Estimating Size and Composition of Biological Communities by Modeling the Occurrence of Species[1]

Robert M. Dorazio, Florida Integrated Science Center, U. S. Geological Survey, Gainesville, Florida 32653, *email:* bdorazio@usgs.gov

J. Andrew Royle, Patuxent Wildlife Research Center, U. S. Geological Survey, Laurel, Maryland 20708, *email:* andy_royle@fws.gov

Abstract

We develop a model that uses repeated observations of a biological community to estimate the number and composition of species in the community. Estimators of community-level attributes are constructed from model-based estimators of occurrence of individual species that incorporate imperfect detection of individuals. Data from the North American Breeding Bird Survey are analyzed to illustrate the variety of ecologically-important quantities that are easily constructed and estimated using our model-based estimators of species occurrence. In particular, we compute site-specific estimates of species richness that honor classical notions of species-area relationships. We suggest extensions of our model to estimate maps of occurrence of individual species and to compute inferences related to the temporal and spatial dynamics of biological communities.

KEY WORDS: Biodiversity; Breeding bird survey; Conservation; Detection heterogeneity; Occurrence heterogeneity; Site occupancy models; Species composition; Species richness.

[1] The authors thank the editors and two referees for helpful comments.

1 Introduction

The conservation or management of biodiversity requires accurate assessments of the number and composition of species in biological communities (Colwell and Coddington, 1994). Interestingly, these quantities are usually unobservable in studies of natural communities. In most situations practical considerations preclude an exhaustive search for every species in the community, and samples are used to *estimate* the number and composition of species. However, samples are unlikely to reveal every species in the community owing to incomplete coverage (particularly when samples represent a small portion of the community) and to imperfect detection of individual plants or animals.

The importance of imperfect detection is well known and has motivated the development of a variety of statistical methods for estimating the number N of distinct species in a community (Bunge and Fitzpatrick, 1993; Nichols and Conroy, 1996). (In the ecological literature N is a measure of a community's *species richness*). In biological communities detection rates often vary considerably among species owing to differences in abundance or behavior of individual species and to differences in an observer's ability to identify a species visually or aurally (Boulinier, Nichols, Sauer, Hines, and Pollock, 1998). These unobservable sources of variation in detection are often specified with mixture models, which include a latent distribution to parameterize the heterogeneity in detection among species (Agresti, 1994; Norris and Pollock, 1996; Coull and Agresti, 1999; Fienberg, Johnson, and Junker, 1999; Pledger, 2000; Basu and Ebrahimi, 2001; Tardella, 2002; Dorazio and Royle, 2003).

Several of these models were developed for estimating the size of a population from repeated observations of individuals in the population (e.g., a capture-recapture study). However, in the context of estimating species richness, these models also may be fit to the detections of species encountered at different locations during a single sampling of a community (Burnham and Overton, 1979; Nichols and Conroy, 1996; Dorazio and Royle, 2003). In this case the locations are selected as a (possibly random) subset of those within a region assumed to contain the community. The sample locations are therefore regarded as replicate units of observation.

At each location a list of species actually detected is recorded, and non-detections of species observed elsewhere in the sample are determined for each location at the completion of the survey. Therefore, for each species observed in the sample, a vector of observations may be constructed to indicate locations where the species was detected (e.g., $(0, 0, 1, 1, 0)$ for detections at locations 3 and 4). This construction of an observation vector for each species tacitly assumes that *every* species in the community occurs and is available to be detected at *every* sample location. In other words, a 0 in the observation vector of a species denotes that the species is not detected at that location, *not* that the species is absent from that location.

This view of sample locations as replicate observations has allowed much progress to be made in studies of avian biology and community ecology (Nichols, Boulinier, Hines, Pollock, and Sauer, 1998a,b; Doherty, Sorci, Royle, Hines, Nichols, and Boulinier, 2003); however, in many, if not most, biological communities, the assumption that all species are present at all sample locations is scientifically untenable. As the size of a region increases (with or without concomitant increases in habitat heterogeneity), we a priori expect that a species may be present at some locations and absent from others; thus, we believe it is far more realistic to consider sampling designs and models that allow the occurrence of a species to vary with location. Furthermore, such considerations permit estimates of the number of species occurring within an area spanned by locations in the sample to increase with the size of the area sampled. This view is consistent with classical notions of species-area relationships (Preston, 1960, 1962a,b; MacArthur and Wilson, 1967; Connor and McCoy, 1979), which provide various explanations for the apparent increase in number of species with an increase in the area sampled.

In this case study we develop a model for estimating the size and composition of a biological community based on an elemental model of occurrence of species that also incorporates imperfect detection of individuals. Modeling the occurrence of species is an interesting problem in its own right (Bayley and Peterson, 2001; MacKenzie, Nichols, Lachman, Droege, Royle, and Langtimm, 2002); however, to our knowledge estimators of community-level attributes, such as species richness, and species-level attributes, such as occurrence, have never been combined in the same modeling framework. In our model, rates of detection *and* occurrence are assumed to vary among species, and every species is *not* assumed to be present at every sample location; therefore, the model honors classical notions of species-area relationships. To estimate rates of detection and occurrence jointly, each of the multiple locations that are representative of the community must be sampled repeatedly and the total duration of the survey must be kept sufficiently short that N may be assumed constant (i.e., local extinctions or colonizations of species are unlikely).

We illustrate this design and our model-based inferences using data from the North American Breeding Bird Survey (summarized in Section 2). Development of the model and construction of several, ecologically-important estimands are described in Section 3. Computational methods for estimating model parameters and functions of these parameters are summarized in Section 4. An analysis of data from the North American Breeding Bird Survey is provided in Section 5. Benefits and extensions of our modeling approach are discussed in Section 6.

2 North American Breeding Bird Survey (BBS)

The BBS is a continental-scale survey of birds that has been conducted since 1966 and includes more than 4000 roadside routes (primary sample units) located in North America (Robbins, Bystrak, and Geissler, 1986; Robbins, Sauer, Greenberg, and Droege, 1989; Sauer, Pendleton, and Peterjohn, 1996). Each route is 39.4 km and contains 50 equally-spaced sites. At each of these sites an observer records the number and identity of each species detected (visually or aurally) within a 3-minute period.

Historically, BBS counts have been analyzed as surrogates of abundance in an attempt to infer spatial and temporal variation in the abundance of different bird species (James, McCulloch, and Wiedenfeld, 1996; Link and Sauer, 1997, 1998). More recently, site-specific counts of an individual species have been quantized to indicate only whether each species was detected at each sampled location, and analyses of these quantal responses have been used in investigations of avian community ecology (e.g., Nichols et al., 1998a,b; Cam, Nichols, Sauer, Hines, and Flather, 2000; Boulinier, Nichols, Hines, Sauer, Flather, and Pollock, 2001; Doherty et al., 2003). Our model of occurrence is likewise developed for site-specific, quantized counts of individual species.

In the conventional BBS sampling protocol, each roadside route is visited only once annually; however, in 1991 several routes were sampled repeatedly during the breeding season to evaluate the variation in bird counts both between and within sites (Link, Barker, Sauer, and Droege, 1994). The data used in our analysis were collected at one of these routes located in Maytown, Alabama (BBS route number 017). This route was visited by the same observer on 11 different days in the month of June. During this time 75 species of birds were detected and there was considerable variation among species in the observed frequencies of detection at each site (Figure 1).

3 Model Description

3.1 Preliminaries and definitions

Let N denote the unknown number of distinct species that occupy a prescribed region, and suppose J representative sites within this region are selected for sampling. If M denotes the total number of species that are present among all J sample locations, we note that $M \leq N$ by definition. As the number of locations J in the sample increases, M approaches N, the total size of the community; therefore, M can be interpreted as an ordinate of a species-area curve (Preston, 1960, 1962a,b; MacArthur and Wilson, 1967; Connor and McCoy, 1979) whose asymptote is N.

We consider surveys wherein each of the J sites is visited several times and the identities of all species detected during each visit is noted. The total duration of the survey must be sufficiently short that N may safely be assumed to remain constant in the time required to complete the survey. Therefore, the traditional "closure" assumption, which precludes an addition (or subtraction) of species in the community as a consequence of local colonization (or extinction) events, is satisfied.

Let x_{ij} denote the number of times that species $i \, (= 1, \ldots, N)$ is detected in K visits to site $j \, (= 1, \ldots, J)$. For clarity, we assume that K is identical at each of the J sites, but a balanced design is not an essential part of the survey. Repeated observations ($K > 1$) at each site *are* crucial, however, because separate parameters for the occurrence and detection of each species are not identifiable in the absence of such replication (see Section 3.2).

Let $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ denote a vector of the J site-specific observations of species i. At the completion of the survey, suppose $n < N$ distinct species are actually detected. For our purposes it is convenient to order the observation vectors as follows:

$$\boldsymbol{x}_i \in \{(K+1)^J - 1 \text{ observable vectors}\} \text{ for } i = 1, \ldots, n$$
$$\boldsymbol{x}_i = \mathbf{0} \text{ for } i = n+1, \ldots, N.$$

This ordering implies a partitioning of the $N \times J$ matrix of observation vectors ($\boldsymbol{X}$) into an observed portion $\boldsymbol{X}_n$ (the first n rows of $\boldsymbol{X}$) and an unobserved portion, which includes species that are undetected in the survey. Based on this partitioning, it is also useful to consider an $N \times J$ matrix of binary indicators $\boldsymbol{Z}$, whose elements denote the presence ($z_{ij} = 1$) or absence ($z_{ij} = 0$) of species i at site j. Note that $\boldsymbol{Z}$ is only partially observed. A species must be present at a site before it can be detected; therefore, z_{ij} must equal 1 if $x_{ij} > 0$. However, if $x_{ij} = 0$, two mutually exclusive possibilities determine the value of z_{ij}: (1) species i is present at site j but undetected ($z_{ij} = 1$), or (2) species i is absent at site j ($z_{ij} = 0$). In Section 3.4 we show that by allowing the occurrence of species to vary spatially (i.e., among sampling sites) through the definition of $\boldsymbol{Z}$, the construction of many ecologically-important estimands is greatly simplified.

3.2 Modelling heterogeneity in occurrence and detection of species

We first develop a model of the site-specific detections of a single species by conditioning on the probabilities of occurrence and detection of that species. Our development is similar to that used in the logistic-normal model of heterogeneous detectability of species (Coull and Agresti, 1999; Fienberg et al., 1999); however, the logistic-normal model conditions only on the site-specific detection probability of each species, implicitly assuming that the species is present and available to be detected at every sample location. In contrast, the model developed here does not assume that each species occurs at each site with probability one.

Let ψ_{ij} denote the probability of occurrence of species i at site j and θ_{ij} denote the probability of detection of species i, given that it occurs at site j. We assume that the indicators of occurrence are independent outcomes of a Bernoulli process with density function

$$p(z_{ij} \mid \psi_{ij}) = \psi_{ij}^{z_{ij}} (1 - \psi_{ij})^{1-z_{ij}}. \tag{1}$$

In addition, we assume that if species i occurs at site j ($z_{ij} = 1$), the number of detections is assumed to have a Binomial(K, θ_{ij}) distribution

$$p(x_{ij} \mid z_{ij}, \theta_{ij}) = \left[\begin{pmatrix} K \\ x_{ij} \end{pmatrix} \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{K - x_{ij}} \right]^{z_{ij}}. \tag{2}$$

On the other hand, if species i is absent at site j ($z_{ij} = 0$), then x_{ij} is assumed to equal zero with probability one. Multiplying (1) and (2) yields the joint density of the observed number of detections and the indicator of occurrence

$$p(x_{ij}, z_{ij} \mid \theta_{ij}, \psi_{ij}) = \left[\begin{pmatrix} K \\ x_{ij} \end{pmatrix} \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{K - x_{ij}} \right]^{z_{ij}} \psi_{ij}^{z_{ij}} (1 - \psi_{ij})^{1-z_{ij}}. \tag{3}$$

However, since z_{ij} is only partially observed, its removal (by summation in (3)) conveniently provides the marginal density of the (fully) observed number of detections

$$p(x_{ij} \mid \theta_{ij}, \psi_{ij}) \;=\; \psi_{ij} \begin{pmatrix} K \\ x_{ij} \end{pmatrix} \theta_{ij}^{x_{ij}} (1 - \theta_{ij})^{K - x_{ij}} + (1 - \psi_{ij})\, I(x_{ij} = 0) \tag{4}$$

where $I(\cdot)$ denotes an indicator function. Note that (4) specifies the density of a zero-inflated binomial outcome. Under this model, if species i is not detected at site j (i.e., $x_{ij} = 0$), species i is either absent (with probability $(1 - \psi_{ij})$) or present but undetected (with probability $\psi_{ij}(1 - \theta_{ij})^K$).

Having developed a model of the site-specific detections of a single species, we now extend the model to combine information among different species in the community. In particular the effects of species- and site-specific differences in rates of occurrence and detection are parameterized on the logit scale as follows:

$$\begin{aligned} \text{logit } \psi_{ij} &= u_i + \alpha_j \\ \text{logit } \theta_{ij} &= v_i + \beta_j \end{aligned}$$

where u_i and v_i denote species-level effects, and α_j and β_j denote site-level effects. The species-level effects are assumed to be centered at zero; therefore, α_j denotes a logit-scale parameter for the mean probability of occurrence among all species at site j, and β_j denotes a logit-scale parameter for the mean probability of detection among all species at site j. A linear combination of parameters and site-level covariates may be substituted for α_j or β_j, assuming of course that such covariates are available and are thought to be informative about the magnitude of ψ_{ij} or θ_{ij}. However, in the absence of site-level covariates, we assume that α_j and β_j have constant values, say α and β, at each of the J sites.

Species-specific differences in the probabilities of occurrence and detection are modeled by specifying a parametric form for the joint distribution of u_i and v_i. For example, we assume $[u_i, v_i \mid \Sigma] \sim \text{Normal}(0, \Sigma)$, which allows us to specify the heterogeneity in occurrence and detection among species using only a few parameters (specifically, σ_u^2, σ_v^2, and σ_{uv}, the unique elements of Σ). These parameters are not constrained in any way (except to ensure positive definiteness of Σ); however, we anticipate that estimates of σ_{uv} will almost surely be positive because probabilities of occurrence and detection are both expected to increase as the abundance of a species increases. For example, if we let A_{ij} denote the abundance of species i at site j and assume $A_{ij} \sim \text{Poisson}(\lambda_{ij})$, then the probability of occurrence is expected to increase with mean abundance λ_{ij} as follows: $\psi_{ij} = \Pr(A_{ij} > 0) = 1 - \exp(\lambda_{ij})$. Similarly, if we let q_{ij} denote the probability of detecting each of the A_{ij} individuals present at site j and assume that such detections are independent, then the probability of detecting any of these individuals increases with A_{ij} as follows: $\theta_{ij} = 1 - (1 - q_{ij})^{A_{ij}}$. Our expectation of positive estimates of σ_{uv} is therefore based on simple, yet entirely reasonable, arguments.

3.3 Likelihood functions

The marginal density of the observed number of detections of species i at site j (4) may be rewritten in terms of the logit-scale parameters as follows:

$$f(x_{ij} \mid u_i, v_i, \alpha_j, \beta_j) = \left(\frac{\exp(\phi_{ij})}{1+\exp(\phi_{ij})}\right) \binom{K}{x_{ij}} \frac{\exp(x_{ij}\eta_{ij})}{\left(1+\exp(\eta_{ij})\right)^K} + \left(\frac{1}{1+\exp(\phi_{ij})}\right) I(x_{ij}=0) \tag{5}$$

where $\phi_{ij} = u_i + \alpha_j$ and $\eta_{ij} = v_i + \beta_j$. Assuming that the J observations of each species are (conditionally) independent, the marginal probability of the observation vector $\boldsymbol{x}_i$ is

$$q(\boldsymbol{x}_i \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{J} f(x_{ij} \mid u_i, v_i, \alpha_j, \beta_j)\right] g(u_i, v_i \mid \Sigma)\, \mathrm{d}u_i\, \mathrm{d}v_i \tag{6}$$

where $g(u_i, v_i \mid \Sigma)$ specifies the bivariate normal density assumed for u_i and v_i. The integrations in (6) are easily approximated with an adaptive form of Gauss-Hermite quadrature (Liu and Pierce, 1994; Pinheiro and Bates, 1995) or with stochastic methods (e.g., Monte Carlo), although these methods can be computationally intensive to implement. More importantly, estimates of the u_i parameters and their uncertainties are actually needed in our problem to estimate quantities that are important in community ecology (see Section 4). Therefore, we derive a likelihood function that uses the marginalization in (6) sparingly and allows estimates of N to be computed as a function of model parameters. This approach is particularly useful in cases where prior information about the probable size of the community is unavailable.

Since the number of distinct species N is unknown and the last $N - n$ rows of $\boldsymbol{X}$ are unobserved, the multinomial likelihood for classifying the $(K + 1)^J$ possible values of $\boldsymbol{x}_i$ must include N as a parameter (Fienberg et al., 1999) as follows:

$$\begin{aligned} p(\boldsymbol{X}_n, \boldsymbol{n}, \boldsymbol{u}, \boldsymbol{v} \mid N, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma) &= \frac{N!}{(N-n)! \prod_h n_h!} \, q(\boldsymbol{0} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma)^{(N-n)} \\ &\cdot \prod_h \left\{ \left[\prod_{j=1}^{J} f(x_{hj} \mid u_h, v_h, \alpha_j, \beta_j) \right] g(u_h, v_h \mid \Sigma) \right\}^{n_h}, \end{aligned} \tag{7}$$

where $n_h = \sum_{i=1}^{n} I(\boldsymbol{x}_i = \boldsymbol{x}_h)$ denotes the number of species that share detection sequence $\boldsymbol{x}_h$ $(h = 1, \ldots, m)$ and $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{n} = (n_1, \ldots, n_m)$ each denote a vector of $m \leq n$ elements that correspond to the distinct values of $\boldsymbol{x}_h$ observed in the sample. Thus, all species that share a common detection sequence $\boldsymbol{x}_h$ are assumed to have the same species effects u_h and v_h.

The multinomial likelihood may be factored into 2 components (Sanathanan, 1972), one for the detections of the n observed species

$$\begin{aligned} p(\boldsymbol{X}_n, \boldsymbol{n}, \boldsymbol{u}, \boldsymbol{v} \mid n, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma) &= \frac{n!}{\prod_h n_h!} \left(\frac{1}{1 - q(\boldsymbol{0} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma)} \right)^{n} \\ &\cdot \prod_h \left\{ \left[\prod_{j=1}^{J} f(x_{hj} \mid u_h, v_h, \alpha_j, \beta_j) \right] g(u_h, v_h \mid \Sigma) \right\}^{n_h}, \end{aligned} \tag{8}$$

and a second for the binomial distribution of n given the unknown size N of the community,

$$p(n \mid N, \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma) = \frac{N!}{(N-n)! \, n!} \left(1 - q(\boldsymbol{0} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma)\right)^{n} q(\boldsymbol{0} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \Sigma)^{N-n}. \tag{9}$$

An important advantage of this factorization is that it allows an estimate of N to be computed as a function of model parameter estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, and $\hat{\Sigma}$ using

$$\hat{N} = \frac{n}{1 - q(\boldsymbol{0} \mid \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\Sigma})} \tag{10}$$

Similarly, in the next section we show that other ecologically important estimands also may be computed as functions of model parameter estimates.

3.4 Ecologically important estimands

In Section 3.1 we defined a partially observed, binary indicator z_{ij} for the occurrence of species i at site j. Suppose counterfactually that this indicator is fully observed for each of the N species in the community. Then, no model of detectability is needed to estimate N, and formulae for calculating many quantities of importance in community ecology are obvious. The practical utility of these formulae is that they provide obvious estimators of ecologically important estimands in the situation where N and z_{ij} are not known but can be estimated.

For example, suppose we wish to estimate the number of species present at the jth site, say N_j. If N was known and $\boldsymbol{Z}$ was fully observed, N_j could be computed by simply adding the elements in the jth column of $\boldsymbol{Z}$: $N_j = \sum_{i=1}^{N} z_{ij}$. However, in practice neither N nor every element of $\boldsymbol{Z}$ is known. To specify an estimator of N_j, let $\mathrm{E}(z_{ij} \mid \boldsymbol{X}_n) = \hat{\psi}_{ij}$ denote the posterior expectation of the indicator of occurrence of species i at site j. Then, the number of species actually present at site j may be estimated as follows:

$$\begin{aligned} \hat{N}_j &= \sum_{i=1}^{n} [z_{ij} \cdot I(x_{ij} > 0) + E(z_{ij} \mid \boldsymbol{X}_n)\, I(x_{ij} = 0)] + \sum_{i=n+1}^{\hat{N}} E(z_{ij} \mid \boldsymbol{X}_n) \\ &= \sum_{i=1}^{n} \Big[1 \cdot I(x_{ij} > 0) + \hat{\psi}_{ij}\, I(x_{ij} = 0)\Big] + (\hat{N} - n)\, \hat{\psi}_{0j} \end{aligned} \qquad (11)$$

Thus, $\hat{N}_j$ is computed by adding the number of species actually detected at site j ($\sum_{i=1}^{n} I(x_{ij} > 0)$) to the expected occurrence at site j of all undetected species, which includes the $\hat{N} - n$ species not detected at any of the J sites in the sample.

A similar process may be used to derive an estimator of the total number of species present among all J sample locations, which we denote by M. If N and $\boldsymbol{Z}$ were known, we would compute M by subtracting from N the number of species not present at any of the J sites: $M = N - \sum_{i=1}^{N} I(\boldsymbol{z}_i = \mathbf{0}))$. In practice we must estimate M by subtracting from $\hat{N}$ the estimated number of species that are not expected to be present at any of the J sites, as follows:

$$\hat{M} = \hat{N} - (\hat{N} - n) \prod_{j=1}^{J} (1 - \hat{\psi}_{0j}) \qquad (12)$$

Indices of similarity in species composition are often used in the classification of biological communities and provide another set of estimands that are easily calculated in terms of occurrence. For example, consider Dice's (1945) index (also called the coefficient of community (Pielou, 1977)) of the similarity in species at sites j and l

$$S_{jl} = \frac{2\ N_{jl}}{N_j + N_l} \tag{13}$$

where N_{jl} denotes the number of species present at both sites j and l. As before, we recognize that if N and $\boldsymbol{Z}$ were known, N_{jl} would be computed using $N_{jl} = \sum_{i=1}^{N} z_{ij}\, z_{il}$. Therefore, in practice we may estimate N_{jl} as follows:

$$\begin{aligned} \hat{N}_{jl} &= \sum_{i=1}^{n} \Bigg[1 \cdot 1\ I(x_{ij} > 0, x_{il} > 0) + 1 \cdot \hat{\psi}_{il}\ I(x_{ij} > 0, x_{il} = 0) \\ &+ \hat{\psi}_{ij} \cdot 1\ I(x_{ij} = 0, x_{il} > 0) + \hat{\psi}_{ij} \cdot \hat{\psi}_{il}\ I(x_{ij} = 0, x_{il} = 0) \Bigg] \\ &+ (\hat{N} - n)\, \hat{\psi}_{0j} \cdot \hat{\psi}_{0l} \end{aligned} \tag{14}$$

which adds the number of species actually detected at both sites j and l ($\sum_{i=1}^{n} I(x_{ij} > 0, x_{il} > 0)$) to the estimated number of species expected to occur at both sites. Therefore, an estimate of simililarity $\hat{S}_{jl}$ is computed by substituting the estimators (11) and (14) into (13).

All of the estimators presented in this section are developed as "plug-in" estimators without any consideration of their uncertainty. In the next section we describe how Bayesian methods may be used to compute inferences for these ecologically important estimands.

4 Computing Estimates of Model Parameters and Ecological Estimands

The model parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and Σ may be estimated by maximizing a marginal likelihood that eliminates (by numerical integration) the latent parameters $\boldsymbol{u}$ and $\boldsymbol{v}$ from the conditional likelihood (8). However, our primary interest lies in computing inferences for *functions* of model parameters, including $\boldsymbol{u}$ and $\boldsymbol{v}$ which are used to estimate the occurrence and detection of individual species. We therefore prefer a Bayesian framework for parameter estimation and inference, using the conditional likelihood (8).

In the absence of site-level covariates that might be used to model differences in the mean rate of occurrence or detection among sites, we assume $\alpha_j = \alpha$ and $\beta_j = \beta$ as described in Section 3.2. We complete the model by assuming mutually independent prior distributions for α, β, and the unique elements of Σ. In particular, we assume a Uniform(0,1) prior for both $\text{logit}^{-1}(\alpha)$ and $\text{logit}^{-1}(\beta)$, which owing to the logit transformation, implies the following prior densities for α and β:

$$\begin{aligned}\pi(\alpha) &= \exp(\alpha)/(1+\exp(\alpha))^2 \\ \pi(\beta) &= \exp(\beta)/(1+\exp(\beta))^2\end{aligned}$$

To specify a prior distribution for the unique elements of Σ, we use the separation strategy of Barnard, McCulloch, and Meng (2000) and assume independent Inverse-Gamma(a,b) priors for σ_u^2 and σ_v^2 ($a = 0.1$ and $b = 10$ denote shape and scale parameters, respectively) and a Uniform(-1,1) prior for the correlation parameter $\rho_{uv} = \sigma_{uv}/\sigma_u\sigma_v$. The Inverse-Gamma$(\epsilon, 1/\epsilon)$ distribution, for some small ϵ, is often used as a default or objective prior of variance parameters, particularly in models that maintain conjugacy between the prior and posterior distributions (e.g., see Carlin and Louis, 2000, p. 149). The Uniform(-1,1) distribution is similarly used to specify prior indifference in the magnitude of ρ_{uv}. Furthermore, to ensure the positive-definiteness of estimates of Σ, we use a log-Cholesky parameterization of Σ (Pinheiro and Bates, 1996), which implies the following one-to-one transformation of variables: $\gamma_1 = \log\sigma_u$, $\gamma_2 = \sigma_v\rho_{uv}$, and $\gamma_3 = \log\sigma_v + 0.5\log(1-\rho_{uv}^2)$. The prior density induced by this change of variables is

$$\pi(\gamma_1,\gamma_2,\gamma_3) = \frac{2w^{0.5-a}}{(b^a\Gamma(a))^2}\ \exp\left[2\Big(\gamma_3 - \gamma_1 a\Big) - b^{-1}\Big(\exp(-2\gamma_1) + w^{-1}\Big)\right]$$

where $w = \gamma_2^2 + \exp(2\gamma_3)$.

The joint posterior density of the model parameters

$$\pi(\alpha, \beta, \gamma_1, \gamma_2, \gamma_3, \boldsymbol{u}, \boldsymbol{v} \mid \boldsymbol{X}_n, \boldsymbol{n}) \;\propto\; p(\boldsymbol{X}_n, \boldsymbol{n}, \boldsymbol{u}, \boldsymbol{v} \mid \alpha, \beta, \gamma_1, \gamma_2, \gamma_3)\; \pi(\alpha, \beta, \gamma_1, \gamma_2, \gamma_3) \tag{15}$$

is proportional to the product of the likelihood of the data $p(\boldsymbol{X}_n, \boldsymbol{n}, \boldsymbol{u}, \boldsymbol{v} \mid \alpha, \beta, \gamma_1, \gamma_2, \gamma_3)$ (defined in (8) using untransformed parameters) and the prior $\pi(\alpha, \beta, \gamma_1, \gamma_2, \gamma_3) = \pi(\alpha)\,\pi(\beta)\,\pi(\gamma_1, \gamma_2, \gamma_3)$. We use Metropolized Gibbs sampling (Robert and Casella, 1999, section 7.3) to compute an arbitrarily large sample from the joint posterior. Although none of the full conditional distributions used in our Gibbs sampler has a familiar form, random draws are easily computed using the Metropolis algorithm and appropriately-scaled Gaussian proposals because the posterior is supported entirely on $\mathbb{R}^{n \times n \times 5}$.

Once the sample of the joint posterior has been computed, it is straightforward to compute a sample from the posterior distribution of N and the ecological estimands derived in Section 3.4. For example, if we let $\tilde{\alpha}$, $\tilde{\beta}$, $\tilde{\sigma}_u^2$ $(= \exp(2\tilde{\gamma}_1))$, $\tilde{\sigma}_{uv}$ $(= \tilde{\gamma}_2 \exp(2\tilde{\gamma}_1))$, and $\tilde{\sigma}_v^2$ $(= \tilde{\gamma}_2^2 + \exp(2\tilde{\gamma}_3))$ denote a random draw from (15), a corresponding draw from the posterior distribution of N is easily computed using (10). All of the other estimands of community structure are functions of N and rates of occurrence ψ_{ij}. For each of the n species actually observed, a random draw from the posterior of occurrence is easily computed using $\tilde{\psi}_{ij} = 1/[1 + \exp(-(\tilde{u}_i + \tilde{\alpha}))]$. However, occurrence ψ_{0j} of the $N - n$ unobserved species is more difficult to compute because u_0 is not a parameter in the conditional likelihood (8). We compute the posterior expectation $\tilde{u}_0 = \mathrm{E}(u_0 \mid \boldsymbol{x} = \boldsymbol{0})$ as follows

$$\tilde{u}_0 = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_0 \left[\prod_{j=1}^{J} f(0 \mid u_0, v_0, \tilde{\alpha}, \tilde{\beta}) \right] \; g(u_0, v_0 \mid \tilde{\Sigma})\, \mathrm{d}u_0\, \mathrm{d}v_0}{q(\boldsymbol{0} \mid \tilde{\alpha}, \tilde{\beta}, \tilde{\Sigma})} \tag{16}$$

which depends on a single draw $(\tilde{\alpha}, \tilde{\beta}, \tilde{\Sigma})$ from the joint posterior. The numerator in (16) may be computed to any desired level of accuracy using Monte Carlo integration and independent draws from the bivariate Normal specified in $g(u_0, v_0 \mid \tilde{\Sigma})$.

5 Analysis of BBS Data

5.1 Estimates of model parameters and N

We estimated the richness and composition of bird species located along BBS route 017 in Maytown, Alabama by fitting our model of occurrence of individual species (Sections 3 and 4). Table 1 contains estimates of N and model parameters based on summary statistics of a simulated sample from the joint posterior distribution. For comparison, we also include maximum likelihood estimates (MLEs), which were computed by maximizing the marginal likelihood that obtains when $\boldsymbol{u}$ and $\boldsymbol{v}$ are integrated numerically from (8).

The MLEs and Bayesian estimates of model parameters and their uncertainties are quite similar owing to the size of the data set ($n = 75$ species and $J = 50$ sites) and to our choice of priors, which are not intended to be highly informative. For example, the MLE of N (91.4) is very close to the posterior median and mean values of N (91.5 and 93.3, respectively). The estimates in Table 1 also indicate that heterogeneity in occurrence among species ($\hat{\sigma}_u = 2.22$) is about twice as high as heterogeneity in detection ($\hat{\sigma}_v = 1.14$) and that detection failures in many bird species are attributed to low rates of occurrence, as opposed to simply low rates of detection. This result is illustrated clearly by computing the marginal distributions of occurrence and detection probabilities implied by the parameter estimates. For example, in the model fitted to the BBS data, we assumed (marginally) that $\text{logit}(\psi_i) \sim \text{Normal}(\alpha, \sigma_u^2)$ and $\text{logit}(\theta_i) \sim \text{Normal}(\beta, \sigma_v^2)$. By transforming from logits to probabilities, it is easily shown that the marginal density functions for the probabilities of occurrence and detection are

$$p(\psi_i \mid \alpha, \sigma_u^2) = \frac{1}{\sqrt{2\pi}\,\psi_i(1-\psi_i)\sigma_u} \exp\left[-\frac{1}{2\sigma_u^2}\Big(\text{logit}(\psi_i) - \alpha\Big)^2\right] \tag{17}$$

and

$$p(\theta_i \mid \beta, \sigma_v^2) = \frac{1}{\sqrt{2\pi}\,\theta_i(1-\theta_i)\sigma_v} \exp\left[-\frac{1}{2\sigma_v^2}\Big(\text{logit}(\theta_i) - \beta\Big)^2\right] \tag{18}$$

respectively. Thus, substituting the estimates of α, β, σ_u, and σ_v into (17) and (18) yields the estimated distributions of occurrence and detection probabilities (Figure 2). Although high probabilities of occurrence are not uncommon, there are many bird species for which $\psi \ll 1$; thus, observations of these species are unlikely, regardless of how easily they might be detected when present. This is an important finding because the ecologically relevant estimands are functions of occurrence, ψ. The detection parameter θ is merely a nuisance parameter in the model.

Our analysis of the BBS data also suggests that the occurrence and detection probabilities of individual species were positively correlated ($\hat{\rho}_{uv} = 0.74$). As noted in Section 3.2, this result can be deduced from relatively simple relationships between the probabilities of occurrence and detection and the abundance of individuals at each sample location.

5.2 Estimates of ecologically important quantities

In addition to species richness of the entire community, we estimated the total number of species present among all 50 sample locations (i.e., M). For the purposes of comparison, we plotted the posteriors of N and M together in Figure 3. The posterior mean of M is 82.3 and there is considerably less uncertainty in M (95% credible interval = 79.3–86.1) than in N. Based on the posterior mean of M, about 7 species of birds were actually present but undetected among the 50 sample locations.

We also estimated the number of species present at each of the 50 sites (Figure 4). The discrepancy between the number of occurring species and the number of species actually observed is quite high at some sites, thus illustrating the effect of accounting for imperfect detection. Comparable results were obtained with estimates of similarity in species composition. For example, we estimated the similarity in species between site 10 and all other sites in the sample (Figure 5). Failure to account for imperfect detection (by substituting the observed numbers of species in (13) instead of the occurring numbers of species) produced biased estimates of similarity at most sites.

6 Discussion

In this case study we have developed and illustrated a model that uses repeated observations of a biological community to estimate the number and composition of species in the community. Estimators of community-level attributes are constructed from model-based estimators of occurrence of individual species that incorporate imperfect detection of individuals. To our knowledge, this is the first model in which community-level and species-level attributes are combined in the same framework.

Our decision to develop a modeling framework that includes both community- and species-level attributes stems from our realization of the importance of sampling design and how it relates conceptually to the definition of a community. For example, we advocate designs that select a *sample of locations* that are thought to be representative of the *population of locations* that encompass the spatial extent of the community. By collecting repeated observations at these sample locations, we have shown that models may be fitted to estimate the number of species in the entire community N, which includes species that are absent (not simply undetected) in the sample of locations. Without repeated observations at each sample location, N can be estimated only if every species is present at every sample location (i.e., if probabilities of species occurrence equal 1 at all sites). In this situation, differences in observed frequencies of detection among species are simply a consequence of their differences in detectability. Every model that specifies only heterogeneity in detection among species (Agresti, 1994; Norris and Pollock, 1996; Boulinier et al., 1998; Coull and Agresti, 1999; Fienberg et al., 1999; Pledger, 2000; Basu and Ebrahimi, 2001; Tardella, 2002; Dorazio and Royle, 2003) implicitly assumes that every species is present at every sample location. Unfortunately, this assumption is unverifiable in the absence of repeated sampling, overly restrictive, and unlikely to be satisfied (even approximately) in natural communities of plants or animals. For example, estimates of the probability of occurrence of many bird species at BBS route 017 are substantially less than 1 (Figure 2).

We believe that it is more reasonable to assume that an individual species may be present at some locations and absent from others. Indeed, this is a natural and widespread view in studies of occurrence of individual species (Bayley and Peterson, 2001; MacKenzie et al., 2002), species-area relationships (Preston, 1960, 1962a,b; MacArthur and Wilson, 1967; Connor and McCoy, 1979), and "patch occupancy" of metapopulations (Hanski, 1992, 1994; Hanski and Gilpin, 1997). We do not mean to imply, however, that our model-based estimates of N are guaranteed to be accurate simply because they allow species occurrence to vary among sample locations. If a community contains species that cannot be detected as a consequence of inadequacies in sampling design or collection methods (e.g., nocturnal animals that cannot be observed in daytime surveys), then our estimates of N will fail to include these species. Thus, sampling deficiencies can have a direct effect on the interpretation of the parameter N, regardless of the model that is used to compute inferences for N. We believe that sampling design and detection methods must be carefully considered in advance of the survey so that every species which is present at a sample location is ensured to have some nonzero probability of being detected.

Our model-based estimates of N are also influenced by the form of the distribution used to specify heterogeneity in species occurrence and detection probabilities. We selected the bivariate normal density $g(u_i, v_i \mid \Sigma)$, but other density functions may also be used to specify variation in u_i and v_i among species. Such alternatives could no doubt have an effect on individual estimates of N because in estimating N we are necessarily extrapolating the number of unobserved species based on patterns of detection and occurrence inferred from the observed species. It is well known that such extrapolations can be sensitive to model structure (Link, 2003).

In developing a model of heterogeneity in species occurrence and detection, we have come to appreciate the shortcomings of trying to estimate species richness with models that assume heterogeneity in detection but not occurrence. In fitting these models to the detections of species that do not occur at all sample locations, we estimate *only* the number of species that occur among the J locations in the sample, which we denote by M (see Section 3.4). In other words, estimates of species richness based on these models will depend on the sample size J, and thus the area sampled, contrary to the commonly held belief that "If detection probability is estimated and explicitly incorporated when estimating species richness, the estimates of richness obtained are independent of the total effort devoted to an area of a given size" (Cam, Nichols, Hines, Sauer, Alpizar-Jara, and Flather, 2002, p. 1121). However, this is not to say that these models respect an underlying species-area relationship induced by the heterogeneity in occurrence among species. They do not. In these models the probability of occurrence is assumed to be one for all species; therefore, the species-area effect can only be accommodated by modifying the data set (that is, by explicitly increasing the number of sites and hence adding more species). Specifically, the models may be fit successively to J, $J+1$, $J+2$, $\ldots$ sites to obtain estimates of species richness that correspond to the sequence of incremental data sets. Under this ad hoc approach, the sampling properties of the estimators are impossible to characterize and the estimators depend on which sites are considered and the order in which they are combined. In contrast, the species-area effect is implied by our model because probabilities of occurrence are not assumed to be one for all species. For any specific data set, one may obtain estimates of the number of species occurring on any number of sites, sampled or not.

Prior to considering the inferential consequences of heterogeneity in occurrence, we (unpublished) developed a model for repeated, site-specific observations of a biological community using the conventional assumption of heterogeneity in detection, but not occurrence, among species. This model is a straightforward extension of the logistic-normal mixture (Coull and Agresti, 1999; Fienberg et al., 1999). We fitted this model to the BBS data in Figure 1 and estimated a species richness of 80.4 (95% profile-likelihood-based confidence interval = 76–91). On fitting our model of heterogeneity in species occurrence, our estimate of M is similar (posterior mean = 82.3) but more precise (95% credible interval = 79.3–86.1). The improved precision may be a consequence of properly partitioning the heterogeneity into its occurrence probability and detection probability components. To explore this possibility, we compared the two estimators of species richness using a small simulation experiment. In this experiment 200 data sets were randomly generated by assuming heterogeneity in species occurrence and detection. The MLEs reported in Table 1 were used as data-generating parameters, with the exception of N, which was assumed to equal 90 species. To save computing time, we also assumed that each of $J = 20$ sites was visited $K = 10$ times. In this experiment the logistic-normal model provided reasonably accurate estimates of M, the number of species present among the J sample locations (estimated bias = 0.9); however, the mean-squared error of this model's estimates was about 40% higher than that computed using the model of heterogeneity in species occurrence and detection (i.e., using (12)). We cannot conclude that the results of this simulation experiment hold generally; however, we do expect improvements in performance of the model of occurrence and detection as the number of sample locations and the number of repeated visits to those locations increase. Of course, one unique feature of this model is that it provides an estimate of the true size of the community N in samples where the number of locations J is limited and $M < N$. This feature is illustrated in Figure 6, which compares the simulated sampling distribution of the model's estimates of N with the simulated sampling distribution of species richness estimates computed using the logistic-normal model of heterogeneity in detection. The median estimate of N computing using the model of occurrence is 91.9 species, which agrees with the true value (90 species) used to simulate the data. In contrast, the logistic-normal model's estimates of species richness are consistently lower than N.

Our model of species occurrence in biological communities provides a realistic starting point for the construction of more complex models. For example, elaborations of the model to include spatial covariates or spatially-dependent priors could be used to predict the occurrence of individual species at unsampled locations. Thus, maps of the occurrence of an individual species (or group of species) could be produced for the entire range of the community. Our model is developed for communities that are "closed" to changes in species composition that may occur as a consequence of local extinctions or colonizations; however, the model could be extended for "open" communities by adding parameters to specify changes in occurrence of individual species. Although some progress has been made in this area (Nichols et al., 1998a,b), inferences based on existing models apply only to those species present in the sampled locations. Failure of these models to include species that are part of the community but occur in unsampled locations obviously could produce misleading inferences about community dynamics. Additional work is needed to develop models of occurrence that provide improved inferences about temporal and spatial dynamics of biological communities.

References

Agresti, A. (1994), "Simple capture-recapture models permitting unequal catchability and variable sampling effort," *Biometrics*, 50, 494–500.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000), "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, 10, 1281–1311.

Basu, S. and Ebrahimi, N. (2001), "Bayesian capture-recapture methods for error detection and estimation of population size: heterogeneity and dependence," *Biometrika*, 88, 269–279.

Bayley, P. B. and Peterson, J. T. (2001), "An approach to estimate probability of presence and richness of fish species," *Transactions of the American Fisheries Society*, 130, 620–633.

Boulinier, T., Nichols, J. D., Hines, J. E., Sauer, J. R., Flather, C. H., and Pollock, K. H. (2001), "Forest fragmentation and bird community dynamics: inference at regional scales," *Ecology*, 82, 1159–1169.

Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. H. (1998), "Estimating species richness: the importance of heterogeneity in species detectability," *Ecology*, 79, 1018–1028.

Bunge, J. and Fitzpatrick, M. (1993), "Estimating the number of species: a review," *Journal of the American Statistical Association*, 88, 364–373.

Burnham, K. P. and Overton, W. S. (1979), "Robust estimation of population size when capture probabilities vary among animals," *Ecology*, 60, 927–936.

Cam, E., Nichols, J. D., Hines, J. E., Sauer, J. R., Alpizar-Jara, R., and Flather, C. H. (2002), "Disentangling sampling and ecological explanations underlying species-area relationships," *Ecology*, 83, 1118–1130.

Cam, E., Nichols, J. D., Sauer, J. R., Hines, J. E., and Flather, C. H. (2000), "Relative species richness and community completeness: birds and urbanization in the mid-Atlantic states," *Ecological Applications*, 10, 1196–1210.

Carlin, B. P. and Louis, T. A. (2000), *Bayes and Empirical Bayes Methods for Data Analysis, second edition*, Boca Raton, Florida: Chapman and Hall.

Colwell, R. K. and Coddington, J. A. (1994), "Estimating terrestrial biodiversity through extrapolation," *Philosophical Transactions of the Royal Society of London B*, 345, 101–118.

Connor, E. F. and McCoy, E. D. (1979), "The statistics and biology of the species-area relationship," *American Naturalist*, 113, 791–833.

Coull, B. A. and Agresti, A. (1999), "The use of mixed logit models to reflect heterogeneity in capture-recapture studies," *Biometrics*, 55, 294–301.

Dice, L. R. (1945), "Measures of the amount of ecologic association between species," *Ecology*, 26, 297–302.

Doherty, Jr., P. F., Sorci, G., Royle, J. A., Hines, J. E., Nichols, J. D., and Boulinier, T. (2003), "Sexual selection affects local extinction and turnover in bird communities," *Proceedings of the National Academy of Sciences*, 100, 5858–5862.

Dorazio, R. M. and Royle, J. A. (2003), "Mixture models for estimating the size of a closed population when capture rates vary among individuals," *Biometrics*, 59, 351–364.

Fienberg, S. E., Johnson, M. S., and Junker, B. W. (1999), "Classical multilevel and Bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society of London A*, 163, 383–405.

Hanski, I. (1992), "Inferences from ecological incidence functions," *American Naturalist*, 139, 657–662.

— (1994), “A practical model of metapopulation dynamics,” *Journal of Animal Ecology*, 63, 151–162.

Hanski, I. and Gilpin, M. E. (1997), *Metapopulation Biology: Ecology, Genetics, and Evolution*, New York: Academic Press.

James, F. C., McCulloch, C. E., and Wiedenfeld, D. A. (1996), “New approaches to the analysis of population trends in land birds,” *Ecology*, 77, 13–27.

Link, W. A. (2003), “Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities,” *Biometrics*, 59, 1123–1130.

Link, W. A., Barker, R. J., Sauer, J. R., and Droege, S. (1994), “Within-site variability in surveys of wildlife populations,” *Ecology*, 75, 1097–1108.

Link, W. A. and Sauer, J. R. (1997), “Estimation of population trajectories from count data,” *Biometrics*, 53, 488–497.

— (1998), “Estimating population change from count data: application to the North American Breeding Bird Survey,” *Ecological Applications*, 8, 258–268.

Liu, Q. and Pierce, D. A. (1994), “A note on Gauss-Hermite quadrature,” *Biometrika*, 81, 624–629.

MacArthur, R. H. and Wilson, E. O. (1967), *The Theory of Island Biogeography*, Princeton, New Jersey: Princeton University Press.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A. (2002), “Estimating site occupancy rates when detection probabilities are less than one,” *Ecology*, 83, 2248–2255.

Nichols, J. D., Boulinier, T., Hines, J. E., Pollock, K. H., and Sauer, J. R. (1998a), “Estimating rates of local species extinction, colonization, and turnover in animal communities,” *Ecological Applications*, 8, 1213–1225.

— (1998b), “Inference methods for spatial variation in species richness and community composition when not all species are detected,” *Conservation Biology*, 12, 1390–1398.

Nichols, J. D. and Conroy, M. J. (1996), “Estimation of species richness,” in *Measuring and Monitoring Biological Diversity. Standard Methods for Mammals*, eds. Wilson, D. E., Cole, F. R., Nichols, J. D., Rudran, R., and Foster, M., Washington, D.C.: Smithsonian Institution Press, pp. 226–234.

Norris, III, J. L. and Pollock, K. H. (1996), “Nonparametric MLE under two closed capture-recapture models with heterogeneity,” *Biometrics*, 52, 639–649.

Pielou, E. C. (1977), *Mathematical Ecology*, New York: John Wiley.

Pinheiro, J. C. and Bates, D. M. (1995), “Approximations to the log-likelihood function in the nonlinear mixed-effects model,” *Journal of Computational and Graphical Statistics*, 4, 12–35.

— (1996), “Unconstrained parametrizations for variance-covariance matrices,” *Statistics and Computing*, 6, 289–296.

Pledger, S. (2000), “Unified maximum likelihood estimates for closed capture-recapture models using mixtures,” *Biometrics*, 56, 434–442.

Preston, F. W. (1960), “Time and space and the variation of species,” *Ecology*, 41, 611–627.

— (1962a), “The canonical distribution of commonness and rarity: Part I,” *Ecology*, 43, 185–215.

— (1962b), “The canonical distribution of commonness and rarity: Part II,” *Ecology*, 43, 410–432.

Robbins, C. S., Bystrak, D., and Geissler, P. H. (1986), “The breeding bird survey: its first fifteen years, 1965–1979,” Resource Publication 157, United States Fish and Wildlife Service, Washington, D.C.

Robbins, C. S., Sauer, J. R., Greenberg, R. S., and Droege, S. (1989), "Population declines in North American birds that migrate to the neotropics," *Proceedings of the National Academy of Sciences (USA)*, 86, 7658–7662.

Robert, C. P. and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer-Verlag.

Sanathanan, L. (1972), "Estimating the size of a multinomial population," *Annals of Mathematical Statistics*, 43, 142–152.

Sauer, J. R., Pendleton, G. W., and Peterjohn, B. G. (1996), "Evaluating causes of population change in North American insectivorous songbirds," *Conservation Biology*, 10, 465–478.

Tardella, L. (2002), "A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity," *Biometrika*, 89, 807–817.

Table 1: Estimates of model parameters and N. Uncertainty in maximum likelihood estimates (MLEs) and Bayesian estimates (posterior means) are indicated by the standard errors in parentheses.

		Posterior	
Parameter	MLE	Median	Mean
α	$-1.51_{(0.47)}$	-1.41	$-1.49_{(0.47)}$
β	$-1.81_{(0.21)}$	-1.84	$-1.85_{(0.22)}$
γ_1	$0.79_{(0.16)}$	0.80	$0.81_{(0.15)}$
γ_2	$0.79_{(0.18)}$	0.84	$0.85_{(0.19)}$
γ_3	$-0.36_{(0.14)}$	-0.27	$-0.27_{(0.15)}$
N	91.4	91.5	$93.3_{(9.0)}$

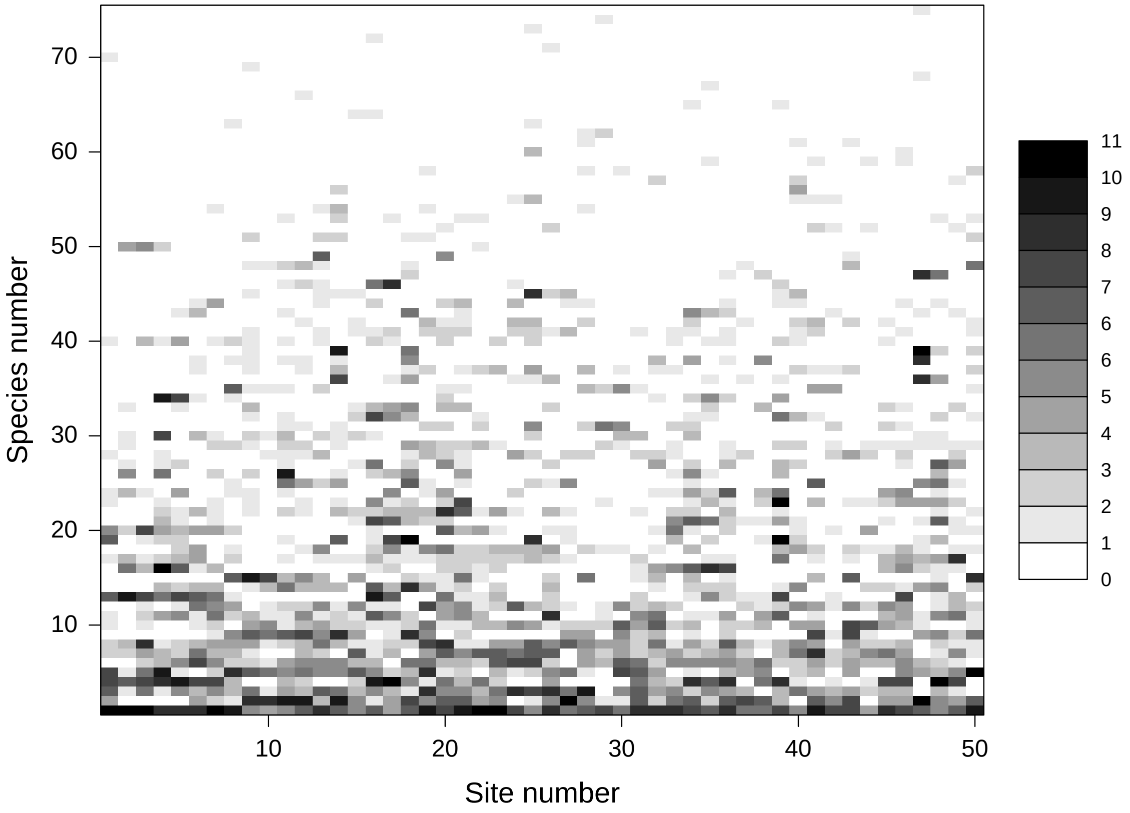

Figure 1: Number of times that each species was detected in 11 visits to each site of BBS route 017 (Maytown, Alabama). Species number increases as the frequency of detection decreases.

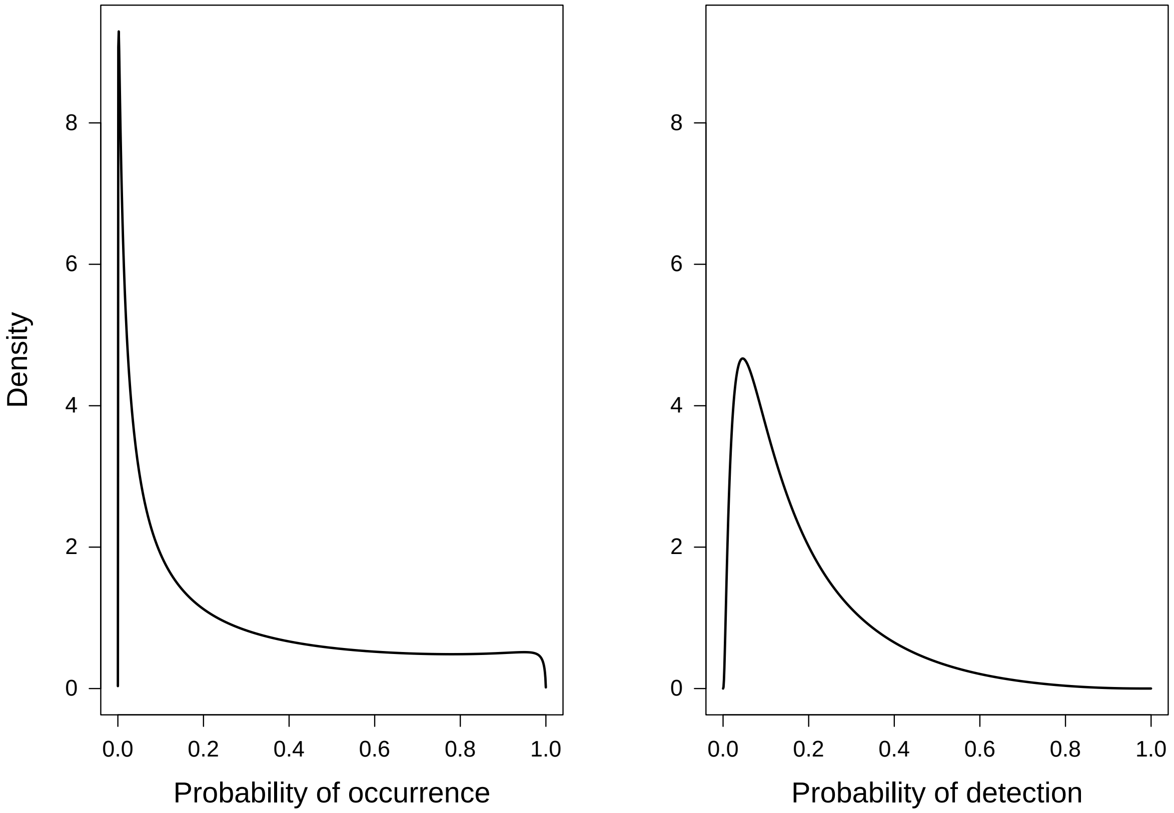

Figure 2: Distributions of probabilities of occurrence and detection based on estimates of model parameters.

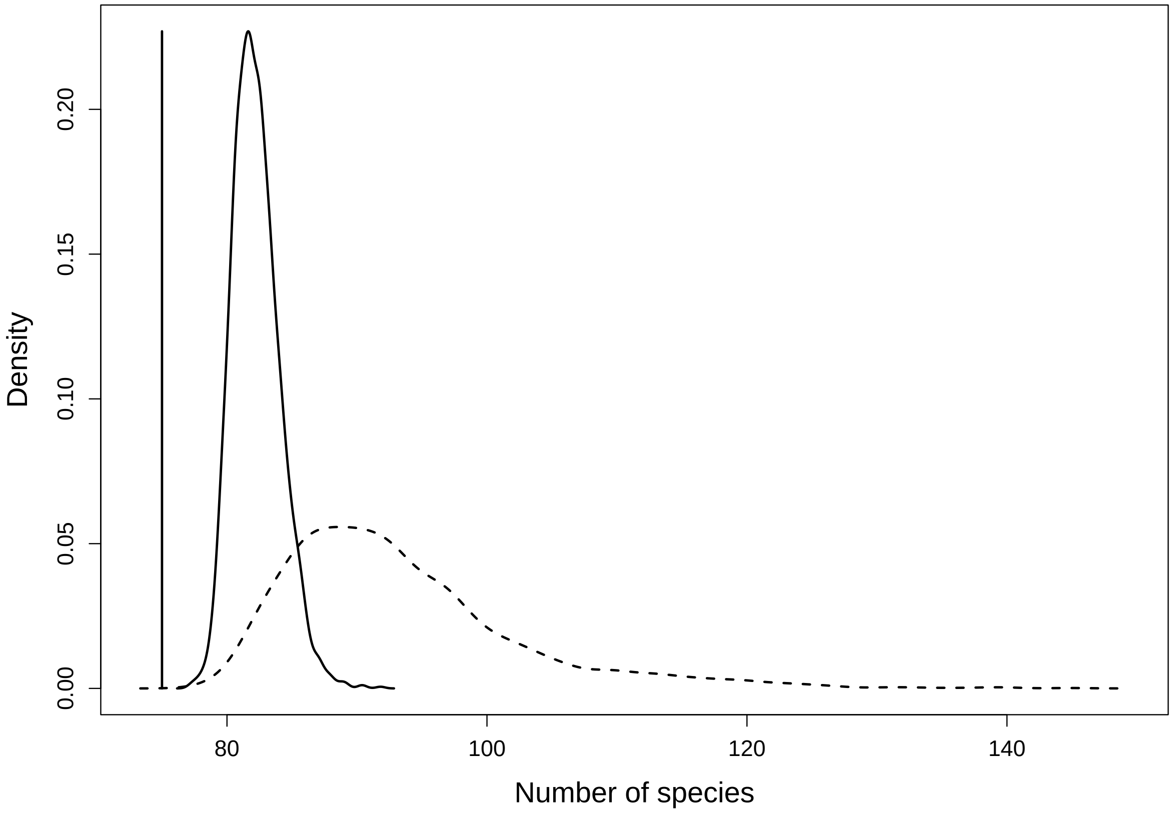

Figure 3: Posterior distributions of the total number of species in the community (N, dashed line), and of the total number of species present among the 50 sample locations (M, solid line). Vertical line indicates the number of species observed in the sample.

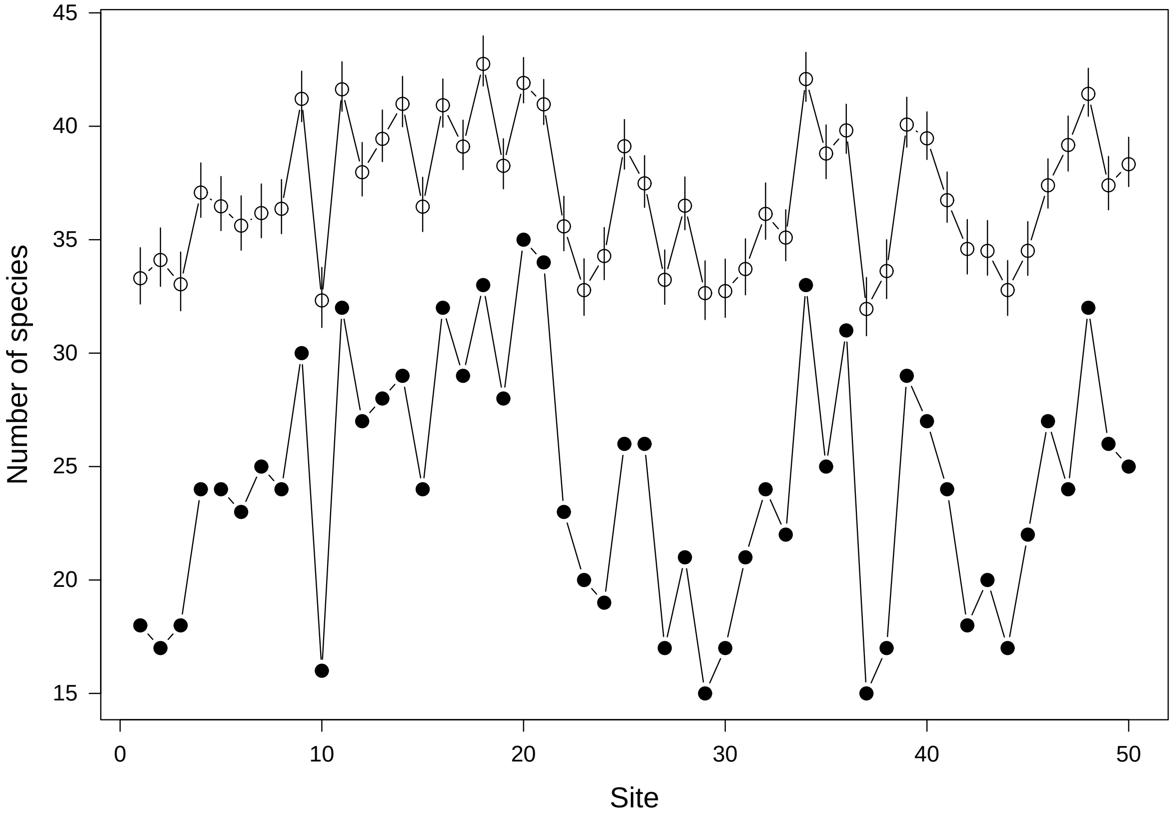

Figure 4: Estimates of the number of species present at each of the 50 sites and 95% credible intervals (open circles with bars). Numbers of species actually observed at each site are shown for comparison (closed circles).

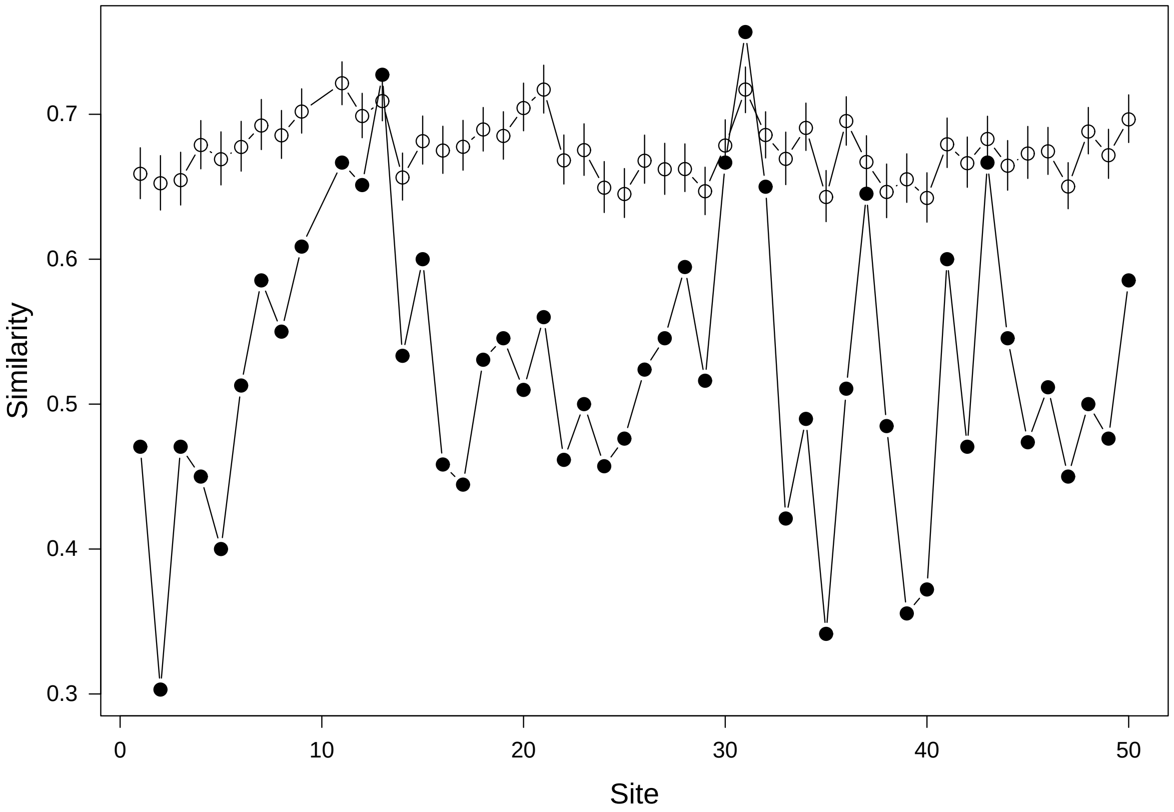

Figure 5: Estimates of similarity in species composition between site 10 and all other sites and 95% credible intervals (open circles with bars). Similarity indices computed using numbers of species actually observed at sites are shown for comparison (closed circles).

.

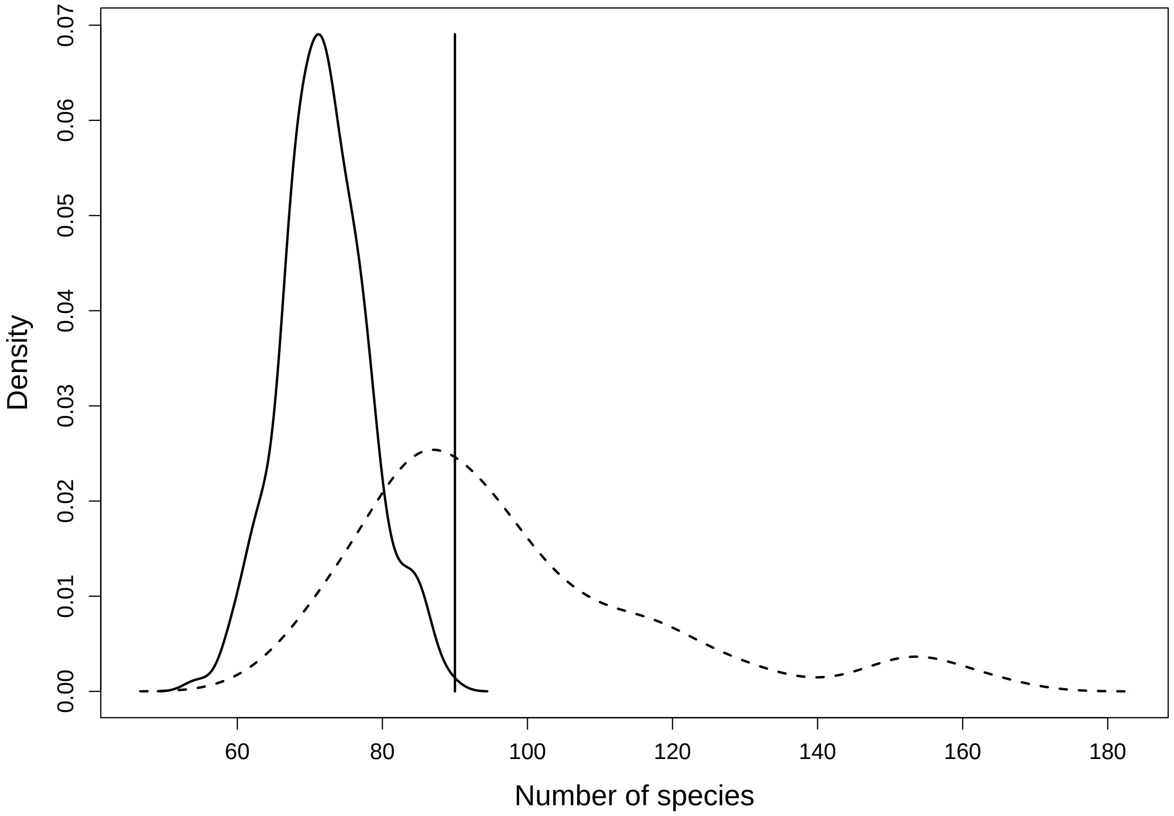

Figure 6: Simulated sampling distributions of 2 estimators of the total number of species N in the community. One distribution is based on our model of heterogeneity in occurrence and detection (dashed line); the other is based on the conventional assumption of heterogeneity in detection only (solid line). Vertical line indicates the true value of N used to simulate the data.