Ch. 18

N-mixture models[1]

"There is no great genius without a mixture of madness."
-- Aristotle

Questions to ponder:

- *How do N-mixture models differ from occupancy models?*
- *How do I gather data for analysis with N-mixture models?*
- *Why are these models called "mixture" models?*
- *How important is the closure assumption for N-mixture models?*

N-mixture model basics

Royle (2004) introduced the concept of *N*-mixture models, which are used to estimate abundance from spatially and temporally replicated surveys. In fact, many small-scale studies or large-scale monitoring efforts are replicated spatially because the ecologist often has a goal to characterize spatial variation in abundance. And, these studies often have multiple visits over time. The end-result of these studies is a set of data, $y_{i,k}$, a set of observed counts from each site *i* and each visit *k*.

The essence of the study design is shown in Figure 18.1 for three sites that are sampled over 4 time periods. In Chapter 15, we saw a similar design for occupancy methods, but the response values for the data were either 1's or 0's (species observed or not observed). Here, the response data is the number (typically counts) of individuals seen at a point.

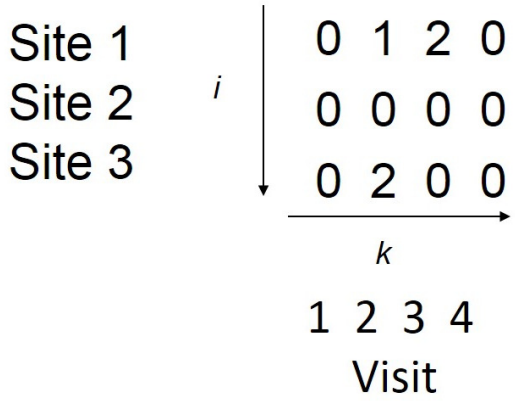

Figure 18.1: *Example design for a study using N-mixture methods with 3 sample sites (I) and 4 visits (k) to each site over time.*

As with previous survey methods, *N*-mixture models are designed to account for incomplete detectability. And, we can see evidence of detectability problems in our data. For Site 1, we can see that the minimum number of animals is 2. We did not see any individuals during visit 1 and 4, and only 1 during visit 2. So, detectability is not perfect. The patterns in our data might lead us to

[1] *With thanks for content to Max Post van der Burg, Therese Donovan, James Hines, and Marc Kéry*

suspect that there may be more animals than two at Site 1. And, that is the basic idea of mixture models—using the patterns of how many animals were seen at a site to estimate the actual number of animals at that site.

Previous to the advent of *N*-mixture models and more rigorous methods to estimate abundance at sites, biologists used the **mean count** (for Site 1, mean count during 4 surveys = 0.75) or **maximum count** (for Site 1, maximum count = 2) to represent the abundance at each point. We sometimes refer to that representation of abundance to compare to the response at other locations as **relative abundance**. But, these types of counts will obviously under-estimate the true abundance if $p \neq 1.0$. Of more concern is the fact that raw counts taken in dense habitats may be done under conditions of lower detection probability than raw counts in sparsely vegetated areas. So, the relative abundance metrics may be biased at best and not comparable at worst.

A mixed landscape

In Chapter 15, we defined the **latent state** of our sampling point as the true state of existence that is hidden or concealed. Specifically to survey sampling, the latent state is the true number of animals at a given sampling site. We conduct surveys to try to estimate how many animals there are, as we do not know the latent state, by definition.

Occupancy models (Chapter 15) assume a binomial distribution of both the latent state and the observations (1's and 0's for present or absent). *N*-mixture models, in similar fashion, assume a binomial distribution for the observations (each animal is either seen or not seen). But, the latent state is assumed to be distributed as Poisson. The Poisson distribution is often used for "count data", and the Poisson distribution can be used to describe the probability of x individuals at a site, given an average number of λ individuals at each site.

Before we look at the distribution in more detail, we can take an example of raw data (summarized as encounter histories) collected at 5 hypothetical sample points during 3 sampling time periods:

Point	Latent State	Data (Encounter Histories)
1	4	0-0-4
2	2	0-1-1
3	3	0-0-0
4	0	0-0-0
5	9	5-6-2

Point 1 had 4 individuals near it (the unknown latent state), and all 4 were seen—but only on the third survey. The two individuals (latent state) at Point 2 were never seen as a group—only one individual seen during the last two time periods. The three individuals at Point 3 were never seen. Point 4 had no individuals (latent state), so none could be detected. And, the observer only saw a portion of the 9 individuals at Point 5 during each of the three surveys.

***N*-mixture analyses assume closure of individuals during all survey periods** (all individuals are present at the same spot through time, which differs from the species-level closure

assumption for occupancy analyses—see Chapter 15), so we can begin to see that our maximum likelihood estimate for detection probability will not be 100% for our 5-point example above. And, we can further guess that the estimate for *N* at each point will be greater than the number of individuals seen during the survey periods.

A bit about the Poisson

It is important to understand the general idea of the Poisson distribution, just as we have described the binomial distribution in previous chapters. The Poisson distribution has one parameter, λ. And, we can define λ as the average number of animals at each location (unknown—the latent state).

The probability density function f(*x*) for the Poisson describes the probability of having *x* animals at a site, when the average number of animals at a site is λ (Figure 18.2). You might conceptualize this by asking the question—*if I go to a random point in my study area, how many animals am I most likely to see*? The peak of the probability density function will give you your answer, and you can see how probable other counts are predicted to be.

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

The Poisson predicts, by its shape and function, that there will be (with much higher chance) λ individuals at each point we sample. That is, the most common number of individuals at sites should be equal to the mean, λ. In contrast, there is a very low probability of seeing a number of individuals that is very far from the mean (either lower or higher). Of course, the number of individuals at a site is not known, and we are trying to estimate λ.

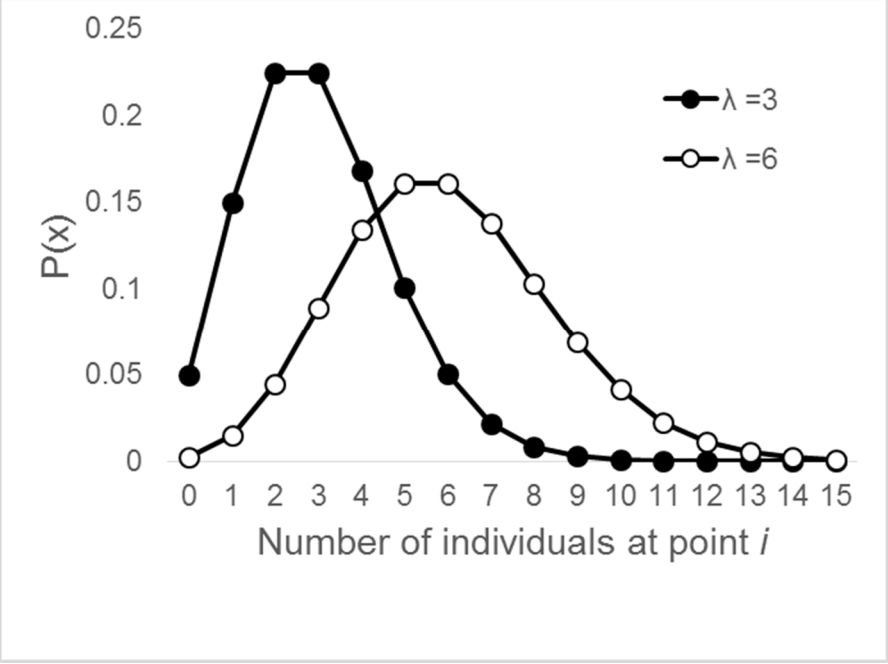

Figure 18.2: *Probability density functions, P(x), describing the distribution of animals at all sample points when the average number of individuals (λ) at all points is either 3 or 6.*

As an example, Figure 18.2 shows two probability density functions—one for λ=3 and one for λ=6. The peaks of the functions maximize at either 3 or 6, so the functions show that it is most probable for sites to have local numbers of animals of either 3 or 6. Although it is possible to have 10 animals at a site when λ=3 or λ=6, but it is very unlikely.

Mixture sampling might initially seem similar to closed mark-recapture methods, because we estimate *N*. However, mixture modeling is conducted in a completely different context than mark-recapture methods—the underlying distribution of animals on the landscape is assumed to vary with some pattern during mixture-based analyses (Figure 18.3, right). And, mixture-based analyses use point-specific samples to estimate the variability in *N*. In contrast, mark-recapture methods use a cohort of marked animals within a study area to estimate the population size for a study site, and *N* is assumed homogeneous for the entire site (Figure 18.3, left).

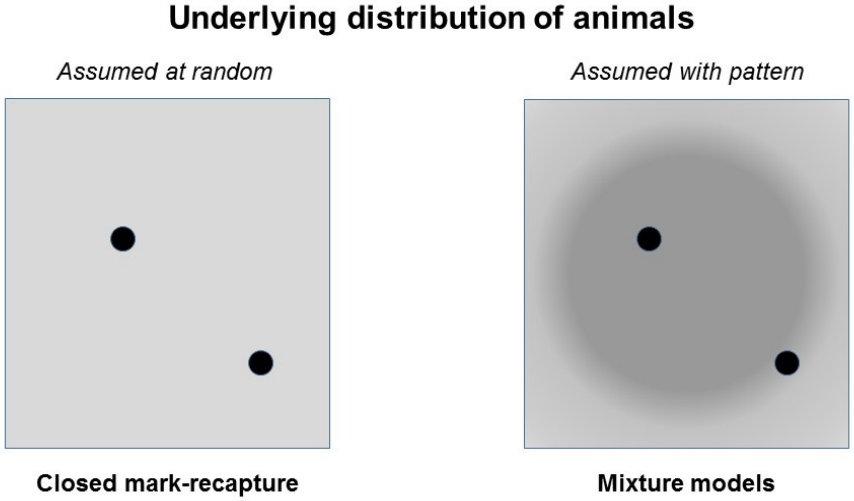

Figure 18.3: *Comparison of the distribution of animals under closed mark-recapture (left) and mixture models (right). The black dots represent sample points. The transition in shading on the right figure indicates changes in population density within the area, while a constant distribution of animals is assumed for closed mark-recapture analyses.*

To further explore this difference, we can transition from a graphical description to a statistical definition—and we will discover the reason for the name, **mixture model**.

Statistically, we can define $E(x_i)$ as the expected count (of your data). In closed mark-recapture models, the expected count of captured animals from a set of traps or nets is a function of the population size, *N*, and the encounter rate, *p*. We can write this as:

$$E(x_i) \sim f(N, p_i)$$

In contrast, *N*-mixture models are structured such that *N* is a function of two distributions. First, we can state, similar to mark-recapture models, that our expected count at a sample site is a function of the population size at the site, *N*, and the encounter rate, *p*. So, we still have:

$$E(x_i) \sim f(N_i, p_i)$$

However, we expect local abundance (at each sample point) to vary across our study's surface, and we define local abundance as λ_i. Thus, the population size, N, is a function of the cumulative local abundance estimates, so

$$N \sim g(\lambda_i)$$

To visualize the mathematics of how we assess a population with N-mixture models, consider Figure 18.4. We see that the local abundance (λ_i) varies spatially (represented by size of light-colored circles)—perhaps a covariate in our analysis would describe why λ_i is larger at some points and smaller at other points. And, we see that the observations (x_i, dark-colored circles) at many locations are incomplete (p<1.0). N-mixture models have become popular methods for analysis of survey data, because the analysis allows a focus on 'ecological processes' or habitat associations rather than on the estimation of density.

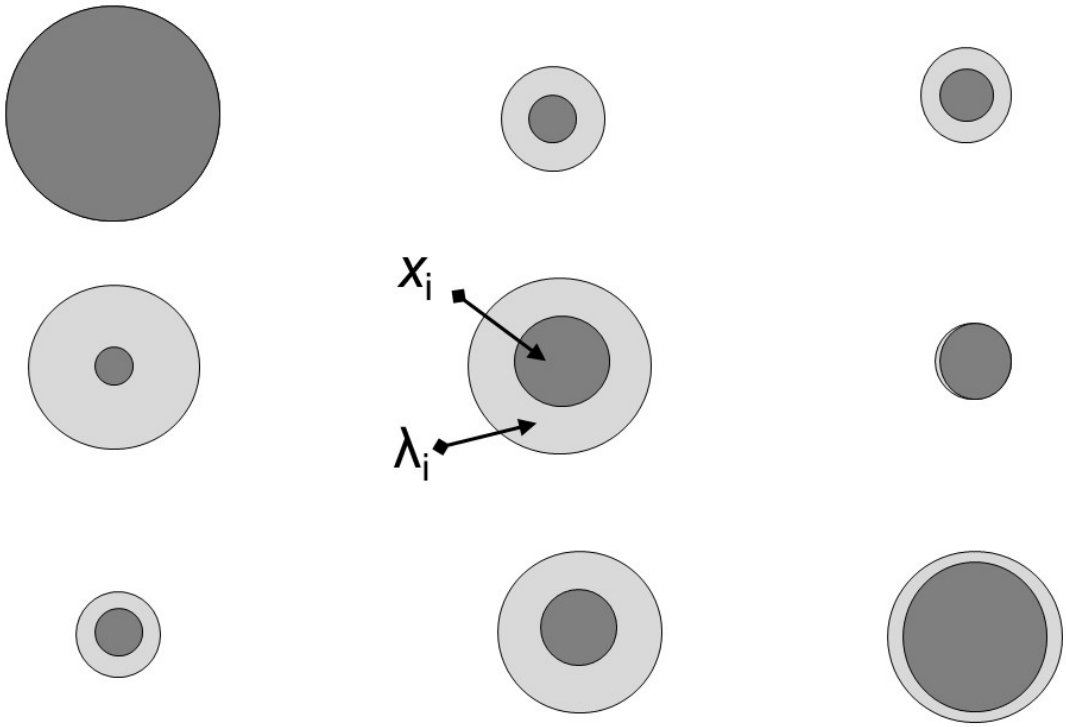

Figure 18.4: *Illustration of the concept of local abundance (λ_i) used in analyses with N-mixture models. Survey sample locations are shown as a grid of points. The general structure for these analyses can be seen as 'metapopulations' each with a local abundance (lighter polygons, λ_i) of animals, of which only a portion are observed during surveys (darker polygons, x_i).*

N-mixture model assumptions

- The distributions assumed (binomial and Poisson) are true
- Observers do not double-count at a sample point within a sampling session
- The number of animals inhabiting one site is random and independent of the number of animals at other sites
- Population is closed between surveys
- Detection is constant across sites, unless modeled by covariates
- All N_i individuals at occasion k have same detection probability p_{ik}

Research designs that use *N*-mixture models for analysis do not require identification of individuals between occasions, as no matching of specific individuals is required (see Chapter 16 on double-observer methods). However, the 'identity' of individuals within a sampling session must be known to the extent that surveyors must keep track of individuals and not double-count during the survey. Because of the assumptions about detection probability, we can see that variation in detection because of distance from the point is not incorporated explicitly into *N*-mixture models (see distance sampling in Chapter 19). Therefore, some ecologists may choose to use fixed-radius sampling to minimize the problem with individuals seen, or not seen, at distances farther from the point.

The closure requirement for *N*-mixture modeling is much more severe than for occupancy modeling. If the closure assumption is violated, estimates of *N* may instead refer to a super-population in some manner, although interpretation of such an estimate is difficult. To ensure closure is not violated, ecologists should use careful planning during the design stage with special attention to the duration of the study. The study period should be short, relative to the movement dynamics expected of the study organism in the target system. If closure problems are realized during or after data collection, ecologists have three options—they may discard data from early or late time periods to shorten the study, they may attempt to use some of the "open" mixture models that are now available (but see our caution below), or they may decide to switch to occupancy modeling to take advantage of the more relaxed closure assumption at the species level, rather than the individual.

Mixing it up: the binomial portion of the likelihood

N-mixture models *are literally a* ***mixture*** *of different distributions.* The likelihood statement for the binomial portion of the *N*-mixture model, below, describes the probability of seeing *y* animals, given a detection probability, *p,* (probability of detecting an animal given it is present and available), and a given population of animals at the site (*N*). The data we collect is *y* in this likelihood description. Our software will use maximum likelihood estimation to estimate *p* given our set of samples.

$$f(y \mid N, p) = \binom{N}{y} p^y (1-p)^{N-y}$$

For a mental exercise, let's assume that $N = 10$. If our probability of detecting an individual is high (e.g., p=0.9), do you think it is more likely that we would see 2 animals or 8 animals of the 10 that are possible to see?

Logically, it makes sense that if we have high detection probabilities, it should be more likely (our working definition of "maximum likelihood") to see 8 of the 10 animals than it would be to see 2 of the 10 animals. And, it is easy enough to check it (the more we use MLE methods, the less scary they seem!):

$$f(y \mid 10,0.9) = \binom{10}{2} 0.9^2 (1-0.9)^{10-2} = 0.0000003645$$

$$f(y \mid 10,0.9) = \binom{10}{8} 0.9^8 (1-0.9)^{10-8} = 0.1937102445$$

Mixing it up: adding the Poisson portion

The entire likelihood statement for the *N*-mixture model is shown here:

$$p(y_{ik} \mid \lambda, p, N_i) = \prod_{i=1}^{R} \left(\prod_{j=1}^{S} Bin(y_{ik} \mid N_i, p) Pois(N_i \mid \lambda) \right)$$

Where

y_{ik}	= observed counts from site *i* and visit *k*
N_i	= site-specific unobserved ('true') abundances
p	= the overall detection rate (estimated from data)
λ	= overall unobserved mean density per site

To understand this likelihood statement, we start on the left side of the equality (=) with "$p(\ldots)$". We read this as the **probability of a count at a site and time, given a mean abundance (λ), detection probability (*p*), and population at the site (*N*)**.

On the right side of the equation, we can see a binomially distributed portion ("*Bin*"). This portion of the likelihood describes the binomial-distributed **probability of detecting *y* animals, given *N* animals at the site and a detection rate (*p*)**.

The Poisson portion ("*Pois*") of the statement describes the Poisson-distributed **probability that there are actually *N* individuals at site *i*, given that the mean abundance across all sites is λ.**

When we throw all of this together, we can see that we should weigh the probability of observing a certain number of animals over *k* visits (binomial portion) by the probability that the site contains *N* animals (Poisson portion). Our software will use maximum likelihood-based methods to estimate values of λ, *p*, and *N* which are the most likely, given your observations (*y*).

Adaptive radiation of mixture modeling: a caution

The use of mixture models is rapidly advancing in ecology. And, recent years have seen a large increase in the number of modifications to the basic methods we describe here. Examples include:

- Mixture methods that allow spatial autocorrelation between sample points (Royle et al. 2007a, Post van der Burg 2011)—the methods we have described in this chapter, above, assume independence among observations from different sample points.

- Mixture methods to estimate species richness (Dorazio and Royle 2005, Royle et al. 2007b, Royle and Dorazio 2008).

- 'Open' *N*-mixture models that allow violations of the basic closure assumption (Dodd and Dorazio 2004, Chandler et al. 2011, Dail and Madsen 2011).

The reader will no doubt be able to find additional modifications with a quick literature search. We caution you with regard to jumping to the newest, most complicated model structure—as with other methods in this book (e.g., multi-state models, robust design models), the complicated methods literally become complicated by adding parameters. "Open" *N*-mixture models have added recruitment and extinction-type parameters, for example—and more parameters to estimate requires you to collect more data, all else equal! We encourage the reader to design your study to use the simplest method possible to answer the question(s) you have. We caution against turning to complicated "open" models, for example, as a remedy for closure violations that could have been prevented with simple changes to the study design.

Conclusion

N-mixture models can be useful for count-surveys that result in a large number of 0's for a species, which can be a result of small populations or low detection rates. The lack of data, in these situations, may make it difficult to use other "standard" approaches such as distance sampling, double-sampling, removal methods, or mark-recapture. *N*-mixture models perform better in these situations because there are multiple visits, during which an animal may be detected that was not detected previously (resulting in a 0-1-2-0 encounter history, perhaps). If species are extremely rare (with many unoccupied sites that result in 0-0-0-0 encounter histories), *N*-mixture models may not be able to estimate parameters—the biologist might consider the use of occupancy models as site occupancy is often more desired for rare species than local abundance predictions. However, stay true to your initial goals for your research!

N-mixture models do have a closure assumption that should not be violated. And, the multiple visits, through time, to multiple sites do require planning and consideration of the effort needed. Other methods, such as distance sampling, removal, or double-observer can provide estimates of either density or abundance with one temporal survey, but the survey must result in large number of detected individuals. And, distance, removal, and double-observer are intensive type of surveys compared to simply counting individuals for use with *N*-mixture models. We

encourage you to evaluate the information in the table at the end of Chapter 14 to assist with your research design and planning.

References

Chandler, R. B., J. A. Royle, and D. I. King. 2011. Inference about density and temporary emigration in unmarked populations. Ecology 92:1429-1435.

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

Dodd, C.K. Jr., and R. M. Dorazio. 2004. Using counts to simultaneously estimate abundance and detection probabilities in a salamander community. Herpetologica 60:468–478.

Dorazio, R. M., and J. A. Royle, J. A. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. Journal of the American Statistical Association 100: 389-398.

Kéry, M. 2013. Introduction to N-mixture models: Short course. EURING Technical Meeting, Athens, Georgia, USA (28 April 2013).

Post van der Burg, M., B. Bly, T. VerCauteren, and A. J. Tyre. 2011. Making better sense of monitoring data from low density species using a spatially explicit modelling approach. Journal of Applied Ecology 48:47-55.

Royle, J. A. 2004. N-Mixture Models for Estimating Population Size from Spatially Replicated Counts. Biometrics 60: 108-115.

Royle, J. A., M. Kéry, R. Gautier, and H. Schmid. 2007a. Hierarchical spatial models of abundance and occurrence from imperfect survey data. Ecological Monographs 77:465-481.

Royle, J. A., R. M. Dorazio, and W. A. Link. 2007b. Analysis of multinomial models with unknown index using data augmentation. Journal of Computational and Graphical Statistics 16:67-85.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Academic Press.

For more information on topics in this chapter

McKenney, H, T. M. Donovan, and J. Hines. 2007. Repeated count model (Royle). In Donovan, T. M. and J. Hines, eds. Exercises in occupancy modeling and estimation. Online: http://www.uvm.edu/envnr/vtcfwru/spreadsheets/occupancy.htm

Citing this primer

Powell, L. A., and G. A. Gale. 2015. Estimation of Parameters for Animal Populations: a primer for the rest of us. Caught Napping Publications: Lincoln, NE.

A biologist conducts a visual and auditory survey for songbirds in a grassland in northern Nebraska, USA. Photo by Silka Kempema, University of Nebraska-Lincoln.