Statistical Reports

Ecology, 99(7), 2018, pp. 1547–1551

On the robustness of N-mixture models

William A. Link,[1,3] Matthew R. Schofield,[2] Richard J. Barker,[2] and John R. Sauer[1]

[1]*USGS Patuxent Wildlife Research Center, Laurel, Maryland 20708 USA*
[2]*Department of Mathematics and Statistics, University of Otago, Dunedin, New Zealand*

Abstract. *N*-mixture models provide an appealing alternative to mark–recapture models, in that they allow for estimation of detection probability and population size from count data, without requiring that individual animals be identified. There is, however, a cost to using the *N*-mixture models: inference is very sensitive to the model's assumptions. We consider the effects of three violations of assumptions that might reasonably be expected in practice: double counting, unmodeled variation in population size over time, and unmodeled variation in detection probability over time. These three examples show that small violations of assumptions can lead to large biases in estimation. The violations of assumptions we consider are not only small qualitatively, but are also small in the sense that they are unlikely to be detected using goodness-of-fit tests. In cases where reliable estimates of population size are needed, we encourage investigators to allocate resources to acquiring additional data, such as recaptures of marked individuals, for estimation of detection probabilities.

Key words: *abundance estimation; Bayesian* P*-value; count data; detection probability;* N*-mixture model; robustness.*

Introduction

The N-mixture model of Royle (2004) and its extensions (e.g., Dail and Madsen 2011) have great appeal to field biologists in that they provide a means for estimating detection probability and population size without the expense, inconvenience, and difficulty of mark–recapture, distance sampling, or other methods of estimation historically used to estimate detection rates. The model is easy to describe and implement, and can be fit to a wide variety of field sampling studies in which data were collected from repeated counts at a series of sampling sites. Counts Y_{ij} are assumed to be conditionally independent binomial random variables with success parameters p_{ij} (parametric functions of observed covariates) and indices N_{ij}; in symbols, $Y_{ij}|p_{ij} \sim B(N_{ij}, p_{ij})$. Typically, index $i = 1, 2, \ldots, n$ distinguishes sites, and index $j = 1, 2, \ldots, r$ distinguishes repeated counts at sites. The values N_{ij} are interpreted as the number of individuals present at site i and time j, and the p_{ij} as associated detection probabilities.

The basic model (Royle 2004) has $N_{ij} \equiv N_i$, where N_i are Poisson random variables with rate parameters λ, denoted $N_i \sim \text{Pois}(\lambda)$. The basic model also assumes constant detection probabilities $p_{ij} \equiv p$. Elaborations on the basic model are straightforward, including alternative distributions for population sizes, and covariate models for detection probabilities.

This methodological simplicity has led to widespread use of N-mixture models by ecologists. As of March 2018, Royle (2004) and Dail and Madsen (2011) have received 747 and 182 citations, respectively. N-mixture based approaches have been suggested as alternative design and analysis models for surveys such as the North American Breeding Bird Survey (Riddle et al. 2010, Hostetler and Chandler 2015).

In this paper, we examine the sensitivity of N-mixture models to violations of assumptions. We stress from the outset that, in principle, N-mixture models work: if the model is correct (an *exact* depiction of the data generating process), if the parameter values are not extreme, and if we have adequate data, the N-mixture models can be estimated with high precision. Though we note with concern a tendency toward excessive optimism about what constitutes adequate data and reasonable parameter values, our focus in this note is on the sensitivity of inference to violations of fundamental model assumptions.

We use the basic N-mixture model (with constant λ and p) as the basis of our evaluations. We present three examples in which data are generated with slight violations of the model's assumptions. In the first, we allow the possibility that individual animals may be accidentally counted twice, violating the binomial assumption. In the second, we allow violation of the constant abundance assumption within sites; in the third, we allow violation of the constant detection probability assumption.

In each case we assess goodness of fit under the basic N-mixture model. The examples we present show that even slight violations of model assumptions can lead to profound biases in estimation. Large departures from model assumptions might be caught by goodness-of-fit testing. However, many of the departures we consider are small enough to evade testing, yet still lead to substantial bias.

These results have implications beyond the basic model. While it may be possible to model some of the departures from the basic model, it may also be impossible to know that

Manuscript received 8 November 2017; revised 28 March 2018; accepted 6 April 2018. Corresponding Editor: Brett T. McClintock.
[3] E-mail: wlink@usgs.gov

a more complex model is needed, or which one is needed. If the basic model cannot be relied upon, neither can its many elaborations. Our conclusion is that for estimating absolute abundance, there is no substitute for mark–recapture analysis: N-mixture modeling relies too heavily on questionable and poorly verifiable model assumptions.

Methods

Baseline model and model fitting

As a shared baseline for our examples, we generated data sets with $n = 50, \lambda = 100$, and $p = 0.42$. For this and all subsequent simulations we analyzed 1,000 replicate data sets with $r = 3$, and 1,000 with $r = 6$.

The sample sizes and parameter values considered in this and the following illustrations may seem generous, especially when compared to values deemed sufficient by some authors. Yamaura et al. (2016) suggest $r = 2, 3$, or 4, $n \geq 20$, $p \geq 0.20$, and $\lambda \geq 0.50$ as "minimal conditions for obtaining adequate performance of community abundance models." Our choice of larger parameter values and more substantial sample sizes was intended to provide favorable circumstances for fitting the simplest N-mixture model, and a reasonable context for evaluating model violations that might be anticipated in practice.

We fit the N-mixture model using Bayesian analysis with vague priors $[\lambda] \propto 1/\lambda$ and $p \sim U(0,1)$. Markov chain Monte Carlo was used to evaluate posterior distributions. MCMC based on full conditionals for λ and p leads to poor mixing, so we used a Metropolis-Hastings algorithm with candidate generation designed to approximate hierarchically centered analysis; details are given in Appendix S1. Mixing was good, with autocorrelations near zero at lag 40; we used chains of length 100,000.

Because posterior distributions for the abundance parameter tend to be positively skewed, we used the posterior median as a point estimator, rather than the posterior mean. The biases we report are positive in all cases, and would be even larger if the posterior mean were used as a point summary. 95% credible intervals were defined by the 2.5th and 97.5th percentiles of posterior distributions.

We also analyzed each data set using package *unmarked* in R (Fiske and Chandler 2011) in order to compare Bayesian results with frequentist. Evaluation of 95% confidence intervals was based on assumptions of asymptotic normality.

Goodness of fit

We used calibrated Bayesian *P*-values to assess goodness of fit. For each posterior sample $(p^{(b)}, N_i^{(b)})$, $b = 1, 2, \ldots, 100,000$ we generated data $Y_{ij}^{(b)} \sim B(N_i^{(b)}, p^{(b)})$, then calculated

$$T^{(b)} = \sum_i \sum_j \frac{(Y_{ij}^{(b)} - N_i^{(b)} p^{(b)})^2}{N_i^{(b)} p^{(b)} (1 - p^{(b)})}.$$

The Bayesian *P*-value, P_B, is the proportion of samples for which $T^{(b)}$ exceeds the value of the same statistic, computed with actual data Y_{ij} rather than $Y_{ij}^{(b)}$. Tests based on Bayesian *P*-values are often conservative, favoring the null hypothesis (Nott et al. 2018). We thus calculated calibrated Bayesian *P*-values P_B^* designed to overcome this conservatism, which we now describe.

Under the null hypothesis, the *P*-value of a continuous test statistic should have a uniform distribution on (0,1). That is, under the null hypothesis, $\Pr(P \leq t) = t$, for all $t \in (0, 1)$. In many cases, Bayesian *P*-values have null distributions that are underdispersed relative to the uniform distribution. Thus we cannot assume that $F(t) = \Pr(P_B \leq t) \equiv t$. We used our baseline simulations to obtain estimates $\hat{F}(t)$. In subsequent simulations, we calculated calibrated Bayesian *P*-values as $P_B^* = \hat{F}(P_B)$. For a fixed level $\alpha = 0.05$ test, determining whether $P_B^* < 0.05$ is accomplished by comparing the observed value P_B to the fifth percentile of Bayesian *P*-values from the baseline simulation. This process of calibrating the Bayesian *P*-value is precisely in the spirit of Hjort et al. (2006), though with the advantage (conferred by application in context of simulations) of having known parameters under the null hypothesis. The use of P_B^* boosts power to detect violations of the null hypothesis while maintaining the appropriate α level of the test.

We chose to use a Bayesian evaluation of goodness of fit to avoid uncertainties arising from use of asymptotic frequentist methods. For instance, during the course of our analyses we encountered evidence that frequentist CI coverage rates actually increased from nominal levels under slight violations of model assumptions. This suggests a weakness of the asymptotic methods. Our message, however, is about the N-mixture models themselves, rather than the methods used to fit them. Rather than getting bogged down in the technical issues, we used Bayesian methods as the most reliable.

Violations of model assumptions

We now turn our attention to three sets of simulations in which the assumptions of the N-mixture model are violated.

Simulation 1: violation of binomial assumption through accidental double counting.—As noted by Barker et al. (2017) the absence of marks on animals may make it impossible to know whether an animal has been counted more than once on a single sampling occasion. Here we examine the consequences of accidental double counts.

N-mixture modeling assumes two possible outcomes per animal on each sampling occasion: the animal is not detected, or detected and recorded as a single animal. Suppose instead that there are three possible outcomes: the animal is not detected (with probability α), detected and correctly recorded as a single animal (with probability β), or detected and incorrectly recorded as two distinct animals (with probability γ). Our intuition might be that if γ is small, this departure from the binomial assumption will be of little consequence.

We generated data with $n = 50$, $r = 3$ and 6, $\lambda = 100$, $\alpha = 0.58$, and γ ranging from 0.02 to 0.10 in increments of 0.02. In each case, the probability an animal is observed is $1 - \alpha = 0.42$, as in the baseline simulation. The baseline cases in which the assumptions of the N-mixture model are met can be described as having $\gamma = 0$.

Simulation 2: unmodeled variation in N_{ij}.—The basic model of Royle (2004) posits $N_{ij} \equiv N_i \sim \text{Pois}(\lambda)$. Suppose instead that

$N_{ij} = \nu_i + A_{ij}$ where $\nu_i \sim \text{Pois}(\theta_1)$ and independently $A_{ij} \sim \text{Pois}(\theta_2)$; the mean abundance is $\lambda = \theta_1 + \theta_2$. The number of animals available for counting at site i varies across "replicates" j. Such variation could result from animal movement between sampling occasions. There is, nonetheless, a common site effect (baseline abundance) so N_{ij} are positively and equi-correlated, with correlation $\rho = \theta_1/(\theta_1 + \theta_2) = \theta_1/\lambda$. The parameter ρ can also be understood as the proportion of variation in population sizes that occurs among sites. The basic N-mixture model assumes that all of this variation occurs among sites; our simulations allow $100(1-\rho)\%$ of this variation to occur within sites, among sampling occasions.

We generated data with $n = 50$, $r = 3$ and 6, $p = 0.42$, $\lambda = 100$, and $\rho = 0.90, 0.80, 0.50, 0.20$, and 0.10. The baseline cases in which the assumptions of the N-mixture model are met have $\rho = 1$.

Simulation 3: unmodeled variation in p_{ij}.—The basic model of Royle (2004) posits $p_{ij} \equiv p$, for $j = 1, 2, \ldots, r$. Suppose instead that p_{ij} are random variables with mean π and variance $\sigma^2 > 0$. Variation in detection probabilities could result from myriad environmental factors; the notion of a constant p is an almost obvious fiction. We generated data with $n = 50$, $r = 3$ and 6, $\lambda = 100$, and $p_{ij} \sim \beta(a, b)$, with parameters a and b chosen so that $\pi = a/(a+b) = 0.42$ and $\sigma = \sqrt{\pi(1-\pi)/(a+b+1)} = 0.01, 0.02, 0.03$, and 0.04. The baseline cases in which the assumptions of the N-mixture model are met can be described as having $\sigma = 0$.

Results

Results for Simulations 1, 2, and 3 are given in Tables 1, 2, and 3, respectively. In the baseline simulations, an N-mixture model was the data generating model. Results for these are included for comparison with each set of simulation results; these are the cases with $\gamma = 0$, $\rho = 1$, and $\sigma = 0$, respectively.

Across all of our simulations, Bayesian and frequentist point estimates of λ were basically in agreement. When the N-mixture model was the data generating model, nominal 95% credible intervals had coverage rates of 95.1% and 95.6%; nominal 95% confidence intervals had coverage rates of 95.2% and 94.2%. Subsequently, we restrict our discussion to Bayesian estimates.

We note that there is some bias in the estimation of λ even in the baseline simulations. This may well be small-sample bias, as it is smaller for the case $r = 6$ than for $r = 3$.

Simulation 1: violation of binomial assumption through accidental double counting.—Estimation of λ is badly biased when $\gamma > 0$. Consider the case $\gamma = 0.02$. Making allowances for the small sample bias of the baseline case (with $\gamma = 0$), the additional bias due to model misspecification is $(125.5/106.5 - 1)100\% = 18\%$ with $r = 3$; with $r = 6$ the additional bias is 21%. This substantial bias arises with only a slight violation of model assumptions: with $\gamma = 0.02$ less than one in twenty animals observed is accidentally counted twice. With $\gamma = 0.04$, the rate of double counts is still low, and the power to detect model inadequacy is low, but the bias due to model misspecification is 34% for $r = 3$ and 39% for $r = 6$.

Misspecification bias is large relative to nominal precision, even in cases where there is little power to detect model misspecification. Thus we might conclude that the null model is satisfactory in cases where credible interval coverage is substantially lower than nominal. With $r = 3$ and $\gamma = 0.04$, there is only a 40.7% chance of rejecting the adequacy of the N-mixture model, and the credible interval coverage is only 56.2%; with $r = 6$ and $\gamma = 0.04$, the probability of rejecting the N-mixture model increases to 63.0%, but the credible interval coverage drops to 16.7%.

For larger values of γ, a test of significance might be useful to screen for model inadequacy. We were interested in knowing whether such screening would reduce the bias of $\hat{\lambda}$, so we computed $\hat{\lambda}^{NS}$, the mean value of the estimator in cases where the model would be deemed acceptable. The bias of estimators remains when the model is not rejected (see Table 1).

Simulation 2: unmodeled variation in N_{ij}.—Results for Simulation 2, in which there is unmodeled variation in population

TABLE 1. Effects of accidental double counts on estimation of mean population abundance.

r	γ	λ_{MLE}	$\hat{\lambda}$	SD ($\hat{\lambda}$)	Nom 95%	$P_B < 0.005$	$P_B^* < 0.05$	$\hat{\lambda}^{NS}$
3	0.00	105.1	106.5	21.8	0.951	0.000	0.050	105.9
3	0.02	124.9	125.5	26.4	0.836	0.008	0.187	124.1
3	0.04	142.6	143.1	32.2	0.562	0.056	0.407	140.7
3	0.06	158.8	159.0	34.9	0.318	0.162	0.657	158.5
3	0.08	176.8	176.7	36.3	0.120	0.290	0.811	170.4
3	0.10	191.5	193.5	40.4	0.044	0.498	0.891	197.4
6	0.00	101.3	101.7	12.4	0.956	0.000	0.050	101.6
6	0.02	122.0	122.1	16.4	0.647	0.004	0.325	120.6
6	0.04	141.9	141.5	19.9	0.167	0.047	0.630	139.1
6	0.06	159.9	159.0	23.1	0.020	0.193	0.851	161.3
6	0.08	178.6	177.2	26.8	0.003	0.480	0.949	170.0
6	0.10	193.7	192.4	29.1	0.000	0.692	0.988	183.3

Notes: A total 1,000 data sets simulated for each of 12 parameter and data configurations. Data consisted of r counts at $n = 50$ sites, with $N_i \sim \text{Pois}(\lambda = 100)$ animals at site i. Individuals counted 0, 1, or 2 times, with probabilities 0.58, $0.42-\gamma$ and γ; $\gamma = 0$ is the baseline case, with data generated according to the basic N-mixture model. Columns labeled λ_{MLE}, $\hat{\lambda}$, and SD($\hat{\lambda}$) give the means of the maximum likelihood estimator, the posterior median and posterior standard deviation, respectively. Column "Nom 95%" gives the coverage rate of nominal 95% credible intervals. Columns labeled $P_B < 0.05$ and $P_B^* < 0.05$ give the proportion of times tests based on the usual and calibrated Bayesian P-values reject the adequacy of the N-mixture model, at nominal level $\alpha = 0.05$. Column $\hat{\lambda}^{NS}$ corresponds to column $\hat{\lambda}$, but with the mean taken only over simulations where $P_B^* > 0.05$.

TABLE 2. Effects of unmodeled variation in abundance within sites.

r	ρ	λ_{MLE}	$\hat{\lambda}$	SD($\hat{\lambda}$)	Nom 95%	$P_B < 0.005$	$P_B^* < 0.05$	$\hat{\lambda}^{NS}$
3	1.00	105.1	106.5	21.8	0.951	0.000	0.050	105.9
3	0.90	119.9	120.6	26.3	0.890	0.000	0.064	119.9
3	0.80	137.6	137.5	33.5	0.777	0.004	0.089	136.1
3	0.50	232.6	234.8	57.6	0.242	0.014	0.160	230.5
3	0.20	741.7	491.5	96.2	0.029	0.034	0.237	462.2
3	0.10	918.0	617.2	108.5	0.011	0.041	0.243	580.0
6	1.00	101.3	101.7	12.4	0.956	0.000	0.050	101.6
6	0.90	114.6	114.9	16.2	0.843	0.000	0.072	114.6
6	0.80	130.5	130.2	21.2	0.582	0.000	0.121	129.8
6	0.50	222.6	218.2	46.7	0.024	0.003	0.201	218.3
6	0.20	670.7	565.9	102.1	0.000	0.033	0.292	559.0
6	0.10	1019.0	1015.2	129.4	0.000	0.049	0.323	993.0

Notes: A total of 1,000 data sets simulated for each of 12 parameter and data configurations. Data consisted of r counts at $n = 50$ sites, with $N_{ij} = v_i + A_{ij}$ animals at site i, occasion j, with $v_i \sim \text{Pois}(\theta_1)$ and $A_{ij} \sim \text{Pois}(\theta_2)$. $\lambda = \theta_1 + \theta_2 = 100$, and $\rho = \theta_1/\lambda$. $\rho = 1$ is the baseline case, with data generated according to the basic N-mixture model. Data are counts $Y_{ij} \sim B(N_{ij}, p)$ with $p = 0.42$. Columns labeled λ_{MLE}, $\hat{\lambda}$ and SD($\hat{\lambda}$), give the means of the maximum likelihood estimator, the posterior median, and posterior standard deviation, respectively. Column "Nom 95%" gives the coverage rate of nominal 95% credible intervals. Columns labeled $P_B < 0.05$ and $P_B^* < 0.05$ give the proportion of times tests based on the usual and calibrated Bayesian P-values reject the adequacy of the N-mixture model, at nominal level $\alpha = 0.05$. Column $\hat{\lambda}^{NS}$ corresponds to column $\hat{\lambda}$, but with the mean taken only over simulations where $P_B^* > 0.05$.

TABLE 3. Effects of unmodeled variation in detection probability.

r	σ	λ_{MLE}	$\hat{\lambda}$	SD($\hat{\lambda}$)	Nom 95%	$P_B < .005$	$P_B^* < 0.05$	$\hat{\lambda}^{NS}$
3	0.00	105.1	106.5	21.8	0.951	0.000	0.050	105.9
3	0.01	110.4	111.6	24.1	0.952	0.001	0.066	110.7
3	0.02	126.2	126.5	29.1	0.846	0.013	0.219	125.3
3	0.03	154.8	154.2	37.5	0.552	0.131	0.584	147.6
3	0.04	200.8	199.6	48.2	0.172	0.579	0.917	200.0
6	0.00	101.3	101.7	12.4	0.956	0.000	0.050	101.6
6	0.01	107.0	107.3	13.6	0.921	0.000	0.113	106.9
6	0.02	123.4	123.3	17.6	0.655	0.003	0.324	121.8
6	0.03	153.7	152.4	24.4	0.131	0.187	0.808	151.7
6	0.04	201.9	198.6	34.7	0.002	0.844	0.992	209.0

Notes: A total of 1,000 data sets simulated for each of 10 parameter and data configurations. Data consisted of r counts at $n = 50$ sites, with $N_{ij} \sim \text{Pois}(\lambda = 100)$ animals at site i. Data are counts $Y_{ij} \sim B(N_i, p_{ij})$; p_{ij} are beta-distributed random variables with mean $\pi = 0.42$ and standard deviation σ. $\sigma = 0$ is the baseline case, with data generated according to the basic N-mixture model. Columns labeled λ_{MLE}, $\hat{\lambda}$ and SD($\hat{\lambda}$) give the means of the maximum likelihood estimator, the posterior median, and posterior standard deviation, respectively. Column "Nom 95%" gives the coverage rate of nominal 95% credible intervals. Columns labeled $P_B < 0.05$ and $P_B^* < 0.05$ give the proportion of times tests based on the usual and calibrated Bayesian P-values reject the adequacy of the N-mixture model, at nominal level $\alpha = 0.05$. Column $\hat{\lambda}^{NS}$ corresponds to column $\hat{\lambda}$, but with the mean taken only over simulations where $P_B^* > 0.05$.

sizes at sites, are the most troubling of our three simulations. Once again, estimation of λ is badly biased when the N-mixture model assumption has been even slightly violated ($\rho < 1$). Coverage rates for nominal 95% CIs drop off quickly and bias increases rapidly as ρ decreases. The Bayesian P-value, even in its calibrated version, shows little ability to detect even the most extreme violations of model assumptions. Even with as much variation between visits as among sites ($\rho = 0.5$), the power to detect inadequacy of the baseline model is only 16% and 20% with $r = 3$ and 6, respectively; in these cases, the expected value of the estimator is more than twice the true value.

Simulation 3: unmodeled variation in p_{ij}.—Our final set of simulations addressed unmodeled variation in detection probability. The levels of variability in these simulations were small. With $\sigma = 0.02$, more than 80% of p_{ij}s fall in the range (0.394,0.446). Nevertheless, $\sigma = 0.02$ is sufficient to produce 19% and 21% additional bias (beyond small sample bias) in estimation of λ, for $r = 3$ and 6, respectively. At the same time, these configurations of parameters provide little power to distinguish the data from such as are generated assuming the N-mixture model is correctly specified.

DISCUSSION

It is (almost) never the case that a mathematical model is a perfect depiction of reality. Often, we are able to say "close enough" and not worry too much. The simulations reported here indicate that such is not the case for N-mixture models. The model specification $Y_{ij} \sim B(N_i, p)$ has three parts: B for binomial, N_i for constant abundance by site, and p for constant detectability. Our simulations address violations of assumption B, N_i, or p. Small, undetectable violations of any of these can lead to substantial biases. A 2% rate of double counts or a standard deviation of 2% in detection rates might seem insubstantial, but each is large enough to produce >20% bias in estimation of mean abundance. To make

matters worse, there is little power to detect such violations of model assumptions through goodness-of-fit tests.

Our attention to the N-mixture models is prompted by their obvious and critical dependence on assumptions in place of data. There is no such thing as a free lunch: extra data have been replaced with extra assumptions, and the assumptions are stringent. Small, undetectable violations of assumptions lead to substantial biases. Similar concerns regarding N-mixture models are being expressed by other authors (Duarte et al. 2018, Knape et al. In Review).

A natural response to the findings reported here is to suggest that one can always build a bigger model. If violating an assumption leads to biases, well, construct a bigger model that describes those violations, and you're all set. But this requires an idea of which assumptions are being violated, and the capacity to detect failures of the base model before the biases become too large. We have not addressed covariate models or other more highly structured models, such as the Dail-Madsen model (Dail and Madsen 2011) in our examples. It seems reasonable to assume that the risks of model misspecification increase as model complexity increases; robustness isn't increased by complexity.

The sensitivity of N-mixture models to violations of model assumptions might arise from their dependence on higher order moments for identifiability. A binomial random variable $X \sim B(N, p)$ has mean $\mu = Np$ and variance $\sigma^2 = Np(1-p)$. It follows that

$$N = \frac{\mu^2}{\mu - \sigma^2}. \quad (1)$$

Taken alone, the mean value of X is inadequate for estimation of N. Eq. 1 shows that it is the relation between mean and variance that makes N an identifiable parameter. Given replicate values, we may substitute sample mean and variance in the right-hand side of Eq. 1 to obtain a moment estimator $\hat{N}$. The estimator is asymptotically normally distributed and consistent. However, unmodeled variation that increases the variance relative to the mean decreases the denominator on the right-hand side of Eq. 1, inflating the estimator.

It was inspection of Eq. 1 that led us to undertake the first of our three sets of simulations. Let Y denote the number of animals recorded, including double counts. With detection probability $p = 1 - \alpha$ and accidental double count parameter γ, it is easily shown that the mean and variance of Y are $N(p + \gamma)$ and $N(p(1-p) + \gamma(3 - \gamma - 2p))$, rather than Np and $Np(1-p)$. Substituting the mean and variance of Y in the right-hand side of Eq. 1 with $p = 0.42$ and $\gamma = 0.02$, one obtains $1.26N$ rather than N. This suggests a 26% bias, a value which was nearly matched in our simulations.

Goodness-of-fit testing is always good practice, but seems especially necessary in context of the N-mixture models. Of course, the failure to find evidence of inadequate fit doesn't confirm the adequacy of a model. What is worrisome is the case where a small, undetectable violation of model assumptions can have large consequences for inference; this appears to be possible with the N-mixture models.

In discussing N-mixture models, Kéry and Royle (2015; Chapter 6) suggest the use of simulation studies as a part of study design, and as means to evaluate potential bias resulting from model misspecification. To their advice, we would add that investigators conduct simulation studies like those done in this paper to see whether small violations of assumptions can be detected. Our results strongly suggest the possibility that minor violations of assumptions - a 2% rate of accidental double counts, or an unmodeled standard deviation of 2% in detection probabilities - can have profound consequences for inference, but be virtually undetectable from the data.

Acknowledgments

We thank Jim Nichols, Evan Cooch, Len Thomas, two anonymous reviewers, and Subject Matter Editor Brett McClintock for their helpful comments. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Literature Cited

Barker, R. J., M. R. Schofield, W. A. Link, and J. R. Sauer. 2017. On the reliability of N-mixture models for count data. Biometrics 74:369–377.

Dail, D., and L. Madsen. 2011. Models for estimating abundance from repeated counts of an open metapopulation. Biometrics 67:577–587.

Duarte, A., M. J. Adams, and J. T. Peterson. 2018. Fitting N-mixture models to count data with unmodeled heterogeneity: bias, diagnostics, and alternative approaches. Ecological Modelling 374:51–59.

Fiske, I., and R. Chandler. 2011. unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. Journal of Statistical Software 43:1–23.

Hjort, N. L., F. A. Dahl, and G. H. Steinbakk. 2006. Post-processing posterior predictive p values. Journal of the American Statistical Association 101:1157–1174.

Hostetler, J. A., and R. B. Chandler. 2015. Improved state-space models for inference about spatial and temporal variation in abundance from count data. Ecology 96:1713–1723.

Kéry, M., and J. A. Royle. 2015. Applied hierarchical modeling in ecology: analysis of distribution, abundance, and species richness in R and BUGS. Volume 1: prelude and static models. Academic Press, Cambridge, Massachusetts, USA.

Nott, D. J., C. C. Drovandi, K. Mengersen, and M. Evans. 2018. Approximation of Bayesian predictive p-values with regression ABC. Bayesian Analysis 13:59–83.

Riddle, J. D., J. H. Pollock, and T. R. Simons. 2010. An unreconciled double-observer method for estimating detection probability and abundance. Auk 127:841–849.

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

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

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.2362/suppinfo