N-Mixture Models for Ecological Data

Fiona Wilson

1 Introduction

A common aim in ecological studies is to estimate the population of a species from survey data. Often the data from surveys is count data because it is cheaper and easier to collect than capture-recapture data. Count data can be collected by scientists and also from citizen science projects which means there can be a lot of data collected for minimal cost. The data is a count of how many individuals were seen at a particular site during the survey and this survey is often repeated at different sites and times. There is no way to know if the same individual is seen on multiple surveys or at multiple sites since these are unmarked individuals. We also don't know how many undetected animals were at the site at the time, so for each survey there is an associated detection probability. A further difficulty with this type of data is that the counts are often quite sparse with a lot of zero values.

To estimate the true population size from the count data we can use N-mixture models (Royle, 2004). These models estimate the population at different survey sites by treating the true population numbers as independent random variables and then a mixture distribution models these random variables while taking into account the probability of detecting an individual. There are different variations of this model and this report will discuss the standard model, a model with open populations (Dail and Madsen, 2011), a model with multiple states (Zipkin, Thorson, et al., 2014) and a spatial dependence model (Zhao et al., 2017).

This report focuses on modelling species abundance (the number of individuals at a site), however there are similar models available which measure species occurrence. Occurrence just informs you of the presence or absence of a species at a site as explained by Royle and Kéry, 2015. Estimating abundance requires more detailed data, but the results can be more informative to understanding population dynamics.

2 Standard N-Mixture Model for Sparse Data

2.1 Model Formulation

The standard N-mixture model proposed by Royle, 2004, is used when we have count data for one species of animal at R different sites and each site is sampled T many times. Let $n_{i,t}$ be the number of individuals counted at site i (for $i = 1, ..., R$) at time t (for $t = 1, ..., T$). In this standard model we assume that the population is closed so there are no changes in the population size (i.e. no births, deaths or migration). We also assume that the number of individuals at each site and time are independent binomial random variables

$$n_{i,t} \sim \text{Bin}(N_i, p)$$

where N_i is the unknown total number of individuals at site i and p is the probability of detection. For now we assume that p is constant across these sites and across individuals. Later models will remove the assumptions of a closed population and constant probability of detection.

The joint likelihood is given by

$$L(\{N_i\}, p|\{n_{i,t}\}) = \prod_{i=1}^{R} \left\{ \prod_{t=1}^{T} \text{Bin}(n_{i,t}; N_i, p) \right\}$$

where $\text{Bin}(n_{i,t}; N_i, p)$ is the binomial likelihood and this can be maximised to find estimates for the abundance parameters N_i and detection probability p. However, this can be quite unstable because there are often a lot of zero counts and there are also a lot of parameters to estimate, some of which may even be zero. This instability is addressed in more detail by Olkin et al., 1981. To improve our estimates, we instead assume that N_i are independent random variables and treat them as a nuisance parameter with prior distribution $f(N_i; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ can be a vector.

2.2 Prior Distributions

There are several commonly used distributions for N_i. The Poisson distribution is often used in ecological settings because it assumes that animals are randomly distributed across the site. However, if animals prefer to live in groups or if the habitat is variable across the site, then the animals are more likely to be clustered instead of randomly dispersed. This causes overdispersion in the data since the variance can be larger than the mean.

In this situation, the negative binomial distribution may be more appropriate. Other possible distributions include the zero inflated Poisson and zero inflated negative binomial distributions. These can be appropriate if there are a lot of zero values in the data, more than the expected number if modelled using a Poisson or negative binomial distribution. For example, if one survey site was mostly inhabitable, then this would generate structural zeroes instead of the expected sampling zeroes.

It is important to note that in general when selecting the appropriate distribution, it is common to use the Akaike Information Criterion (AIC) or goodness of fit test to compare the final models. However, there have been repeated studies with ecological data such as those by Joseph et al., 2009, and Kéry et al., 2005, which show that these statistical tests may select the negative binomial model, but the estimates then produced are not realistic when you consider the ecological setting. This is still an issue today which does not have a clear solution or explanation, and so most papers indicate that alongside these statistical tests, appropriate knowledge of the ecological processes should be used to assess whether estimates and predictions are realistic (Royle and Kéry, 2015).

Using the chosen prior distribution for N_i, the likelihood is now

$$L(p,\theta|\{n_{i,t}\}) = \prod_{i=1}^{R}\left\{\sum_{N_i=\max_t n_{i,t}}^{\infty}\left(\prod_{t=1}^{T}\text{Bin}(n_{i,t};N_i,p)\right)f(N_i;\theta)\right\}. \tag{1}$$

To maximise this likelihood, usually we take the summation over a large K (instead of an infinite summation), however in practice K should be chosen carefully to ensure the resulting estimates have stabilised. This parameter sensitivity is highlighted in the data analysis in Section 4. There are alternative forms of the likelihood in Section 2.4 which avoid the challenge of choosing an appropriate K. Given the maximum likelihood estimates, the expected abundance can be estimated in several ways. If you just want to estimate the total abundance across the whole location (not site-specific abundance), this can be estimated simply using the area of the sites and the mean abundance per sample unit estimated from the prior distribution. For example, if N_i has a Poisson distribution, then $\hat{\lambda}$ is the mean abundance per site and if there are R sites, then the estimated total abundance of the sampled sites is just $N = \hat{\lambda}R$. If instead we want to estimate the site-specific abundance, we can use Bayes Theorem to estimate the posterior distribution of N_i conditional on $\hat{\theta}$,

$$\mathbb{P}(N_i = k|n_{i,1},n_{i,2},...,n_{i,T},\hat{\theta},\hat{p}) = \frac{\mathbb{P}(n_{i,1},.n_{i,2},...,n_{i,T}|N_i=k,\hat{p})\mathbb{P}(N_i=k;\hat{\theta})}{\sum_{k=0}^{\infty}\mathbb{P}(n_{i,1},n_{i,2},...,n_{i,T}|N_i=k,\hat{p})\mathbb{P}(N_i=k;\hat{\theta})}.$$

This approach is known as the 'plug-in' empirical Bayes procedure. With this posterior distribution, we can calculate the expected value for N_i at each site, as well as the corresponding variance and confidence intervals.

2.3 Bayesian Approach

An alternative method is to use a fully Bayesian approach where you would also provide priors for the model parameters p and θ as well as for N_i, though since it is likely that little is known about these, vague priors can be used (Madsen and Royle, 2023). This approach usually uses Markov Chain Monte Carlo (MCMC) simulation to generate a large sample from the joint posterior distribution from which we can calculate point estimates.

One advantage of using this Bayesian approach is that it is easier to add random effects into the covariate models whereas for the maximum likelihood approach we have to integrate over the random effects when computing the likelihood. Common R software such as `unmarked` (Kellner et al., 2023; Fiske and Chandler, 2011) which is used later in the data analysis can incorporate limited random effects, but it is not complete. However, in general using maximum likelihood estimation is quicker and computationally more efficient, especially with larger datasets, because we don't have to iteratively sample from the data. Furthermore, in practice the maximum likelihood based software currently available to ecologists is easier to understand and input the data, whereas often the Bayesian approach requires more understanding of priors and posteriors to accurately input the data and interpret the results.

2.4 Alternative formulations without an infinite sum

As mentioned previously, the exact likelihood given in equation 1 includes an infinite sum and it can be difficult to truncate this sum as the population estimates can be sensitive to the choice of K. One alternative formulation used by Dennis et al., 2015, rewrites a Poisson N-mixture model as a multivariate Poisson model (or equivalently replaces a negative binomial N-mixture model with a multivariate negative binomial model). The explicit form for a likelihood provided in their paper does not require K as there is no infinite sum. However, this approach can be more computationally complex and so if we have enough sampling times (from simulations done in the paper for the Poisson distribution, enough times was generally $T > 3$), then the parameter estimates from the maximum likelihood stabilise anyway.

Another alternative is suggested by Haines, 2016, where she writes the likelihood in a closed form using the generalized hypergeometric function. However, again this is computationally complex to evaluate as not all programming languages have functions to calculate it and instead it has to evaluated from the closed form.

2.5 Adding covariates

When analysing population numbers across a large area, it is unlikely that each survey site was exactly the same or that each survey was performed at the same time and in the same conditions. There may be some underlying environmental factors which affect the number of individuals living at a certain site or which affect the probability of detecting them during a survey. To calculate more accurate abundance estimates, we want to include these influences as covariates in our N-mixture model using link functions. If we are using a Poisson N-mixture model we would add covariates to the estimation of λ_i if the sites vary a lot and we believe this may impact the abundance. For example, changes in elevation, forest cover and a bird's route length were used in a study on willow tits (*Parus montanus*) with data collected by the Swiss Breeding Bird Survey (Royle et al., 2005). It is common to use a log link function of the form

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i$$

where x_i is the value of a covariate at site i as this ensures that λ_i are positive.

Similarly, covariates can be added to $p_{i,t}$ if we believe the detection probability varies due to factors such as the duration of the survey, the time of day or the weather. To ensure that $p_{i,t}$ is in the range $[0, 1]$ we use a logit function of the form

$$\text{logit}(p_{i,t}) = \alpha_0 + \alpha_1 x_{i,t}$$

where $x_{i,t}$ is the value of the covariate at site i at time t. By modelling these as fixed effects, we can still apply the maximum likelihood method as mentioned previously.

A different way to model the spatial variation in expected abundance in a Poisson N-mixture model is to treat λ_i as a random variable. It is reasonable to assume it has a Gamma distribution because this ensures λ_i is positive and it is a flexible distribution. Then N_i has a negative binomial distribution (or Gamma-Poisson distribution) and so it can be modelled as before (White and Bennetts, 1996).

2.6 Non-identifiability

If there is only one sampling occasion, then we can have the problem of model non-identifiability (Madsen and Royle, 2023). Assuming that $N_i \sim \text{Poisson}(\lambda)$ are independent and identically distributed, then the marginal distribution of n_i is now $\text{Poisson}(\lambda p)$ where p is still the detection probability. Obviously if we had perfection detection (i.e. $p = 1$) then we have no problem, but in practice when $p < 1$, the model is non-identifiable because from our observations we only know λp instead of λ and p. However, Sólymos et al., 2012, suggest a method to still use this data to produce abundance estimates in certain circumstances. Though it hasn't yet been proved, simulations suggest that if there are continuous covariates, at least one of which is unique to either the abundance rate or detection probability, then the model would be identifiable. This method calculates the conditional likelihood function for the coefficients of the covariates to reduce the confounding and then produce abundance estimates. However, if there are no covariates or only discrete covariates, then the model is still non-identifiable.

3 Model Extensions

3.1 Open Population

The models previously discussed assume that we have a closed population, however this is often only a valid assumption if surveys are done on an isolated population over a short period of time. If these models were used to estimate the abundance of open populations, then the estimates are likely to be very biased because if the point count varies a lot, the model will assume this is due to high non-detections or detections instead of a varying population. This could indicate a lower probability of detection and so the estimates of the total population would be too high. Instead we should use a dynamic model which allows for a different population size at each time point. We discuss the model developed by Dail and Madsen, 2011.

Again we assume there are R survey sites and T times of surveys. Let $n_{i,t}$ be the number of individuals counted and let $N_{i,t}$ be the true number of individuals at site i at time t. We assume the point counts have the distribution

$$n_{i,t} \sim \text{Bin}(N_{i,t}, p)$$

and we need a joint prior for $N_{i,1}, ..., N_{i,T}$. We still assume that the population at the R

sites are independent and that each animal detection is independent. Therefore we have the joint likelihood function,

$$L(p,\lambda,\gamma,\omega|\{n_{i,t}\}) = \prod_{i=1}^{R}\left[\sum_{N_{i,1}=n_{i,1}}^{\infty}\ ...\ \sum_{N_{i,T}=n_{i,T}}^{\infty}\left(\prod_{t=1}^{T}\text{Bin}(n_{i,t};N_{i,t},p)\right)\times f(\{N_{i,1},...N_{i,T}\};\theta)\right].$$

We simplify this by assuming that at a given site, the Markov property applies. That is, for any $t \geq 2$, we assume $N_{i,t}$ only depends on $N_{i,t-1}$. We now have the joint prior,

$$f(\{N_{i,1},...,N_{i,T}\};\theta) = f(N_{i,1};\theta)\prod_{t=2}^{T} f(N_{i,t};N_{i,t-1},\theta)$$

where each $f(N_{i,t};\theta)$ is the distribution of animals at site i and so can be one of the distributions discussed in Section 2.2.

To model the dynamic population we introduce the random variables $S_{i,t}$ and $G_{i,t}$. Let $S_{i,t}$ be the number of survivors, that is the number of animals that were at site i at time $t-1$ and were then still at the same site at time t. And let $G_{i,t}$ be the number of gains, that is the number of new animals who weren't at site i at time $t-1$, but who were there at time t. We note that since we do not mark and identify individuals, we cannot distinguish between different reasons for decreases in population (e.g. death or emigration), nor between different reasons for increases (e.g. births or immigration). Now let ω be the survival probability (this is usually the apparent survival probability since individuals may enter and leave sites, but we are not tracking individuals, just the change in the point count) and let γ be the arrival rate, then we have the conditional distributions

$$\begin{aligned} S_{i,t}|N_{i,t-1} &\sim \text{Bin}(N_{i,t-1},\omega) \\ G_{i,t}|N_{i,t-1} &\sim \text{Poisson}(\gamma(N_{i,t-1})) \end{aligned}$$

for $t = 2,...T$ and where $\gamma(N_{i,t-1})$ means that γ depends on $N_{i,t-1}$. Therefore given the abundance at $t-1$, the abundance at t is just the sum of these random variables,

$$N_{i,t}|N_{i,t-1} = (S_{i,t} + G_{i,t})|N_{i,t-1}.$$

We assume that $S_{i,t}$ and $G_{i,t}$ are independent and so we calculate the conditional proba-

bility for $N_{i,t}|N_{i,t-1}$ as

$$\mathbb{P}(N_{i,t} = k|N_{i,t-1} = j) = P_{j,k} = \sum_{c=0}^{\min(j,k)} \mathbb{P}(S_{i,t} = c|N_{i,t-1} = j)\mathbb{P}(G_{i,t} = k - c|N_{i,t-1} = j)$$

for $j, k = 1, 2, \cdots$. In this summation we are multiplying the probability of c survivors by the probability of $k - c$ gains given that the abundance at the previous time was j. We sum over all possible c values to calculate the total conditional probability.

For simplicity, we now assume a Poisson prior on $f(N_{i,1}; \theta)$ and so we write the joint likelihood as

$$L(p, \lambda, \gamma, \omega|\{n_{i,t}\}) = \prod_{i=1}^{R} \left[\sum_{N_{i,1}=n_{i,1}}^{\infty} \cdots \sum_{N_{i,T}=n_{i,T}}^{\infty} \left\{ \left(\prod_{t=1}^{T} \text{Bin}(n_{i,t}; N_{i,t}, p) \right) \times \frac{e^{-\lambda}\lambda^{N_{i,1}}}{N_{i,1}!} \cdot \prod_{t=2}^{T} P_{N_{i,t-1},N_{i,t}} \right\} \right].$$

As with the standard N-mixture model, when maximising this likelihood we typically choose a large finite K for the summation instead of using the infinite sumations. Again, the choice of K can impact the estimates, so it needs to be chosen with care.

Similarly to the standard model, we will maximise the likelihood to find parameter estimates and then use these to estimate the abundance. Analogously, we can either estimate the total abundance across the sites or the site-specific abundance at each time point. If we want the total abundance, we can recursively calculate it using

$$\begin{aligned} \widehat{N}_{.,1} &= R\hat{\lambda}, \\ \widehat{N}_{.,t} &= \hat{\omega}\widehat{N}_{.,t-1} + R\hat{\gamma}. \end{aligned}$$

On the other hand if we want to estimate the site-specific abundance at each time t we can use the 'plug-in' empirical Bayes procedure again. We can use an improper prior $f(N_{i,t}) \propto 1$ (other more informative priors could work too, but they may not give a closed form expression) and then we have

$$\mathbb{P}(N_{it} = k|n_{i,t}, \hat{p}) = \binom{k}{n_{i,t}} \hat{p}^{n_{i,t}+1}(1 - \hat{p})^{k-n_{i,t}}.$$

We can estimate $N_{i,t}$ by the mean of this distribution, so

$$\mathbb{E}(N_{i,t}|n_{i,t},\hat{p}) = \frac{n_{i,t}}{\hat{p}} + \frac{1-\hat{p}}{\hat{p}}.$$

This open population model is known as the generalised N-mixture model because it can be reduced to the standard closed population model by setting the parameters $\omega = 1$ and $\gamma = 0$. This is an advantage because it means the assumption of a closed (or open) population can be tested using these nested models if there is a large number of sites, R, as explained by Dail and Madsen, 2011. This is particularly useful because often the closure assumption is made for point counts collected during one season due to biological reasons, however this model provides a method to statistically test this assumption.

Despite the benefits of this model, there are still certain assumptions being made which could be problematic. Firstly, we've assumed that the sites are independent. This is a common assumption in our models, but may not be accurate if animals are forced to move between sites. There are other models which model the spatial dependence of population abundance such as that developed by Zhao et al., 2017.

Another assumption made is that $S_{i,t}$ and $G_{i,t}$ are independent, however this assumption is not necessary and was just a simplification. One model for dependence suggested by Dail and Madsen, 2011, is to let $G_{i,t}|(S_{i,t} = s)$ be a Poisson random variable with mean $\gamma = se^{\varphi}$. This implies that if there are more survivors, there are also likely to be more gains which could be reasonable if a species values safety in numbers or is in its breeding season.

3.2 Multi-State Model

Another extension of the model is to include multiple states for a species, for example stratifying the data into adults and juveniles, or males and females. This allows ecologists to better understand the population structure and there may be covariates which are state dependent, for example during breeding season males may be more visible while females are brooding. One suggested approach to model these states was developed by Zipkin, Thorson, et al., 2014, and is an extension of the open population model in Section 3.1.

Let $n_{i,j,t}$ be the number of individuals counted at site i in state j at time t. The states could be quite specific such as male juveniles ($j = 1$), female juveniles ($j = 2$), male adults ($j = 3$) and female adults ($j = 4$). Here we just consider two states juveniles ($j = 1$) and adults ($j = 2$). We still assume that $n_{i,j,t} \sim \text{Bin}(N_{i,j,t}, p_j)$ where $N_{i,j,t}$ is the true

population at site i of state j at time t and p_j is the probability of detecting an individual who is of state j. We assume that $N_{i,j,1}$ has a standard distribution as mentioned in Section 2.2 such as $N_{i,j,1} \sim \text{Poisson}(\lambda_j)$. We need to allow transitions between states because if surveys are done over a long time period, individuals may develop from juveniles to adults. Therefore the distribution of $N_{i,j,t}$ for $t \geq 2$ must also depend on these transitions.

As for the open population model, we consider random variables for the survivors ($S_{i,1,t}$, $S_{i,2,t}$) and the gains ($G_{i,1,t}$), as well as one for the state transitions ($T_{i,j,t}$). For now we assume that gains are due to births of juveniles instead of immigration of adults. So we have,

$$\begin{aligned} S_{i,1,t} &\sim \text{Bin}(N_{i,1,t-1}, \omega_1) \\ S_{i,2,t} &\sim \text{Bin}(N_{i,2,t-1}, \omega_2) \\ T_{i,1,t} &\sim \text{Bin}(S_{i,1,t}, \varphi) \\ G_{i,1,t} &\sim \text{Poisson}(\gamma(N_{i,2,t-1})) \end{aligned}$$

where ω_1 and ω_2 are the state-specific survival probabilities, φ is the probability that a surviving juvenile transitions to an adult, and γ is the arrival rate. Therefore the state-specific abundances at a site i are

$$\begin{aligned} N_{i,1,t} &= G_{i,1,t} + S_{i,1,t} - T_{i,1,t} \\ N_{i,2,t} &= S_{i,2,t} + T_{i,1,t}. \end{aligned}$$

The paper then used the Bayesian approach with MCMC to estimate the parameters.

We have considered just two states where individuals can only move from juveniles to adults (they can't move from adults to juveniles). More generally a transition matrix T could be used to allow transitions between more states (such as an individual moving back and forth between breeding and non-breeding states) and the model could also incorporate states which can't be transitioned between (such as male and female). Further, it can be extended to include immigration, whereas currently our model just includes gains to juvenile counts from births. This could be included by adding another gain random variable $G_{i,2,t} \sim \text{Poisson}(\gamma_2)$ for example as done by Sirén et al., 2024.

Another important addition to the model discussed by Zipkin, Sillett, et al., 2014, is that sometimes an individual may be observed, but the surveyor can't determine which state the individual is in. However, this imperfect observation is still useful information to include in a model as they contribute to the overall population. The proposed method

is to model these uncertain counts separately and use both probabilities of detection as well as probabilities of correctly classifying an individual when defining the point count distributions. For example, we would also include $n_{i,3,t}$ which is the number of individuals of unknown age and add a probability c for being able to correctly classify an individual. However, by increasing the number of parameters to determine, this can increase the amount of data required to provide identifiable estimates.

3.3 Spatial Model

In the original open population model, we assumed that we could not distinguish between different increases (births and immigration) or between different decreases (deaths and emigration) in counts. However, the model proposed by Zhao et al., 2017, separates these causes and uses the spacing of survey sites to model the migration of animals between them. We partition the local area such that each survey site is in just one partition. We say that two sites are adjacent if their partitions share a side and we believe that animals are more likely to move between adjacent sites. We still assume that $n_{i,t} \sim \text{Bin}(N_{i,t}, p_i)$ where $N_{i,t}$ is the true population at site i (for $i = 1, ..., R$) at time t and p_i is the probability of detecting an individual at site i. We assume that $N_{i,1}$ has a standard distribution as mentioned in Section 2.2 such as $N_{i,1} \sim \text{Poisson}(\lambda)$. We now model the true survival, the reproduction, the emigration and the immigration by the following random variables for $t \geq 2$:

$$
\begin{aligned}
S_{i,t} &\sim \text{Bin}(N_{i,t-1}, \omega) && \text{where } \omega \text{ is the survival rate} \\
R_{i,t} &\sim \text{Poisson}(\gamma(N_{i,t-1})) && \text{where } \gamma \text{ is the reproduction rate} \\
E_{i,t} &\sim \text{Bin}(S_{i,t}, \kappa) && \text{where } \kappa \text{ is the rate of emigration} \\
I_{i,t} &\sim \text{Poisson}\left(\sum_{j=1}^{R} w_{i,j} E_{j,t}\right) && \text{where } w_{i,j} \text{ are the weights of the movements}
\end{aligned}
$$

There are different options for the weights, however the proposed method is

$$
\begin{aligned}
w_{i,j} &= \frac{1}{n_j^{adj}} && \text{if sites } i \text{ and } j \text{ are adjacent} \\
w_{i,j} &= 0 && \text{otherwise}
\end{aligned}
$$

where n_j^{adj} is the number of sites adjacent to site j. With this model the abundance at each site i and time t is given by

$$N_{i,t} = S_{i,t} + R_{i,t} - E_{i,t} + I_{i,t}.$$

In the paper by Zhao et al., 2017, they again used a Bayesian approach to estimate the parameters via MCMC.

The assumption in this model that individuals only move between adjacent sites is more likely to be appropriate for species with limited mobility and if sites are sufficiently far apart. If animals frequently move large distances, this would be an inaccurate assumption. In future models, instead of just allowing movement between adjacent sites, other forms of connectivity could be used when we have knowledge of the landscape. For example, if there are well-trodden paths or migration routes between sites, then animals would be more likely to move between these sites even if the distance is greater.

3.4 Other Extensions

There are yet more extensions to the standard N-mixture model available, however they will not be discussed in detail here. For example, there are multi-species models which take into account the correlations between species (Mimnagh et al., 2022). There are also models which include temporary emigration when individuals leave and then re-enter survey sites (Chandler et al., 2011). Most of the dynamic models discussed here were available as multi-season models where k surveys were conducted for each year t (usually to ensure robust design they required $k > 1$).

4 Data Analysis

As part of a conservation project in 2022, count data was collected by the author during animal surveys in three sites in the Lokobe National Park on Nosy Be, Madagascar. The three sites were Ampasipohy which is part of a local tribes' land and was surveyed on 7 days, and Kindro and Ramy which are the different trails in the main national park and were both surveyed on 9 days, however across the full data set there are 13 survey days as different sites were surveyed on different days. The data was collected on herpetofauna and for this report we consider the count data for the brown mantella frog (*Mantella ebenaui*) and the Nosy Be plated lizard (*Zonosaurus subunicolor*).

To estimate the number of individuals at each survey site using the standard N-mixture model discussed in Section 2.1, we used the package `unmarked` in R from Kellner et al., 2023, and Fiske and Chandler, 2011. Ideally each site would have been surveyed on every survey day resulting in 13 time points for all sites, however they weren't, due to time and manpower constraints which is a common issue in ecological data. However, the R package used can manage unbalanced datasets which may have missing data for certain sites.

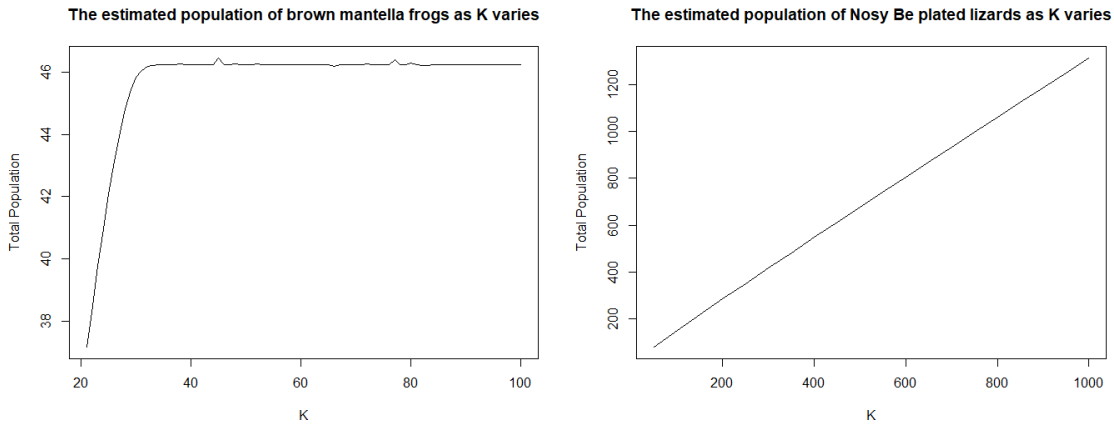

Figure 1: Plots showing the estimated total populations as K varies

First we consider the brown mantella frog dataset. For this data, the means and variances of our count data at each site were quite similar, so we used a Poisson distribution to model the total number of species at each site. We also checked whether a zero inflated Poisson was more appropriate by fitting both models, however using the AIC to compare model fits we determined that the Poisson model was most appropriate. We selected the parameter $K = 50$ since for choices of K above 40 the total abundance stabilises as shown in the first plot of Figure 1. We used the `pcount` function to model our data. Figure 2 shows the posterior distributions of the species abundance at each site. These distributions show the uncertainty in our abundance estimates and how this varies by site. The estimates for site 2 are the most certain, but it still has a low probability of being the true abundance. This is a common challenge with count data, particularly for small datasets where there is more variation in the data.

The expected abundances and 95% confidence intervals are included in Table 1 and the overall estimated probability of detection was 0.219. From the posterior distributions and the table, clearly there are wide confidence intervals for these estimates due to the limited count data available. More surveys would be needed to reduce these confidence intervals.

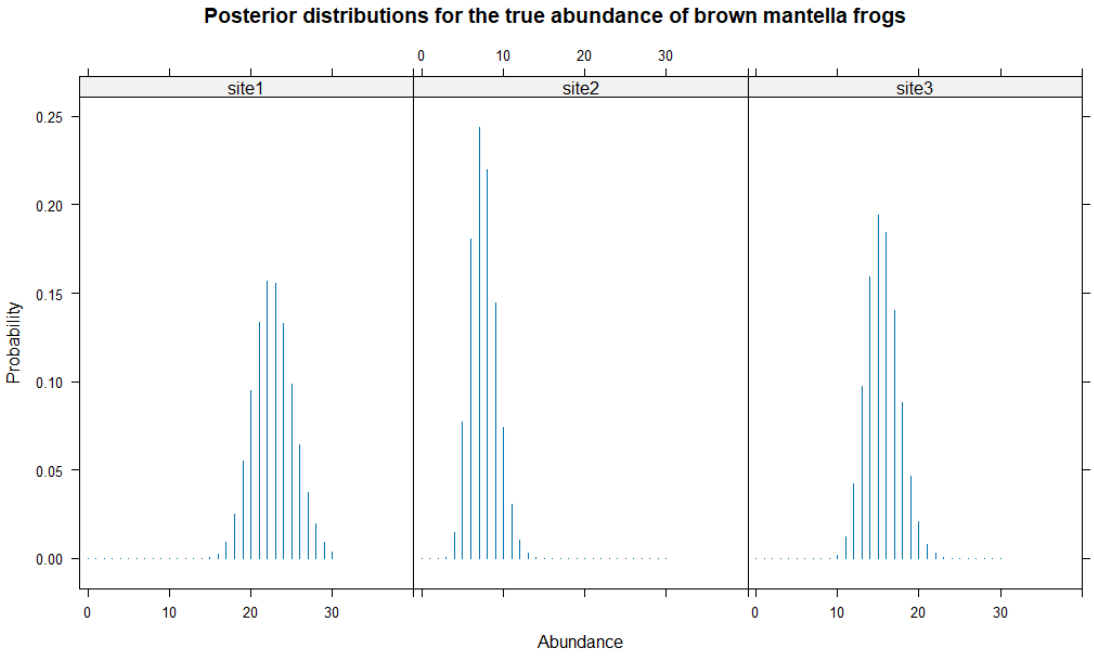

Figure 2: The posterior distributions for the true abundance of brown mantella frogs at site1 (Ampasipohy), site2 (Kindro) and site3 (Ramy)

	Mean	95% Confidence Interval
Ampasipohy	22.72	(18,28)
Kindro	7.55	(5,11)
Ramy	15.56	(12,20)

Table 1: The estimated true abundance of brown mantella frogs at each site

We also added a covariate to the probability of detection using the logit link. It was the number of minutes after sunrise when we started the survey. We thought this may have impacted the results because the temperatures can get very warm in the middle of the day which might decrease the chance of seeing these frogs as they seek shade when it is hottest. However, the estimated abundance for each site was very similar (in fact the confidence intervals were exactly the same) and when we calculated the AIC values, the model with covariates had an AIC value of 118.62 and the model without covariates had 119.69. This was not a significant difference and so the time after sunrise did not seem to have a significant impact on the probability of detection.

We then tried to repeat this analysis with the Nosy Be plated lizard data. The variances of the count data were much higher than the means, so we wanted to use a negative

binomial distribution. However, when choosing the value of K, the total population estimate didn't stabilise for $K = 1000$ and instead just increases as K increases as shown in the second plot in Figure 1. This suggests that any estimates from this model would be inaccurate as they are entirely dependent on the value of K chosen. This is likely due to the small amount of data available and thus its large variance. This is the problem when survey data is collected on a small-scale, any estimates are very sensitive to the choice of parameters. We again tried fitting it with the time after sunrise covariate to see if this improved the issue, however K still didn't stabilise. To produce accurate abundance estimates, more data is needed.

5 Conclusions

In the last 20 years there has been a lot of development of different N-mixture models to estimate species abundance from count data. It is becoming more and more important to accurately understand animal population numbers if we want to implement correct measure to protect them. There is also a lot of public support for such surveys, so involving the public to collect the count data is a valuable resource and much cheaper than running capture-recapture surveys. After the first N-mixture model was developed, there have been many proposed extensions, however in practice many ecologists still use the standard model, so work needs to be done to make other models and coding packages more accessible to them.

As mentioned throughout this report, all of these models have their limitations and points for further development. For example, the standard model is reliant on the choice of parameter K which, if not chosen carefully, will significantly change the population estimates. For some of the more complicated models, a Bayesian approach was suggested, however this again requires better understanding of priors and posteriors to make best use of the model. There could also be further work on spatial dependence models, especially if there is local environmental knowledge available.

There is also a need for new models to be developed which combine existing ideas. For example, developing a multi-season model which also includes the correlations between multiple species. Although, this report focused on models just using count data, if a model was developed to combine count data with capture-recapture data, this could further improve abundance estimates. Capture-recapture data is expensive to collect, but when it is available, it can provide more accurate population estimates, so it would be beneficial to have a model which can use both types of data to estimate species abundance.

The continued development of N-mixture models which better reflect the true biological processes will allow ecologists to make more accurate estimates from incomplete survey data. This is increasingly important as we focus on understanding and protecting biodiversity.

References

Chandler, R. B., Royle, J. A., & King, D. I. (2011). Inference about density and temporary emigration in unmarked populations. *Ecology*, *92*(7), 1429–1435.

Dail, D., & Madsen, L. (2011). Models for Estimating Abundance from Repeated Counts of an Open Metapopulation. *Biometrics*, *67*(2), 577–587.

Dennis, E. B., Morgan, B. J. T., & Ridout, M. S. (2015). Computational Aspects of N-Mixture Models. *Biometrics*, *71*(1), 237–246.

Fiske, I. J., & Chandler, R. B. (2011). unmarked: An R Package for Fitting Hierarchical Models of Wildlife Occurrence and Abundance. *Journal of Statistical Software*, *43*(10), 1–23.

Haines, L. M. (2016). Maximum Likelihood Estimation for N-Mixture Models. *Biometrics*, *72*(4), 1235–1245.

Joseph, L. N., Elkin, C., Martin, T. G., & Possingham, H. P. (2009). Modeling Abundance Using N-Mixture Models: The Importance of Considering Ecological Mechanisms. *Ecological Applications*, *19*(3), 631–642.

Kellner, K. F., Smith, A. D., Royle, J. A., Kéry, M., Belant, J. L., & Chandler, R. B. (2023). The unmarked R package: Twelve years of advances in occurrence and abundance modelling in ecology. *Methods in Ecology and Evolution*, *14*(6), 1408–1415.

Kéry, M., Royle, J. A., & Schmid, H. (2005). Modeling Avian Abundance From Replicated Counts Using Binomial Mixture Models. *Ecological Applications*, *15*(4), 1450–1461.

Madsen, L., & Royle, J. A. (2023). A review of N-mixture models. *WIREs Computational Statistics*, *15*(6), e1625.

Mimnagh, N., Parnell, A., & Prado, E. (2022). Bayesian multi-species N-mixture models for unmarked animal communities. *Environmental and Ecological Statistics*, *29*(4), 755–778.

Olkin, I., Petkau, A. J., & Zidek, J. V. (1981). A Comparison of n Estimators for the Binomial Distribution. *Journal of the American Statistical Association*, *76*(375), 637–642.

Royle, J. A. (2004). N-Mixture Models for Estimating Population Size from Spatially Replicated Counts. *Biometrics*, *60*(1), 108–115.

Royle, J. A., & Kéry, M. (2015). *Applied Hierarchical Modeling in Ecology: Analysis of Distribution, Abundance and Species Richness in R and BUGS: Volume 1: Prelude and Static Models* [Pages 3-4, 257-264]. Elsevier Science and Technology.

Royle, J. A., Nichols, J. D., Kéry, M., & Ranta, E. (2005). Modelling Occurrence and Abundance of Species When Detection is Imperfect. *Oikos*, *110*(2), 353–359.

Sirén, A. P. K., Hallworth, M. T., Kilborn, J. R., Bernier, C. A., Fortin, N. L., Geider, K. D., Patry, R. K., Cliché, R. M., Gifford, S. J., Wixsom, S., Morelli, T. L., & Wilson, T. L. (2024). Monitoring Animal Populations With Cameras Using Open, Multistate, N-Mixture Models. *Ecology and Evolution*, *14*(12), e70583.

Sólymos, P., Lele, S., & Bayne, E. (2012). Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. *Environmetrics*, *23*(2), 197–205.

White, G. C., & Bennetts, R. E. (1996). Analysis of Frequency Count Data Using the Negative Binomial Distribution. *Ecology*, *77*(8), 2549–2557.

Zhao, Q., Royle, J. A., & Boomer, G. S. (2017). Spatially explicit dynamic N-mixture models. *Population Ecology*, *59*(4), 293–300.

Zipkin, E. F., Sillett, T. S., Grant, E. H. C., Chandler, R. B., & Royle, J. A. (2014). Inferences about population dynamics from count data using multistate models: a comparison to capture–recapture approaches. *Ecology and Evolution*, *4*(4), 417–426.

Zipkin, E. F., Thorson, J. T., See, K., Lynch, H. J., Grant, E. H. C., Kanno, Y., Chandler, R. B., Letcher, B. H., & Royle, J. A. (2014). Modeling structured population dynamics using data from unmarked individuals. *Ecology*, *95*(1), 22–29.