Ecology, 87(4), 2006, pp. 842–854
© 2006 by the Ecological Society of America

ESTIMATING SPECIES RICHNESS AND ACCUMULATION BY MODELING SPECIES OCCURRENCE AND DETECTABILITY

Robert M. Dorazio,[1,4] J. Andrew Royle,[2] Bo Söderström,[3] and Anders Glimskär[3]

[1]*U.S. Geological Survey, Florida Integrated Science Center, Department of Statistics, University of Florida, P.O. Box 110339, Gainesville, Florida 32611-0339 USA*
[2]*U.S. Geological Survey, Patuxent Wildlife Research Center, 12100 Beech Forest Road, Laurel, Maryland 20708 USA*
[3]*Department of Conservation Biology, Swedish University of Agricultural Sciences, P.O. Box 7002, 750 07 Uppsala, Sweden*

Abstract. A statistical model is developed for estimating species richness and accumulation by formulating these community-level attributes as functions of model-based estimators of species occurrence while accounting for imperfect detection of individual species. The model requires a sampling protocol wherein repeated observations are made at a collection of sample locations selected to be representative of the community. This temporal replication provides the data needed to resolve the ambiguity between species absence and nondetection when species are unobserved at sample locations. Estimates of species richness and accumulation are computed for two communities, an avian community and a butterfly community. Our model-based estimates suggest that detection failures in many bird species were attributed to low rates of occurrence, as opposed to simply low rates of detection. We estimate that the avian community contains a substantial number of uncommon species and that species richness greatly exceeds the number of species actually observed in the sample. In fact, predictions of species accumulation suggest that even doubling the number of sample locations would not have revealed all of the species in the community. In contrast, our analysis of the butterfly community suggests that many species are relatively common and that the estimated richness of species in the community is nearly equal to the number of species actually detected in the sample. Our predictions of species accumulation suggest that the number of sample locations actually used in the butterfly survey could have been cut in half and the asymptotic richness of species still would have been attained. Our approach of developing occurrence-based summaries of communities while allowing for imperfect detection of species is broadly applicable and should prove useful in the design and analysis of surveys of biodiversity.

Key words: biodiversity; conservation; detection heterogeneity; occurrence heterogeneity; site-occupancy models.

Introduction

Species richness provides a fundamental measure of community status in quantitative assessments of biological diversity. Species richness is used in the development of ecological theory (MacArthur and Wilson 1967, Hubbell 2001) and in applied problems focused on the conservation of biodiversity. Importantly, the richness of species in a community usually cannot be observed directly because a complete enumeration of every species is seldom feasible. In practice, a sample of the community is collected or observed, and species richness is estimated from the data obtained in the sample. Therefore, the accuracy of an estimate of species richness depends on both the method of data collection (which includes the sampling design) and the statistical model used to analyze the data.

Although natural communities are composed of individual plants or animals, a representative sample of the community is difficult to obtain using individual-based encounters (see Gotelli and Colwell [2001] for some exceptions). In many surveys, a community is divided into spatial sample units (e.g., quadrats or plots) that collectively include every species in the community. Careful attention to spatial scale (i.e., the grain and extent of sampling) is obviously required to define the community and its species unambiguously (Peterson and Parker 1998). Historically, ecologists were advised to study communities in regions considered to be "relatively homogeneous" with respect to habitat (MacArthur 1972); however, such regions are almost never found in practice because some degree of habitat heterogeneity exists at every spatial scale of practical importance. Furthermore, important connections may exist among the spatial variation in habitat, the community's structure, and the population dynamics of individual species (Hanski and Gilpin 1997, Wiens 2002). These connections are increasingly appreciated in modern ecology and should be considered in designing surveys of natural communities.

Once a region of sampling is selected in such surveys, the finite number N of distinct species in the community is determined unambiguously because the spatial extent

Manuscript received 23 May 2005; revised 30 September 2005; accepted 11 October 2005. Corresponding Editor: N. G. Yoccoz.
[4] E-mail: bdorazio@usgs.gov

of the community is bounded. Though N is sometimes used to denote "local" species richness (e.g., Colwell and Coddington 1994), this distinction is really unnecessary because the region selected for sampling provides the spatial context needed to define and interpret N. But how should N be estimated once the community has been sampled?

Various statistical approaches, many developed for use in nonecological problems, have been used to estimate species richness. These approaches may be classified into four categories: (1) extrapolation of species-accumulation curves (Gotelli and Colwell 2001, Ugland et al. 2003), (2) parametric models of the apparent species-abundance distribution (Pielou 1977:chapter 18), (3) nonparametric models based on sampling theory (Bunge and Fitzpatrick 1993), and (4) community analogs of capture–recapture models of demographically closed populations (Nichols and Conroy 1996, Boulinier et al. 1998, Dorazio and Royle 2003). A review or comparison of all of these approaches is beyond the scope of our paper.

Our view of the estimation problem is that species richness, species accumulation, and other attributes of community structure are most naturally formulated using models of individual species occurrence that explicitly account for the imperfect detection of a species during sample collection (Dorazio and Royle 2005). We combine community-level and species-level attributes in the same modeling framework, allowing either community-level or species-level estimands to be evaluated as needed in specific problems. This versatility is not shared by any of the existing methods of estimating species richness. For example, estimating N from the asymptote of an empirical species-accumulation curve, one of the most popular approaches, is intuitively attractive because it honors the axiomatic increase in species number with increases in area sampled (Connor and McCoy 1979, Coleman 1981, Coleman et al. 1982, McGuinness 1984, Rosenzweig 1985, Lomolino 2000); however, this approach may fail in communities that contain many rare species (Fisher 1999, Novotny and Bassett 2000) or in communities of species that are common but difficult to detect. Some statistical approaches attempt to overcome these difficulties by explicitly assuming that all species are not equally well detected (Burnham and Overton 1979, Nichols and Conroy 1996, Norris and Pollock 1996, Coull and Agresti 1999, Pledger 2000, Basu and Ebrahimi 2001, Tardella 2002, Dorazio and Royle 2003, Mao and Colwell 2005, Mao et al. 2005). Many of these approaches were developed for estimating the size of a population from the repeated observations of marked individuals in capture–recapture surveys. In the context of estimating species richness, the detections of species encountered at different sample locations are analogs of the recaptures of marked individuals at different sample times. Therefore, for each species observed in the survey, 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). The sample matrix of these vectors (one for each observed species) is sometimes called an "incidence matrix" (Colwell et al. 2004); however, it is clear that a zero may indicate that a species is absent at that location or that the species is present but undetected at that location. In other words, there is an inherent ambiguity between detection and occurrence that cannot be resolved by viewing sample locations as replicate observations. Existing models of such incidence matrices either assume that every species is present at every sample location, though possibly undetected, or they admit an explicit confounding between species presence and detectability. In either case, these models cannot be used to construct occurrence-based summaries of community-level attributes (Dorazio and Royle 2005).

We believe it is necessary to consider sampling designs and models that allow the occurrence of a species to vary with location while accounting for imperfect detectability; therefore, additional sampling is required to resolve the ambiguity between species occurrence and detection. In particular, each sample location must be sampled repeatedly, and the total duration of the survey must be kept sufficiently short that local extinctions or colonizations of species are unlikely. The latter constraint is needed to ensure that site-specific species occurrence and N remain constant during the survey.

In this paper, we describe a statistical model for estimating the richness and accumulation of species based on elemental models of species occurrence and detection. Our previous efforts (Dorazio and Royle 2005) were limited to the estimation of species richness and similarity. Here, we develop a reparameterized model that allows species-accumulation curves to be estimated from estimates of species richness and species occurrence. Such curves may be used to compare the richness of different communities at comparable levels of sampling effort (Colwell and Coddington 1994, Gotelli and Colwell 2001), to improve the efficiency of future community surveys (Soberón and Llorente 1993, Colwell and Coddington 1994), or to select priority areas for conservation in the design of natural species reserves (Margules et al. 2002, Cabeza et al. 2004). We illustrate our method by estimating species accumulation curves of avian and butterfly communities.

Protocol for Sampling Communities

We describe here a protocol for sampling communities based on presence–absence data (more correctly, detection–nondetection data) that can be applied to many taxa and in many settings. In this sense, the protocol is quite general and can be applied in many surveys of biodiversity. We assume that the community to be surveyed is divided into spatial sample units (i.e., distinct locations) and that a proper consideration of spatial scale is exercised in defining the community (as noted in *Introduction*). Furthermore, we assume that

appropriate procedures (e.g., randomization, stratification, or clustering) have been used to select a sample of units that is representative of the community. In this way, we ensure that inferences derived from the sample are in fact relevant to the community.

Once the sample of locations has been selected, each location must be visited repeatedly (at least twice), recording the list of species actually detected. Ideally, species should be detected at each location using the same level of effort on all sampling occasions. For example, the time spent sampling, the method(s) of detection, and even the identities of observers should be standardized to the extent that this is possible. While standardization helps to reduce variation among sampling occasions in the probability of detecting individuals of each species, our model of species detection can be extended to accommodate situations that prevent the use of standardized sampling protocols. A final requirement of our sampling protocol is that the entire survey must be completed within a sufficiently short time that local extinctions or colonizations cannot change the composition of species that occupy a sample location. In other words, each species is assumed to be present or absent at each sample location and this state is assumed to remain fixed during the survey. In more conventional models the assumption of closure to changes in species composition is usually made for the entire community of N species. In our model, closure is assumed for each sample location. Kendall (1999) provides a good discussion of the consequences of violating the closure assumption.

The purpose of temporal replication at each sample location is to provide the information needed to estimate the probability of detecting each species (given that it is present) separately from its probability of occurrence. By using temporal replication at each site, we essentially remove the ambiguity of observed zeroes that occur at locations where a species is not detected. (Recall that such nondetections can arise because individuals of the species are truly absent at a sample location or because these individuals are present but undetected.) The ambiguity can be resolved by properly modeling the temporally and spatially replicated species detections. Although a minimum of two visits is needed at each sample location, a higher number of temporal replicates (subject to the restriction on total survey duration) is obviously desirable, particularly in communities dominated by species that are difficult to detect. In the next sections, we briefly describe two communities that were sampled using the protocol that we advocate.

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 located in North America (Robbins et al. 1986, 1989, Sauer et al. 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-min period. 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. The data used in our analysis were collected at one of these routes located in Maytown, Alabama, USA (BBS route number 017). This route was visited by the same observer on 11 different days in the month of June.

Survey of butterflies in Swedish grasslands

In July 1997, a survey was completed to estimate the number of butterfly species present within a region of grazed seminatural grasslands located in south-central Sweden (Söderström et al. 2001, Vessby et al. 2002). These grasslands occur as small (mean size 6 ha) habitat fragments and are a common part of the agricultural landscape in Sweden. Twenty grasslands thought to be representative of the region were selected for sampling. A square route (100×100 m) was located within each grassland and used to delineate a 400-m transect for sampling. Surveyors walked along each route at constant speed (20 m/min) and counted all butterfly species detected within 1 m of the transect. Each of the 20 routes was visited on 18 different days in the month of July.

Model Description

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-accumulation curve 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 *Model*

TABLE 1. An $N \times J$ matrix of species- and site-specific detections, $\mathbf{X}$, and a partially observed matrix $\mathbf{Z}$ (also $N \times J$), whose elements indicate species- and site-specific occurrence.

	Site j							
	Observed				Partially observed			
Species i	1	2	$\cdots$	J	1	2	$\cdots$	J
1	x_{11}	x_{12}	$\cdots$	x_{1J}	z_{11}	z_{12}	$\cdots$	z_{1J}
2	x_{21}	x_{22}	$\cdots$	x_{2J}	z_{21}	z_{22}	$\cdots$	z_{2J}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{nJ}	z_{n1}	z_{n2}	$\cdots$	z_{nJ}
$n+1$	0	0	$\cdots$	0	$z_{n+1,1}$	$z_{n+1,2}$	$\cdots$	$z_{n+1,J}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
N	0	0	$\cdots$	0	z_{N1}	z_{N2}	$\cdots$	z_{NJ}

description: Modeling heterogeneity in occurrence and detection of species).

Let $x_i = (x_{i1}, x_{i2}, \ldots, x_{iJ})$ 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: $\mathbf{x}_i \in \{(K+1)^J - 1 \text{ observable vectors}\}$ for $i = 1, \ldots, n$; $\mathbf{x}_i = \mathbf{0}$ for $i = n+1, \ldots, N$. This ordering implies a partitioning of the $N \times J$ matrix of observation vectors ($\mathbf{X}$) into an observed portion $\mathbf{X}_n$ (the first n rows of $\mathbf{X}$) and an unobserved portion, which includes species that are undetected in the survey (Table 1). Based on this partitioning, it is also useful to consider an $N \times J$ matrix of binary indicators $\mathbf{Z}$, whose elements denote the presence ($z_{ij} = 1$) or absence ($z_{ij} = 0$) of species i at site j. Note that $\mathbf{Z}$ 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 *Model description: Predicting species accumulation as a function of species occurrence*, we show that allowing the occurrence of species to vary spatially (i.e., among sample sites) through the definition of $\mathbf{Z}$ simplifies our calculation of species accumulation curves, as well as other estimands of ecological interest (e.g., similarity in species composition).

Modeling 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 (Coull and Agresti 1999, Fienberg et al. 1999); however, the logistic-normal model conditions only on the site-specific detection probability of each species.

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 probability mass function

$$p(z_{ij}|\psi_{ij}) = \psi_{ij}^{z_{ij}}(1-\psi_{ij})^{1-z_{ij}}. \quad (1)$$

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

$$p(x_{ij}|z_{ij}, \theta_{ij}) = \left[\binom{K}{x_{ij}}\theta_{ij}^{x_{ij}}(1-\theta_{ij})^{K-x_{ij}}\right]^{z_{ij}}. \quad (2)$$

In contrast, if species i is absent at site j ($z_{ij} = 0$), then x_{ij} is assumed to equal zero with probability one.

Dorazio and Royle (2005) showed that removal of z_{ij} (by summation) conveniently provides the marginal probability of the observed number of detections:

$$p(x_{ij}|\theta_{ij}, \psi_{ij}) = \psi_{ij}\binom{K}{x_{ij}}\theta_{ij}^{x_{ij}}(1-\theta_{ij})^{K-x_{ij}} + (1-\psi_{ij})I(x_{ij}=0) \quad (3)$$

where $I(\cdot)$ denotes an indicator function, which equals one when its argument is true and is zero otherwise. Note that Eq. 3 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 $\psi_{ij}(1 - \theta_{ik})^K$). Eq. 3 has been used in models of site occupancy (MacKenzie et al. 2002) for individual species; therefore, in that sense, our model specifies a multispecies, site-occupancy model.

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: logit $\psi_{ij} = u_i + \alpha_j$ and logit $\theta_{ij} = v_i + \beta_j$, 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}. In the absence of site-level covariates (as in our avian and butterfly surveys), 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 $[u_i, v_i \,|\, \boldsymbol{\Sigma}] \sim \text{normal}(\mathbf{0}, \boldsymbol{\Sigma})$, 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 the 2 $\times$ 2 matrix $\mathbf{\Sigma}$).

Estimating model parameters and species richness

Let $f(x_{ij} \mid u_i, v_i, \alpha, \beta)$ denote the conditional probability of the observed number of detections of species i at site j given the logit-scale parameters for ψ_{ij} and θ_{ij}. This is obtained by substituting into Eq. 3 the logit-scale parameters for ψ_{ij} and θ_{ij} (cf. Eq. 5 in Dorazio and Royle [2005]). Adopting a conventional treatment of random effects, a likelihood function for the fixed parameters (α, β, $\mathbf{\Sigma}$) may be derived by integrating a conditional likelihood over the distribution of random effects parameters. For example, if we assume that the J observations of each species are conditionally independent, the marginal probability of the observation vector $\mathbf{x}_i$ is

$$q(\mathbf{x}_i|\alpha, \beta, \mathbf{\Sigma}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\prod_{j=1}^{J} f(x_{ij}|u_i, v_i, \alpha, \beta)\right] \times g(u_i, v_i|\mathbf{\Sigma})\, du_i\, dv_i \quad (4)$$

where $g(u_i, v_i \mid \mathbf{\Sigma})$ specifies the bivariate normal density assumed for u_i and v_i. Although the integrations in Eq. 4 can be approximated using numerical quadrature (Liu and Pierce 1994, Pinheiro and Bates 1995) or stochastic methods (e.g., Monte Carlo), these techniques can be computationally intensive to implement. Furthermore, and more importantly, estimates of the u_i parameters and their uncertainties are needed to calculate species occurrence and other ecologically relevant quantities. Therefore, Dorazio and Royle (2005) adopted a Bayesian framework for parameter estimation and inference, making only minimal use of the marginalization in Eq. 4. They showed that the multinomial likelihood based on the detections of the n observed species and their logit-scale contributions to detection and occurrence is

$$p(\mathbf{X}_n, \mathbf{n}, \mathbf{u}, \mathbf{v}|\alpha, \beta, \mathbf{\Sigma}) = \frac{n!}{\prod_h n_h!}\left[\frac{1}{1 - q(\mathbf{0}|\alpha, \beta, \mathbf{\Sigma})}\right]^n \times \prod_h \left\{\left[\prod_{j=1}^{J} f(x_{hj}|u_h, v_h, \alpha, \beta)\right] g(u_h, v_v|\mathbf{\Sigma})\right\}^{n_h} \quad (5)$$

where $n_h = \Sigma_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}_h)$ denotes the number of species that share detection sequence $\mathbf{x}_h$ ($h = 1, \ldots, m$) and $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{n} = (n_1, \ldots, n_m)$ each denote a vector of $m \leq n$ elements that correspond to the distinct values of $\mathbf{x}_h$ observed in the sample.

Here, we develop an alternative approach based on a reparameterization of the model that allows it to be fitted by Markov chain Monte Carlo (MCMC) sampling without any numerical integration. This reparameterization is easily specified in the BUGS language (Gilks et al. 1994) and can be fitted to data using WinBUGS software (*available online*).[5] (We provide the necessary WinBUGS code in the Supplement.) To motivate our reparameterization, note that if N was known, specification of the hierarchical model for community structure would be complete without having to consider the "undetected" portion of the community in the likelihood, and there would be no difficulty in fitting the model using MCMC. The difficulty with N being unknown is that the dimension of the parameter vectors $\mathbf{u}$ and $\mathbf{v}$ (and thus, $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$) changes each time another MCMC draw of the parameter N is computed. To obtain a model in which the dimension of the parameters is constant, we create a supercommunity of species, one that comprises the n observed species and an arbitrarily large, but known, number of unobserved species for which $\mathbf{x}_i = \mathbf{0}$ ($i = n + 1, n + 2, \ldots, N, N + 1, \ldots, S$). The supercommunity size S is fixed, and thus the dimension of the parameter vectors is constant (i.e., not a function of N). In taking this approach, we do not directly estimate N as a parameter. Instead, we introduce an additional latent indicator variable, say w_i, which takes the value 1 if a species in the supercommunity is a member of the community available to be sampled and 0 otherwise. We assume that $\{w_i\}$ are independent, Bernoulli-distributed, random variables indexed by parameter Ω. Obviously, w_i is observed for $i = 1, 2, \ldots, n$, but not otherwise. By introducing the supercommunity of latent variables $\{w_i\}$ into the model, we effectively transform the problem of estimating N into the equivalent problem of estimating $\Sigma_{i=1}^{S} w_i$, which, of course, depends on the estimated value of Ω.

Our reparameterized model does require that S be assigned a sufficiently high value; however, in practice it is a simple matter to assess the adequacy of any particular choice of S. Recall that $E(N) = S\Omega$ and that Ω is bounded between 0 and 1; therefore, estimates of Ω will necessarily decline as higher values of S are chosen. If the assigned value of S is too low, the posterior distribution of Ω will be concentrated near the upper limit of its range, and we risk underestimating the true value of N. The obvious solution is to increase S until the posterior of Ω is centered well below its upper limit. However, higher values of S also imply higher computational costs, so some care is advised in assigning too high a value to S.

The concept of a supercommunity of S species may seem artificial, but it also can be motivated quite naturally. Suppose we observe n avian species after sampling an individual BBS route somewhere in the United States. We could apply the conventional approach wherein species richness N is viewed as an unknown, multinomial index to be estimated (Dorazio and Royle 2005); however, it seems entirely reasonable to specify some maximal species list, perhaps composed of all known avian species in North America. Surely the number of bird species living in proximity to an

[5] ⟨http://www.mrc-bsu. cam.ac.uk/bugs/⟩

individual BBS route could not exceed the total number of known avian species in North America!

We complete our reparameterized model by assuming mutually independent prior distributions for Ω, α, β, and $\boldsymbol{\Sigma}$. In particular, we assume a uniform(0,1) prior for Ω, $\text{logit}^{-1}(\alpha)$, and $\text{logit}^{-1}(\beta)$; 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($\varepsilon$, $1/\varepsilon$) 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 2000:149). Similarly, we use the uniform distribution to specify our prior indifference in the mean probabilities of detection and occurrence, in the Ω parameter, and in the correlation parameter ρ_{uv}. This set of priors was used in both of the analyses described in *Analysis of data sets*.

Predicting species accumulation as a function of species occurrence

We are interested in predicting the relationship between the expected number of species that occur in some prescribed region as a function of the area of that region. Given our sampling frame, "area" is equivalent to the aggregate number of discrete spatial units selected from those that define the spatial extent of the community. Unlike empirical species-accumulation curves, our prediction is not confined to the particular set of locations in the sample. If we adopt that convention, the predicted species–area relationship will depend on *which* locations are considered and on the order in which they are aggregated. We view this as an inherent limitation of rarefying (i.e., interpolating) and extrapolating empirical species-accumulation curves (Gotelli and Colwell 2001, Ugland et al. 2003). In addition, we believe that predictions of species accumulation should account for uncertainty in estimates of N and species occurrence; therefore, in this section we develop the posterior-predictive distribution of the species-accumulation curve.

In *Model description: Preliminaries and definitions*, we defined a matrix $\mathbf{Z}$ whose elements indicate the occurrence (i.e., presence/absence) of each of the N species in the community at each of the J sample locations. Similarly, let $\tilde{\mathbf{Z}}$ denote a $N^* \times L$ matrix whose elements indicate whether each species occurs at each of L unsampled locations (spatial units). (We use an asterisk, as in N^*, to denote a random draw from the simulated sample of the joint posterior distribution; therefore, each predicted matrix of species occurrences $\tilde{\mathbf{Z}}$ is associated with a single draw from the simulated sample of the joint posterior.) According to Eq. 1, predicting an element of $\tilde{\mathbf{Z}}$, say $\tilde{z}$, is simply a matter of computing a random draw from a Bernoulli($\tilde{\psi}$) distribution, where $\tilde{\psi}$ denotes the predicted probability of occurrence of species i in unit j. Similarly, predicting the occurrence probability $\tilde{\psi}$ is done by computing a random draw from the normal(α^*, σ_u^{2*}) distribution and transforming the result to a probability (using logit^{-1}). Therefore, predictions of occurrence $\tilde{\mathbf{Z}}$ ultimately depend on estimates of N and the model parameters used to specify heterogeneity in species occurrence probabilities. These predictions can be expressed formally using the posterior-predictive density of a single element of $\tilde{\mathbf{Z}}$:

$$p(\tilde{z}|\mathbf{X}_n, \mathbf{n}) = \int_{\tilde{\psi}\times\Theta} p(\tilde{z}|\,\tilde{\psi}) p(\,\tilde{\psi}|\Theta)\pi(\Theta|\mathbf{X}_n, \mathbf{n})\; \mathrm{d}\tilde{\psi}\; \mathrm{d}\Theta \tag{6}$$

where $\pi(\Theta \,|\, \mathbf{X}_n, \mathbf{n})$ denotes the joint posterior density of model parameters and N. Note that our predictions of species occurrence $\tilde{z}$ are not simply point estimates. By integrating over the posterior uncertainty in N and model parameters (as expressed in the joint density $\pi(\Theta \,|\, \mathbf{X}_n, \mathbf{n})$), we compute a distribution of predictions $\tilde{\mathbf{Z}}$ that conditions only on the observed data ($\mathbf{X}_n$, $\mathbf{n}$). We use the method of composition (Tanner 1996) rather than direct integration to compute the sample of $\tilde{\mathbf{Z}}$ values.

A sample from the posterior-predictive distribution of species-accumulation curves is readily computed once we have computed a sample from the posterior-predictive distribution of species occurrences. Let $\tilde{M}_l$ denote the cumulative number of distinct species that are predicted to be present among the first l spatial units. In other words, $\tilde{M}_l$ denotes an ordinate of the species-accumulation curve with abscissa l ($= 1, \ldots, L$). For a particular draw of $\tilde{\mathbf{Z}}$, we compute$\tilde{M}_l$ by summing only those rows where species are predicted to be present, that is, $\tilde{M}_l = \Sigma_{i=1}^{N^*} I(\tilde{z}_{i\cdot} > 0)$ where $\tilde{z}_{i\cdot} = \Sigma_{j=1}^{l} \tilde{z}_{ij}$. Repeating this calculation for each value of l yields a single draw from the posterior-predictive distribution of species accumulations $[\tilde{M}_1, \ldots, \tilde{M}_L \,|\, \mathbf{X}_n, \mathbf{n}]$. The sample of predicted species-accumulation curves is computed by repeating this sequence of calculations for each posterior-predicted draw of $\tilde{\boldsymbol{Z}}$.

In the Supplement, we provide code for fitting the model and for computing species richness and accumulation using the freely available software packages, R (R Development Core Team 2004) and WinBUGS. Our code uses the R package R2WinBUGS (Sturtz et al. 2005) to execute WinBUGS while running a session in R. We also provide the data observed in our avian and butterfly surveys.

Analysis of Data Sets

Breeding bird survey

Seventy-five species of birds were detected in the survey, and there was considerable variation among species in the observed frequencies of detection at each site (Fig. 1). We computed the posterior distribution of species richness using the species- and site-specific detections of all birds observed along the BBS route.

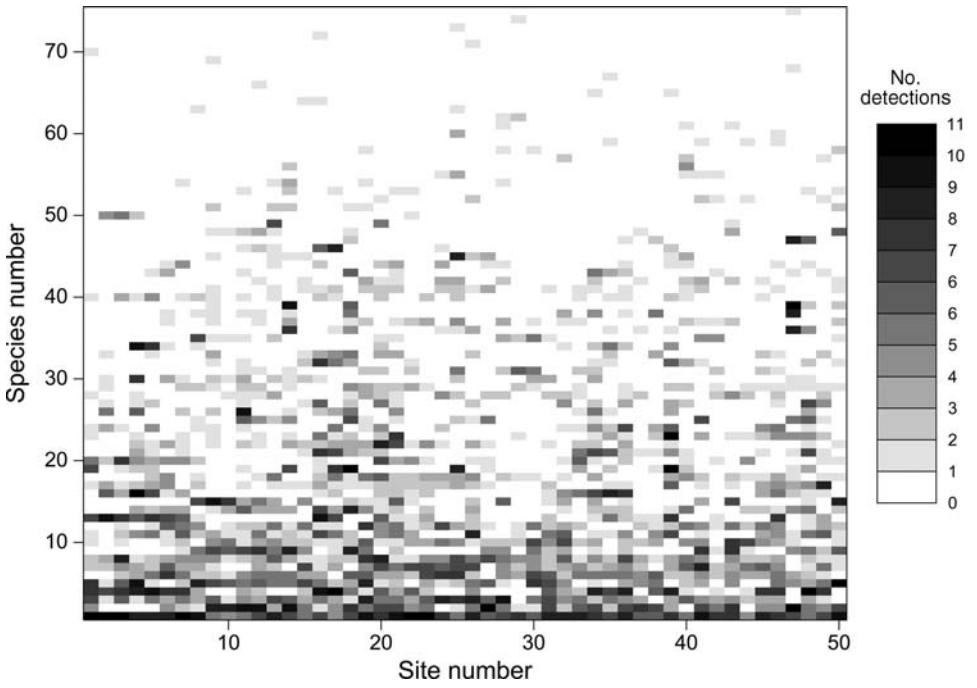

FIG. 1. Number of times that each bird species was detected in 11 visits to each site of BBS route 017 (Maytown, Alabama, USA). Species are numbered from the most detectable to the least detectable.

The estimated size of the community exceeds the number of species observed in the sample by a substantial margin (Fig. 2). In fact, the posterior probability that the avian community comprises only $N = 75$ species is essentially zero, and the estimated median and mean values of species richness are 90.0 and 93.0, respectively.

These results are consistent with estimated levels of heterogeneity in species occurrence and detection. For example, the marginal distributions of species-specific probabilities of occurrence and detection implied by our estimates of the model parameters (Fig. 3) suggest that detection failures in many bird species are attributed to low rates of occurrence, as opposed to simply low rates of detection. In other words, a substantial portion of the community includes relatively uncommon species; therefore, it is not surprising that the estimated total number of species in the community exceeds the number of species observed in the sample.

We also computed the accumulation of species that is predicted in samples of 1–100 sites (Fig. 4). The

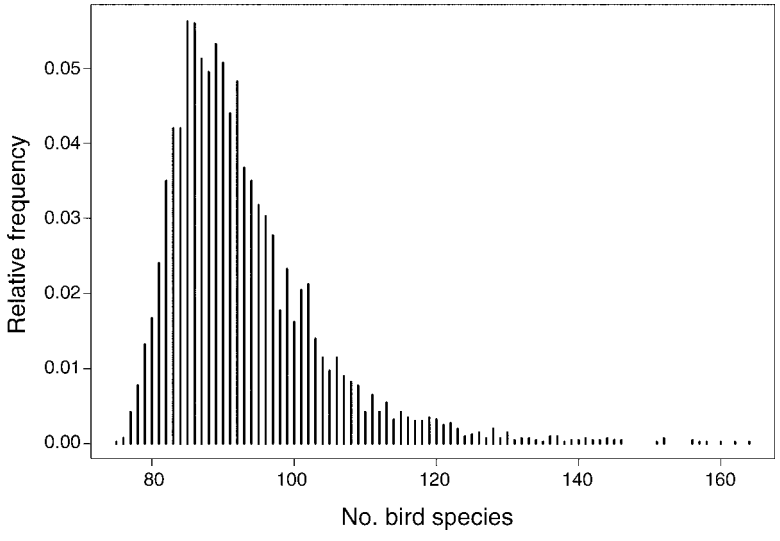

FIG. 2. Posterior distribution of species richness in the community of breeding birds.

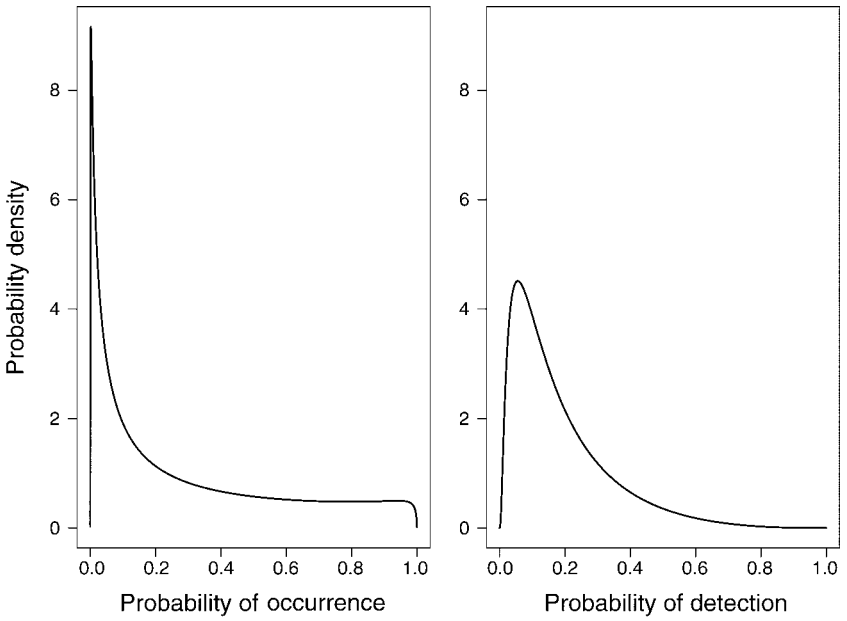

FIG. 3. Distributions of probabilities of occurrence and detection of bird species based on estimates of model parameters ($\hat{\alpha} = -1.50$, $\hat{\sigma}_u = 2.20$, $\hat{\beta} = -1.81$, $\hat{\sigma}_v = 1.07$.

predicted species-accumulation curve fails to reach the asymptotic richness of the community even if the number of sample sites (and corresponding area sampled) is twice as large as that actually used in the BBS. Again, this result stems from the estimated rarity of many bird species.

Butterfly survey

Twenty-eight species of butterflies were detected in the survey, and, as in the BBS, there was considerable variation among species in the observed frequencies of detection at each site (Fig. 5). Unlike the avian community, our posterior estimates of species richness (median = 28.0, mean = 28.5) are close to the observed number of butterfly species. In fact, the posterior probability that $N > 30$ species is only 0.05 (Fig. 6).

These results are consistent with obvious differences in the estimates of occurrence of butterfly and bird species. The marginal distribution of estimated probabilities of occurrence of butterfly species (Fig. 7) suggests that many species are relatively common and that repeated sampling within a site can substantially

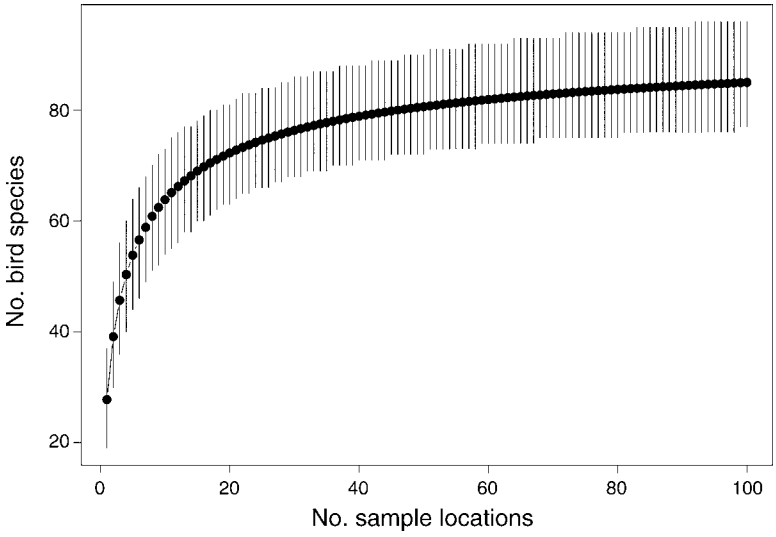

FIG. 4. Predicted species-accumulation curve for the community of breeding birds. Each point along the curve corresponds to an estimate of the mean of the posterior-predictive distribution. Error bars indicate 90% prediction intervals.

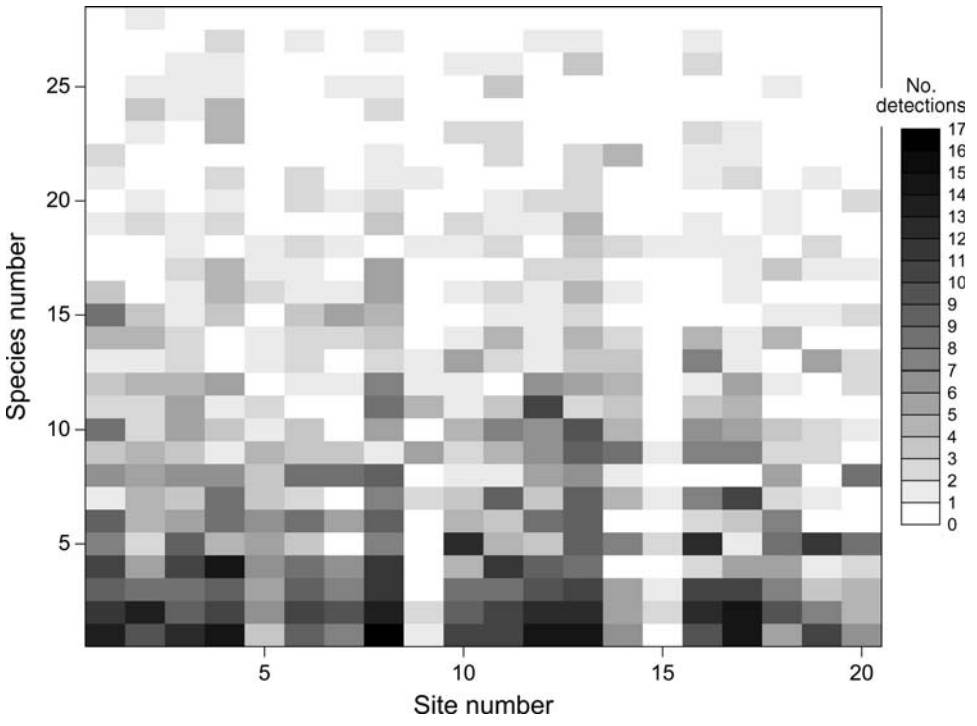

FIG. 5. Number of times that each butterfly species was detected in 18 visits to each route. Species are numbered from the most detectable to the least detectable.

increase the chances of detecting these species, even for those whose probability of detection is not particularly high. Therefore, it is not surprising that the estimated size of the butterfly community is nearly equal in magnitude to the number of species detected in the sample.

We also computed the accumulation of species predicted in samples of 1–30 sites (Fig. 8). The predicted species-accumulation curve reaches the asymptotic richness within 10 sample sites, well below the number of sites used in the survey. Again, this result stems from the relatively high estimates of occurrence of many butterfly species.

DISCUSSION

In this paper, we have described a sampling protocol and statistical model for computing estimates of species richness and species accumulation. A strength of our

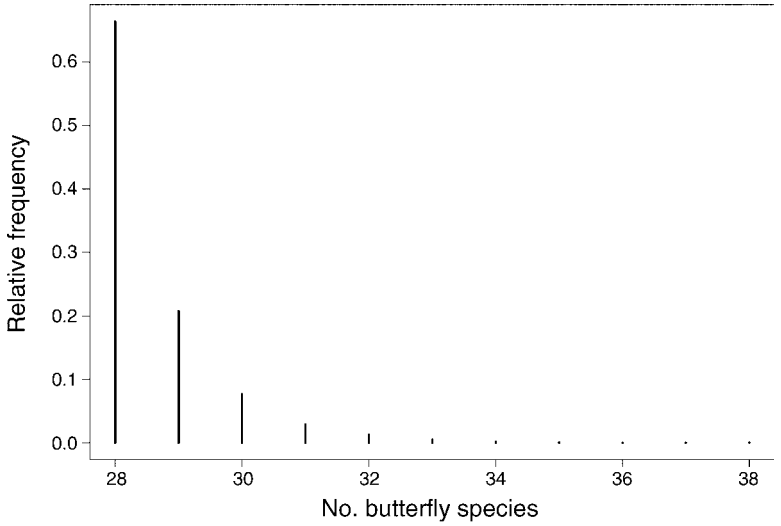

FIG. 6. Posterior distribution of species richness in the community of butterflies.

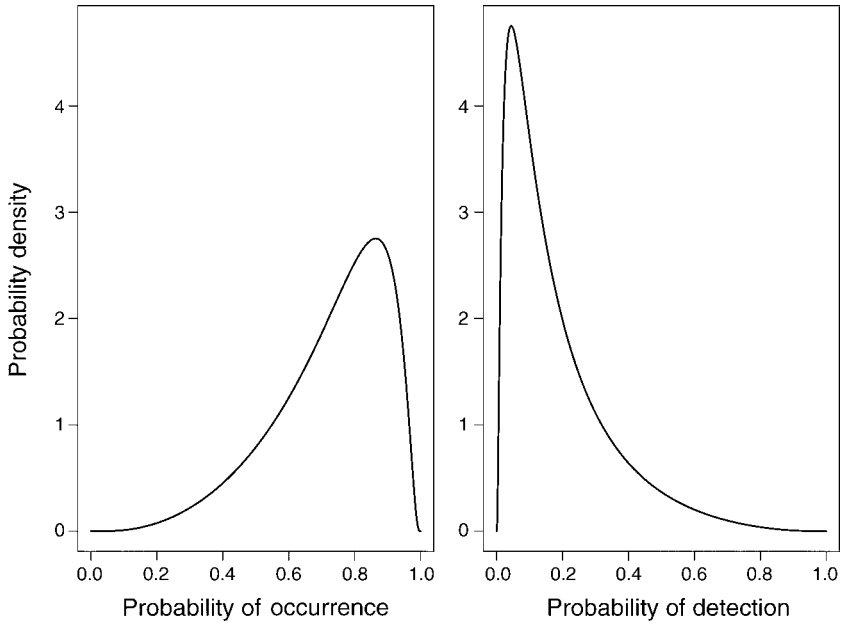

FIG. 7. Distributions of probabilities of occurrence and detection of butterfly species based on estimates of model parameters ($\hat{\alpha} = 1.17$, $\hat{\sigma}_u = 0.96$, $\hat{\beta} = -1.87$, $\hat{\sigma}_v = 1.15$.

approach is that community-level and species-level attributes are combined in the same modeling framework; thus, community-level attributes, such as species richness and accumulation, may be expressed naturally as a function of occurrence of individual species.

The sampling protocol that we advocate requires repeated observations at sample locations that are selected to be representative of those locations that encompass the spatial extent of the community. Repeated sampling (i.e., temporal replication) at each location provides the information needed to determine probabilistically whether a species is absent at each location or present but undetected (MacKenzie et al. 2002). The separation of species occurrence and detectability is what allows us to estimate the number of species N in the entire community, as well as other occurrence-based, community-level estimands. Consequently, our model-based estimators of species richness and accumulation represent improvements over existing methods; however, our estimators are not guaranteed to be accurate simply because they allow species occurrence to vary among sample locations. If a community

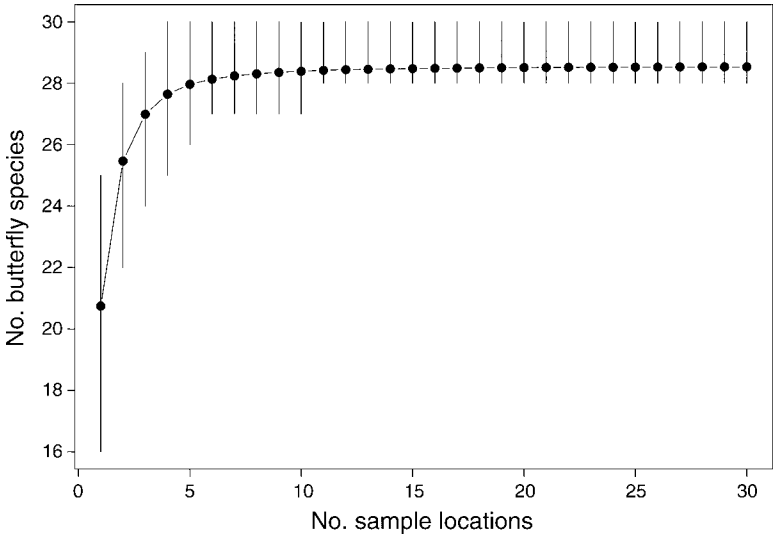

FIG. 8. Predicted species-accumulation curve for the community of butterflies. Each point along the curve corresponds to an estimate of the mean of the posterior-predictive distribution. Error bars indicate 90% prediction intervals.

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. Sampling deficiencies exert a direct effect on the interpretation of the parameter N, regardless of the model or method of estimation. We believe that sampling design and detection methods must be carefully considered in advance of the survey, so that every species that is present at a sample location is ensured to have some nonzero probability of being detected.

Our model-based estimates of N are influenced also by the form of the distribution used to specify heterogeneity in species occurrence and detection probabilities. We selected the bivariate normal distribution, but other parametric or semiparametric forms may also be considered. Such alternatives no doubt can influence the estimated value of N because, in computing 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 and that conventional diagnostics for assessing a model's goodness of fit cannot be relied upon for selecting a model that provides valid inferences for N (Dorazio and Royle 2003, Link 2003). This problem is especially acute in communities that are suspected to contain many rare species. In fact, Mao and Colwell (2005) have suggested that model-based estimates of N for these communities should be regarded as lower bounds that improve upon the observed number of species n, a negatively biased estimate of N.

Our model-based predictions of species accumulation are computed for a hypothetical sequence of locations that are spatially distinct from the sample locations. Therefore, while our predictions of species-accumulation depend on sample data, they do not depend on a particular ordering of the sample locations. As noted earlier (in *Model description: Predicting species accumulation as a function of species occurrence*), we view this dependence as an inherent limitation of interpolating (rarefying) or extrapolating empirical species-accumulation curves (Gotelli and Colwell 2001, Ugland et al. 2003). Furthermore, such dependence is usually not desired in practice. For example, in comparing the species richness of different communities at some common level of sampling effort (Colwell and Coddington 1994, Gotelli and Colwell 2001, Colwell et al. 2004), one normally does not want the comparison to be affected by the location or number of units in the original sample. Another advantage of our predictions of species accumulation is that they account for the uncertainty in estimating species richness and species occurrence. By adopting a Bayesian framework for inference and prediction, errors in estimation are automatically incorporated in the prediction intervals of our model-based, species-accumulation curve (Figs. 4 and 8).

Species accumulation curves can be used to improve the efficiency of future community surveys (Soberón and Llorente 1993, Colwell and Coddington 1994), so it is natural to inquire about the effects of sample size (number of spatial and temporal replicates) on our model-based estimates of species richness and accumulation. In *Protocol for sampling communities*, we noted that the spatial coverage of the sample must be sufficient to ensure that the sample is representative of the entire community. We also emphasized that a minimum of two visits are needed at each sample location so that the conditional probability of detection of a species (that is, given it is present) may be estimated separately from its probability of occurrence. But how many temporal replicates are needed? Should the number of temporal replicates vary among sample locations? Unfortunately answers to these questions depend on the underlying probabilities of detection of each species, and these probabilities usually are not known prior to completing the survey. However, less abundant species are generally more difficult to detect, so we may conclude that a higher number of temporal replicates is obviously desirable in communities that are thought to include several rare species. The detectability of a species also can vary with its habitat or behavior (e.g., as in "call surveys" of amphibians where species are detected aurally). In these circumstances, prior knowledge about differences in detectability seems essential before one could recommend a particular number or spatial pattern of temporal replicates. Perhaps a sequential sampling design is needed wherein the collection of each temporal replicate is followed immediately by data analysis to determine if estimates of model parameters and species richness satisfy prescribed levels of precision (or other design criteria). The development of such "stopping rules" for additional sampling could provide enormous practical benefits (e.g., time and cost savings) in routine assessments of biological diversity.

Acknowledgments

We thank two anonymous referees for their review comments, which helped to improve the exposition of our ideas.

Literature Cited

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

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

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

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

Cabeza, M., M. B. Araújo, R. J. Wilson, C. D. Thomas, M. J. R. Cowley, and A. Moilanen. 2004. Combining probabilities

of occurrence with spatial reserve design. Journal of Applied Ecology **41**:252–262.
Carlin, B. P., and T. A. Louis. 2000. Bayes and empirical Bayes methods for data analysis. Second edition. Chapman and Hall, Boca Raton, Florida, USA.
Coleman, B. D. 1981. On random placement and species–area relations. Mathematical Biosciences **54**:191–215.
Coleman, B. D., M. A. Mares, M. R. Willig, and Y. H. Hsieh. 1982. Randomness, area, and species richness. Ecology **63**: 1121–1133.
Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B **345**:101–118.
Colwell, R. K., C. X. Mao, and J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology **85**:2717–2727.
Connor, E. F., and E. D. McCoy. 1979. The statistics and biology of the species–area relationship. American Naturalist **113**:791–833.
Coull, B. A., and A. Agresti. 1999. The use of mixed logit models to reflect heterogeneity in capture–recapture studies. Biometrics **55**:294–301.
Dorazio, R. M., and J. A. Royle. 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics **59**:351–364.
Dorazio, R. M., and J. A. Royle. 2005. Estimating size and composition of biological communities by modeling the occurrence of species. Journal of the American Statistical Association **100**:389–398.
Fienberg, S. E., M. S. Johnson, and B. W. Junker. 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.
Fisher, B. L. 1999. Improving inventory efficiency: a case study of leaf-litter ant diversity in Madagascar. Ecological Applications **9**:714–731.
Gilks, W. R., A. Thomas, and D. J. Spiegelhalter. 1994. A language and program for complex Bayesian modelling. Statistician **43**:169–178.
Gotelli, N. J., and R. K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters **4**:379–391.
Hanski, I., and M. E. Gilpin. 1997. Metapopulation biology: ecology, genetics, and evolution. Academic Press, New York, New York, USA.
Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.
Kendall, W. L. 1999. Robustness of closed capture–recapture methods to violations of the closure assumption. Ecology **80**: 2517–2525.
Link, W. A. 2003. Nonidentifiability of population size from capture–recapture data with heterogeneous detection probabilities. Biometrics **59**:1123–1130.
Liu, Q., and D. A. Pierce. 1994. A note on Gauss–Hermite quadrature. Biometrika **81**:624–629.
Lomolino, M. V. 2000. Ecology's most general, yet protean pattern: the species–area relationship. Journal of Biogeography **27**:17–26.
MacArthur, R. H. 1972. Geographical ecology: patterns in the distribution of species. Harper and Row, New York, New York, USA.
MacArthur, R. H., and E. O. Wilson. 1967. The theory of island biogeography. Princeton University Press, Princeton, New Jersey, USA.
MacKenzie, D. I., J. D. Nichols, G. B. Lachman, S. Droege, J. A. Royle, and C. A. Langtimm. 2002. Estimating site occupancy rates when detection probabilities are less than one. Ecology **83**:2248–2255.
Mao, C. X., and R. K. Colwell. 2005. Estimation of species richness: mixture models, the role of rare species, and inferential challenges. Ecology **86**:1143–1153.
Mao, C. X., R. K. Colwell, and J. Chang. 2005. Estimating the species accumulation curve using mixtures. Biometrics **61**: 433–441.
Margules, C. R., R. L. Pressey, and P. H. Williams. 2002. Representing biodiversity: data and procedures for identifying priority areas for conservation. Journal of Biosciences **27**: 309–326.
McGuinness, K. A. 1984. Equations and explanations in the study of species–area curves. Biological Reviews of the Cambridge Philosophical Society **59**:423–440.
Nichols, J. D., and M. J. Conroy. 1996. Estimation of species richness. Pages 226–234 *in* D. E. Wilson, F. R. Cole, J. D. Nichols, R. Rudran, and M. Foster, editors. Measuring and monitoring biological diversity. Standard methods for mammals. Smithsonian Institution Press, Washington, D.C., USA.
Norris, J. L., III, and K. H. Pollock. 1996. Nonparametric MLE under two closed capture–recapture models with heterogeneity. Biometrics **52**:639–649.
Novotny, V., and Y. Bassett. 2000. Rare species in communities of tropical insect herbivores: pondering the mystery of singletons. Oikos **89**:564–572.
Peterson, D. L., and V. T. Parker, editors. 1998. Ecological scale: theory and applications. Columbia University Press, New York, New York, USA.
Pielou, E. C. 1977. Mathematical ecology. John Wiley, New York, New York, USA.
Pinheiro, J. C., and D. M. Bates. 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics **4**:12–35.
Pledger, S. 2000. Unified maximum likelihood estimates for closed capture–recapture models using mixtures. Biometrics **56**:434–442.
R Development Core Team. 2004. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Robbins, C. S., D. Bystrak, and P. H. Geissler. 1986. The breeding bird survey: its first fifteen years, 1965–1979. Resource publication 157. United States Fish and Wildlife Service, Washington, D.C., USA.
Robbins, C. S., J. R. Sauer, R. S. Greenberg, and S. Droege. 1989. Population declines in North American birds that migrate to the neotropics. Proceedings of the National Academy of Sciences (USA) **86**:7658–7662.
Rosenzweig, M. L. 1985. Species diversity in time and space. Cambridge University Press, Cambridge, UK.
Sauer, J. R., G. W. Pendleton, and B. G. Peterjohn. 1996. Evaluating causes of population change in North American insectivorous songbirds. Conservation Biology **10**:465–478.
Soberón, J. M., and J. B. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. Conservation Biology **7**:480–488.
Söderström, B., B. Svensson, K. Vessby, and A. Glimskär, 2001. Plants, insects and birds in semi-natural pastures in relation to local habitat and landscape factors. Biodiversity and Conservation **10**:1839–1863.
Sturtz, S., U. Ligges, and A. Gelman. 2005. R2WinBUGS: a package for running WinBUGS from R. Journal of Statistical Software **12**(3)::1–16.
Tanner, M. A. 1996. Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. Third edition. Springer-Verlag, New York, New York, USA.
Tardella, L. 2002. A new Bayesian method for nonparametric capture–recapture models in presence of heterogeneity. Biometrika **89**:807–817.

Ugland, K. I., J. S. Gray, and K. E. Ellingsen. 2003. The species-accumulation curve and estimation of species richness. Journal of Animal Ecology **72**:888–897.

Vessby, K., B. Söderström, A. Glimskär, and B. Svensson. 2002. Species-richness correlations of six different taxa in Swedish seminatural grasslands. Conservation Biology **16**:430–439.

Wiens, J. A. 2002. Predicting species occurrences: progress, problems, and prospects. Pages 739–749 *in* J. M. Scott, P. J. Heglund, M. L. Morrison, J. B. Haufler, M. G. Raphael, W. A. Wall, and F. B. Samson, editors. Predicting species occurrences: issues of accuracy and scale. Island Press, Washington, D.C., USA.

SUPPLEMENT

R and WinBUGS code for fitting the model of species occurrence and detection and example data sets (*Ecological Archives* E087-050-S1).