proportion of simulations for which estimates were stable relative to the numerical cutoff  $K$  and for which the optimization routine converged. (b) Average difference in AIC between each model and the fitted correct B-NB model. (c) Relative bias in estimated mean abundance. Black line gives estimates not adjusted for imperfect detection. The red line gives the theoretical bias of the BB-P model by matching moments. (d) Proportion of Wald confidence intervals (90%) for the covariate effect that cover the true value (0). (e) Proportion of simulations for which a normality test (Shapiro) computed from site-sum  $r_q$  residuals was rejected at the 10% level. (f) Proportion of simulations for which a normality test (Shapiro) computed from observation  $r_q$  residuals was rejected at the 10% level



**FIGURE 5** Results for binomial Poisson (B-P, blue), binomial ZIP (B-ZIP, green), binomial NB (B-NB, magenta), and beta-binomial Poisson (BB-P, orange) models fitted to data simulated from a Poisson mixture with beta-binomial detection (scenario 2) as a function of the amount of overdispersion in detection  $\delta$ . Grey lines give the reference level in each panel. (a) Proportion of simulations for which estimates were stable relative to the numerical cutoff  $K$  and for which the optimization routine converged. (b) Average difference in AIC between each model and the fitted true BB-P model. (c) Relative bias in estimated mean abundance. Black line gives estimates not adjusted for imperfect detection. The red line gives the theoretical bias of the B-NB model by matching moments. (d) Proportion of Wald confidence intervals (90%) for the covariate effect that cover the true value (0). (e) Proportion of simulations for which a normality test (Shapiro) computed from site-sum  $r_q$  residuals was rejected at the 10% level. (f) Proportion of simulations for which a normality test (Shapiro) computed from observation  $r_q$  residuals was rejected at the 10% level

### 2.4.3 | Approximating the BB-P N-mixture model with a B-NB model

The inability of the goodness-of-fit checks to diagnose lack of fit of the B-NB model in scenario 2, the small difference in AIC between this model and the true BB-P model for moderate values of  $\delta$ , and the collapse at large values of  $\delta$ , can be understood through approximating the BB-P model with a B-NB model. Barker, Schofield, Link, and Sauer (2018) recently used moment matching to show that Poisson and negative binomial N-mixture models with a binomial detection model can be approximated by a double Poisson regression model, the latter lacking any notion of a latent abundance. Using moment matching, we show in Supporting Information Appendix S1 that an N-mixture model with beta-binomial detection and a Poisson abundance mixture can be approximated by another N-mixture model with binomial detection and a negative binomial abundance mixture where the abundance is inflated as long as  $\delta^2 < p/(\lambda - \lambda p)$ . In other words, data from a BB-P model will look identical to data from a B-NB model with higher abundance in terms of means, variances and covariances for such values. Because of this, it is difficult to distinguish between overdispersion in the detection probability and overdispersion in abundance.

## 3 | DISCUSSION

N-mixture models provide an appealing framework for learning about absolute rather than relative abundance of populations from count data alone, but this comes at the price of a very strong reliance on model assumptions. Count data by themselves contain only minimal information about absolute abundances (Knappe & Korner-Nievergelt, 2015; Barker et al., 2018) and our results, and some results of previous studies (Martin et al., 2011; Toribio, Gray, & Liang, 2012; Duarte et al., 2018), show that this leads to N-mixture models often being sensitive to even small amounts of model mis-specification. As a result, estimates of abundance and detection probability can be severely biased and inference about effects of covariates misleading if model assumptions are not met to a satisfactory degree. In light of this, finding a model that adequately fits the data is a necessity for reliable inferences about abundance using N-mixture models. The diagnostic tools proposed here are designed to evaluate the goodness of fit of N-mixture models, with a particular focus on overdispersion.

Our results show sensitivity of estimated abundances to overdispersion in the abundance mixture and, as previously shown (Martin et al., 2011), in the detection probability if the overdispersion is not accounted for. Not accounting for overdispersion in the abundance mixture leads to strong underestimation of absolute abundance while not accounting for random variation in the detection probability leads to strong overestimation of abundance, at overdispersion values commonly found in ecological count data. In our simulations, site-sum *r*<sub>q</sub> residuals were effective in detecting lack of fit caused by overdispersion in the abundance mixture unless overdispersion

was low. However, even relatively low levels of overdispersion lead to considerable underestimation of abundance. We found it more challenging to detect lack of fit due to overdispersion in detection. Lack of fit of a binomial detection model due to random variation in the detection probability among sites and visits was only reliably found at levels of overdispersion where bias in abundance was already large. *R*<sub>q</sub> residuals had no power to detect lack of fit of the negative binomial model even when abundance was overestimated by over 300%, but had some power to detect lack of fit of the binomial Poisson and ZIP models for high levels of overdispersion in detection. Like for lack of fit due to overdispersion in abundance, low levels of overdispersion can correspond to strong bias in estimated abundance.

Problems with detecting lack of fit due to variation in the detection probability occur despite a large sample size of 200 sites and 5 repeat visits in our simulation, and are not simply due to a poor choice of goodness of fit checks. The problems arise because of a fundamental similarity between alternative model structures for the same data. We show in Supporting Information Appendix S1 that the first- and second-order moments of the negative binomial N-mixture model can be matched exactly to the moments of a beta-binomial Poisson N-mixture model for small to intermediate variability in the probability of detection. This correspondence explains why detecting lack of fit is problematic for this model since higher order moments are needed to separate between them. Barker et al. (2018) used moment matching to show that data generated from a double Poisson model can be very similar to data from the binomial Poisson and negative binomial N-mixture models. Our results show that for overdispersed data, we do not need to go outside of the N-mixture framework to find alternative models that produce similar data but correspond to different inferences. This is concerning for the robustness of estimates of abundance using the N-mixture model, including the beta-binomial N-mixture model, as most abundance count data would be expected to contain at least some overdispersion (or sometimes underdispersion) in both abundance and detection, due, for example, to heterogeneity in spatial distribution of organisms, animal behaviour, or external conditions affecting detectability.

Overdispersion in detection in simulated data led to failure of the negative binomial N-mixture model such that it provided practically infinite estimates of abundance. This happened in our simulations when the moment matching of the negative binomial N-mixture model to the beta-binomial model suggested a negative detection probability. Thus, our results suggest that the negative binomial model can fail to provide finite estimates of abundance, a problem that has been commonly observed in case studies and in simulations (Dennis et al., 2015; Kéry & Royle, 2016), due to incorrectly attributing overdispersion to the among site variation in abundance. Recently, Kéry (2018), using bird data, found that infinite estimates occur frequently in negative binomial models. Since cases with infinite estimates of abundance can be diagnosed, he argued that when the negative binomial N-mixture model provides finite estimates of abundance, those estimates can be trusted. Our results, however, suggest that even when the negative binomial model produces finite estimates, a good fit, and low

AIC values, it can still strongly overestimate abundance (Figure 5). In cases where the negative binomial model produced infinite estimates of abundance Kéry (2018) found that it typically also had the lowest AIC value among the three models he fitted, indicating that those are cases where the data are overdispersed relative to Poisson and ZIP models. One strategy for dealing with infinite estimates from the negative binomial model would be to ignore them and resort to simpler models even if they have higher AIC values (Joseph et al., 2009). Given the potential for strongly biased estimates when overdispersion is ignored, this appears like a risky strategy.

The bias of the N-mixture model under mis-specification depends on parameter values. We used a moderately low detection probability ( $p = 0.25$ ) and a high abundance ( $\lambda = 10$ ) in our simulations. The moment matching suggests that if the detection probability is higher or abundances lower, the biases will be smaller and the N-mixture model more robust. The problem is that in practice these quantities are unknown. It seems tempting to rely on estimated detection probabilities and abundances from a fitted model to determine that one is in the parameter region where estimates are more robust. Unfortunately, it is clear from the simulations that such an approach is not reliable. In scenario 1, estimated detection probabilities under models ignoring overdispersion in abundance were much higher than the detection probabilities used to simulate the data (with detection probabilities estimated at up to 58%). Our suggestion is to instead fit multiple N-mixture models with and without overdispersion to the same data. In the parameter region where the N-mixture model is more robust, the different models are expected to provide similar estimates. In cases where the different models give similar abundances and fit the data well, the estimation issues discussed here may therefore be less of a problem.

The goodness-of-fit checks discussed here for binomial N-mixture models are easily extended to multinomial N-mixture models (Kéry & Royle, 2016). Site-sum  $r_q$  residuals and overdispersion metrics (Supporting Information Appendix S1) may, for example, be defined for the sum of counts over all the observed categories of the multinomial. In distance sampling, this equates to the total number of individuals detected across all distances at each site (Johnson, Laake, & Ver Hoef, 2010).

### 3.1 | Conclusions

Ill fitting binomial N-mixture models are highly likely to provide misleading estimates of abundance, detection and effects of covariates. In some cases, abundance estimates from a poorly fitting model may be more far off than estimates that are not corrected for nondetection. Thorough assessment of goodness of fit through multiple checks should therefore become a standard part of their use. The goodness-of-fit checks presented here can be used to this end and are available in an R-package `nmixgof` compatible with `unmarked`.

Estimates from models with an apparently good fit may still be highly sensitive to the accuracy of the distributional assumptions of the model. This is particularly the case if data

are overdispersed relative to the basic Poisson or ZIP models with binomial detection and can lead to alternative N-mixture model specifications providing different abundance estimates but similar fit. How useful abundance estimates from well fitting N-mixture models are, therefore, to some extent comes down to how much faith one has in that the parametric assumptions of the fitted model reflect reality. Our view is in agreement with Barker et al. (2018) that such a strong reliance on a particular model specification makes inferences about abundance dubious. Without additional information about detection probabilities, count data are often better suited to estimate indices of relative abundance rather than absolute abundance (Link & Sauer, 1997; Knappe, 2016). If estimates of absolute abundance are necessary, our recommendation is to complement the detection submodel by additional information or, if none are available, to carefully assess goodness of fit and to investigate robustness by checking the similarity among inferences from multiple N-mixture model specifications.

### ACKNOWLEDGEMENTS

We thank Marc Kéry and anonymous reviewers for providing valuable comments. J.K. was funded by grant 621-2012-4076 from the Swedish Research Council VR and D.A. by grant 942-2015-287 from FORMAS. The wetland survey was funded by grants from FORMAS and the Swedish EPA to T.P. F.B. was supported by LabEx COTE (ANR-10-LABX-45).

### AUTHORS' CONTRIBUTIONS

J.K. planned the study, performed analyses and wrote the first draft. Å.B. digitized the Northern shoveler data. All authors contributed to revisions.

### DATA ACCESSIBILITY

The `nmixgof` package is available from CRAN (<https://cran.r-project.org/package=nmixgof>) and github (<https://github.com/jknape/nmixgof>). The Northern shoveler data are archived at Zenodo, <http://doi.org/10.5281/zenodo.1303825>, and are also included in the `nmixgof` package.

### ORCID

Jonas Knappe  <http://orcid.org/0000-0002-8012-5131>

Frédéric Barraquand  <http://orcid.org/0000-0002-4759-0269>

Alejandro Ruete  <http://orcid.org/0000-0001-7681-2812>

### REFERENCES

- Barker, R. J., Schofield, M. R., Link, W. A., & Sauer, J. R. (2018). On the reliability of N-mixture models for count data. *Biometrics*, 74, 369–377. <https://doi.org/10.1111/biom.12734>

- Dennis, E. B., Morgan, B. J., & Ridout, M. S. (2015). Computational aspects of N-mixture models. *Biometrics*, 71, 237–246. <https://doi.org/10.1111/biom.12246>
- Duarte, A., Adams, M. J., & Peterson, J. T. (2018). Fitting N-mixture models to count data with unmodeled heterogeneity: Bias, diagnostics, and alternative approaches. *Ecological Modelling*, 374, 51–59. <https://doi.org/10.1016/j.ecolmodel.2018.02.007>
- Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5, 236–244. <https://doi.org/10.1080/10618600.1996.10474708>
- Fiske, I., & Chandler, R. (2011). Unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*, 43, 1–23. <https://doi.org/10.18637/jss.v043.i10>
- Haines, L. M. (2016). Maximum likelihood estimation for N-mixture models. *Biometrics*, 72, 1235–1245. <https://doi.org/10.1111/biom.12521>
- Hunter, E. A., Nibbelink, N. P., & Cooper, R. J. (2017). Divergent forecasts for two salt marsh specialists in response to sea level rise. *Animal Conservation*, 20, 20–28. <https://doi.org/10.1111/acv.12280>
- Johnson, D. S., Laake, J. L., & Ver Hoef, J. M. (2010). A model-based approach for making ecological inference from distance sampling data. *Biometrics*, 66, 310–318. <https://doi.org/10.1111/j.1541-0420.2009.01265.x>
- Joseph, L., Elkin, C., Martin, T., & Possingham, H. (2009). Modeling abundance using N-mixture models: The importance of considering ecological mechanisms. *Ecological Applications*, 19, 631–642. <https://doi.org/10.1890/07-2107.1>
- Kéry, M. (2018). Identifiability in N-mixture models: A large-scale screening test with bird data. *Ecology*, 99, 281–288. <https://doi.org/10.1002/ecy.2093>
- Kéry, M., & Royle, J. A. (2016). *Applied hierarchical modeling in ecology*. Boston, MA: Academic Press.
- Kéry, M., Royle, J., & Schmid, H. (2005). Modeling avian abundance from replicated counts using binomial mixture models. *Ecological Applications*, 15, 1450–1461. <https://doi.org/10.1890/04-1120>
- Knape, J. (2016). Decomposing trends in Swedish bird populations using generalized additive mixed models. *Journal of Applied Ecology*, 53, 1852–1861. <https://doi.org/10.1111/1365-2664.12720>
- Knape, J., & Korner-Niervergelt, F. (2015). Estimates from non-replicated population surveys rely on critical assumptions. *Methods in Ecology and Evolution*, 6, 298–306. <https://doi.org/10.1111/2041-210x.12329>
- Knape, J., & Korner-Niervergelt, F. (2016). On assumptions behind estimates of abundance from counts at multiple sites. *Methods in Ecology and Evolution*, 7, 206–209. <https://doi.org/10.1111/2041-210x.12507>
- Ladin, Z. S., D'Amico, V., Baetens, J. M., Roth, R. R., & Shriver, W. G. (2016). Predicting metapopulation responses to conservation in human-dominated landscapes. *Frontiers in Ecology and Evolution*, 4, 122. <https://doi.org/10.3389/fevo.2016.00122>
- Leask, K. L., & Haines, L. M. (2014). The beta-Poisson distribution in Wadley's problem. *Communications in Statistics - Theory and Methods*, 43, 4962–4971. <https://doi.org/10.1080/03610926.2012.744047>
- Linden, A., & Mäntyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, 92, 1414–1421. <https://doi.org/10.1890/10-1831.1>
- Link, W. A., & Sauer, J. R. (1997). Estimation of population trajectories from count data. *Biometrics*, 53, 488–497. <https://doi.org/10.2307/2533952>
- Link, W. A., Schofield, M. R., Barker, R. J., & Sauer, J. R. (2018). On the robustness of N-mixture models. *Ecology*, 99, 1547–1551.
- Loy, A., Follett, L., & Hofmann, H. (2016). Variations of q-q plots: The power of our eyes!. *The American Statistician*, 70, 202–214. <https://doi.org/10.1080/00031305.2015.1077728>
- Martin, J., Royle, J. A., Mackenzie, D. I., Edwards, H. H., Kéry, 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. <https://doi.org/10.1111/j.2041-210x.2011.00113.x>
- Richards, S. A. (2008). Dealing with overdispersed count data in applied ecology. *Journal of Applied Ecology*, 45, 218–227. <https://doi.org/10.1111/j.1365-2664.2007.01377.x>
- Romano, A., Costa, A., Basile, M., Raimondi, R., Posillico, M., Scinti Roger, D., ... De Cinti, B. (2017). Conservation of salamanders in managed forests: Methods and costs of monitoring abundance and habitat selection. *Forest Ecology and Management*, 400, 12–18. <https://doi.org/10.1016/j.foreco.2017.05.048>
- Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics*, 60, 108–115. <https://doi.org/10.1111/j.0006-341x.2004.00142.x>
- Royle, J. A., & Dorazio, R. M. (2006). Hierarchical models of animal abundance and occurrence. *Journal of Agricultural, Biological, and Environmental Statistics*, 11, 249–263. <https://doi.org/10.1198/108571106x129153>
- Studds, C. E., Kendall, B. E., Murray, N. J., Wilson, H. B., Rogers, D. I., Clemens, R. S., ... Fuller, R. A. (2017). Rapid population decline in migratory shorebirds relying on Yellow Sea tidal mudflats as stopover sites. *Nature Communications*, 8, 14895. <https://doi.org/10.1038/ncomms14895>
- Toribio, S. G., Gray, B. R., & Liang, S. (2012). An evaluation of the Bayesian approach to fitting the N-mixture model for use with pseudo-replicated count data. *Journal of Statistical Computation and Simulation*, 82, 1135–1143. <https://doi.org/10.1080/00949655.2011.572881>
- ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88, 2766–2772. <https://doi.org/10.1890/07-0043.1>
- Warton, D. I., Lyons, M., Stoklosa, J., & Ives, A. R. (2016). Three points to consider when choosing a LM or GLM test for count data. *Methods in Ecology and Evolution*, 7, 882–890. <https://doi.org/10.1111/2041-210x.12552>
- Warton, D. I., Stoklosa, J., Guillera-Aroita, G., MacKenzie, D. I., & Welsh, A. H. (2017). Graphical diagnostics for occupancy models with imperfect detection. *Methods in Ecology and Evolution*, 8, 408–419. <https://doi.org/10.1111/2041-210x.12761>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

**How to cite this article:** Knape J, Arlt D, Barraquand F, et al. Sensitivity of binomial N-mixture models to overdispersion: The importance of assessing model fit. *Methods Ecol Evol*. 2018;9:2102–2114. <https://doi.org/10.1111/2041-210X.13062>