

A Robustness Study of N-Mixture Models in Estimating Animal Population Sizes

Michael Hemmer

Faculty Sponsor: Sherwin Toribio, Department of Mathematics

INTRODUCTION

The size of an animal population, also known as “abundance”, is a crucial parameter that carries significance within the field of statistics, as well as many other academic disciplines, like biology and zoology. This figure is also relevant to people outside academia, from government officials considering lists of endangered species to potential contributors to special interest groups, such as the World Wildlife Fund. It is apparent that there is a diverse body that is impacted by the estimate of abundance.

In the past, the most popular method that has been used to estimate abundance is the capture-recapture method. The essence of this procedure is to capture two separate samples of animals, marking and releasing individuals in the initial sample and observing the proportion of marked animals in the second, independently obtained sample. While this process is rather straightforward and has been widely accepted by researchers, it can be fairly demanding. It requires a high level of physical effort on the part of the capturers and it also may not be practical for certain animal species that are difficult to capture.

An alternative to the capture-recapture method is the simple count method. The simple count method estimates abundance from a collection of simple counts of observations of the animal of interest. While this process is much more convenient in the sense that a researcher can observe animals rather than physically capture them, especially for extensive examinations, it is susceptible to errors when estimating parameters.

A solution to the shortcomings of these methods is to use the N-Mixture model, as proposed by J. Andrew Royle. This model utilizes the observation counts obtained from the simple count method, regarding those counts as ones derived from a binomial distribution, using a parameter for abundance at site i , N_i , and a parameter for the probability of individual detection, p . If an experiment consists of R sites and T repeated observation periods, then all of the observations will constitute an $R \times T$ matrix of counts. Royle proposed that, if it is safe to assume that site-specific population sizes are not of direct interest, those site abundance levels could be integrated out of the likelihood function under a prior distribution, $f(\lambda)$. If we could estimate values for p and λ , which are the detection probability and mean number of animals per site, respectively, then it is plausible to estimate total abundance across all of the sites in our experiment. From this secondary likelihood function, it is possible to obtain estimates for p and λ using maximum likelihood estimation (MLE).

While Royle’s method does provide a fairly convenient and efficient approach to estimating abundance, there is at least one assumption that is worth investigating. As stated before, Royle treated each of the observation counts as a variable taken from a binomial distribution, which means that the total size of the population at that site remains constant throughout the T occasions of observation. It would be reasonable to believe that these abundance levels could be a fluctuating parameter, one that does not remain stagnant during changes in mortality rate, birth rate, abundance levels of predators, food scarcity and other factors. It is important to test Royle’s process if this closure assumption is not upheld and therefore make an evaluation on the robustness of this method as a whole.

For the remainder of this paper, we will study the robustness of estimating abundance using N-Mixture models as proposed by Royle. Section 2 describes the details of the N-Mixture model and the process of estimating the model’s parameters. Section 3 explains a simulation study that was used to evaluate the N-Mixture model’s robustness. The fourth and final section provides the results and conclusions of the study and several final remarks.

N-MIXTURE MODEL

In the N-Mixture model, as it was previously stated above, the observation counts will be considered to have been derived from a binomial distribution with a site-abundance parameter N_i and detection probability p . The site abundance levels, or N_i ’s, can be integrated out of the likelihood function via a prior distribution, provided the population size at an individual site is not necessary for the experiment. If we can assume that the animals are physically distributed evenly across the region in our study, then the Poisson distribution, with mean λ , could serve

as a prior distribution for N_i . Following these assumptions and the procedure of integrating N_i out of the likelihood function, the revised likelihood function is:

$$\left\{ \sum_{N_i = \max \{n_{it}\}}^{\infty} \left(\prod_{t=1}^T \text{Bin}(n_{it}; N_i, p) \right) \text{Pois}(N_i; \lambda) \right\}$$

In order to estimate p and λ , and in turn estimate total abundance for our study area, we will use the values that maximize this likelihood function.

To test the robustness of this model, it would be insightful to compare the estimates for p and λ obtained from the above likelihood function with those obtained from a likelihood function with increased variability. This additional variation is used to simulate changes to the animal abundance level over the course of the T sampling occasions.

ROBUSTNESS TEST USING SIMULATED DATA

Royle's method of N-Mixture models was tested by utilizing data simulated by the computer statistical software R. There were $R=(20,50)$ sampling sites that were used, and each site had an abundance level that was derived from a Poisson distribution with means of $\lambda=(2,5)$. There were $T=5$ sampling occasions used at each site and the detection probabilities of $p=(0.25,0.50)$ were considered. This provided a total of 8 scenarios with unique parameter-value combinations. 4 of these sets of circumstances were selected to further examine for the continuation of the study and they were 1) $R=20, \lambda=2, p=0.25$; 2) $R=50, \lambda=2, p=0.25$; 3) $R=20, \lambda=5, p=0.50$; 4) $R=50, \lambda=5, p=0.50$. A total of 500 iterations were executed and values of p and λ were estimated by using the MLE process for each of the 4 above scenarios (**Table 1**). The median of the 500 values for λ were recorded and used to estimate total abundance. These results were compared with those of Royle's in his research article.

Table 1. Estimates of site abundance (λ) and total abundance (N) using MLE in a simulation and Royle's published results. $T = 5$ sampling occasions for every scenario.

p	λ	R	N	Source	λ -Median	\hat{N}
0.25	2	20	40	Simulation	2.1059	42.118
				Royle	2.0385	40.77
0.25	2	50	100	Simulation	1.99	99.5
				Royle	2.0064	100.32
0.5	5	20	100	Simulation	5.001	100.02
				Royle	4.9875	99.75
0.5	5	50	250	Simulation	5.042	252.1
				Royle	4.9922	249.61

Since our simulations were essentially the same as Royle performed himself, it is no surprise that our results are very similar. Overall, the estimates for total abundance were fairly accurate, even to within one organism when surveying a region holding 100 individuals. The median estimate for λ was more accurate when $\lambda=5$. There were times Royle's estimates were more accurate, which is to be expected since Royle performed 10,000 iterations for each scenario, which implies that his results would be more accurate to the true values of λ and N.

To test the robustness of the N-Mixture model in estimating animal population size, additional variability (henceforth known as δ) was added within the data, specifically within the counts observed at an individual site for each sampling period. Specifically, a measured amount of variability was incorporated in the simulation process by adding and subtracting a randomly generated value from the binomial distribution, with parameters fixed according to the desired amount of variation.

It is worth noting here that with the addition and subtraction of randomly selected binomial values to a similar value generated from the same distribution, it is conceivable in the case when the true $\lambda = 2$ that the final result may be less than zero. While this is a possibility, it happens naturally on a very rare basis. However, considerations were made in the simulation process so that if the count were to be a negative number, a value of zero would be used for the observation count.

Simulations were performed for the same 4 scenarios as in Table 1. There were two levels of increased variability, $\delta = 0.10$ & 0.20 , that were considered for each scenario. As before, simulations of 500 iterations were calculated for both levels of δ under each scenario and the median MLE estimate(s) for λ (and p) was recorded and total abundance was estimated (Table 2).

Table 2. Simulated estimates for mean site abundance (λ) and total abundance (N) with increased amounts of variation (δ). $T = 5$ sampling occasions for all of the above scenarios.

p	λ	R	N	δ	λ -Median	N hat
0.25	2	20	40	0	2.1059	42.118
				0.10	12.479	249.58
				0.20	79.856	1597.12
0.25	2	50	100	0	1.99	99.5
				0.10	8.931	446.55
				0.20	12.304	615.2
0.5	5	20	100	0	5.001	100.02
				0.10	9.323	186.46
				0.20	19.482	389.64
0.5	5	50	250	0	5.042	252.1
				0.10	5.902	295.1
				0.20	13.329	666.45

The results clearly indicate that as additional variability is introduced into the data generation process, the estimates for mean site abundance and total abundance become positively biased, meaning the estimated values are higher than their true value. As the level of variation (δ) increases, the estimates for both λ and N deviate further and further from the values we expect. There seems to be one case, when $\delta = 0.10$ for the 4th scenario, where the estimates of abundance do not deviate as extremely as the rest of the simulated approximations. The amount of deviation in the abundance estimate seems to decrease as either the number of sites (R) increases.

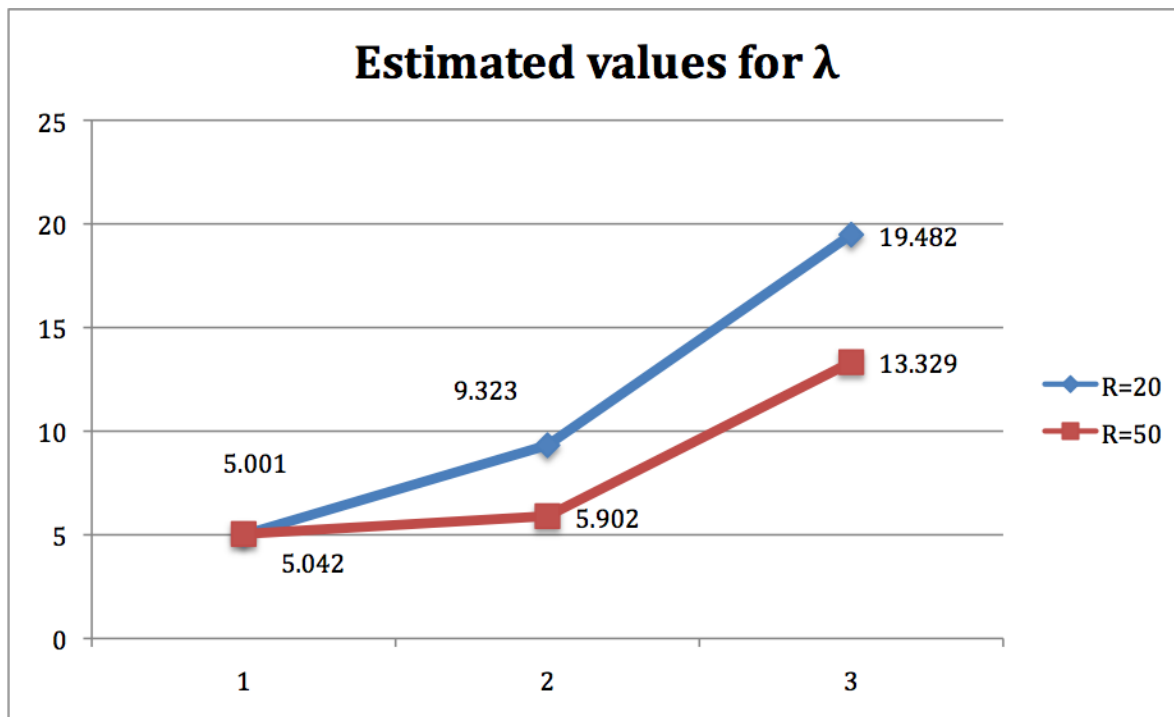


Figure 1. Estimated values for λ from simulated data, as variability is incrementally increased

CONCLUSION AND FINAL REMARKS

This paper tested the robustness of the N-Mixture model method, as proposed by Royle in his 2004 article, when the assumption of an animal population that was closed to external forces was not upheld. This intentional violation introduced a certain level of additional variability in the observation counts at an individual site at a single sampling occasion. After ensuring our estimation process yielded results comparable to that of Royle when no inconsistency was added, we were prepared to test the affects of increased variability on the estimates of abundance.

After conducting a simulation study, it was possible to see two trends. The first is that Royle's N-Mixture model method in estimating abundance is unreliable when the closure assumption is not true. In other words, if the population in question experiences any changes throughout the duration of the experiment, the abundance estimates resulting from the N-Mixture model method may not be accurate. The second pattern that can be observed is that the estimates for abundance improve as the number of sites used in the experiment increases.

We conclude that any future research using the N-Mixture model to estimate abundance must exercise caution if the size of the animal population at each site may change during the experiment. We would also encourage any future research that estimates or utilizes the parameter of abundance to incorporate more observation sites in their experiment as compared to less.

There are several final remarks to make on this research. Due to time constraints, only 500 iterations could be performed for each scenario, and not all of the scenarios were even considered during this experiment. The process that was used to maximize the modified likelihood function may also be erroneous at times. It is possible that the value that was returned as that which maximizes the likelihood function may only be considered a local maximum and not one that is the global maximum, the point that returns the greatest value of the function for any input. A goal of future research on this topic would be to include more iterations in the simulations, include all 8 possible scenarios of R , λ and p , and investigate the maximization process used in this research.