SPECIAL FEATURE: REVIEW

Bayesian data analysis in population ecology: motivations, methods, and benefits

Robert M. Dorazio[1]

Abstract During the 20th century ecologists largely relied on the frequentist system of inference for the analysis of their data. However, in the past few decades ecologists have become increasingly interested in the use of Bayesian methods of data analysis. In this article I provide guidance to ecologists who would like to decide whether Bayesian methods can be used to improve their conclusions and predictions. I begin by providing a concise summary of Bayesian methods of analysis, including a comparison of differences between Bayesian and frequentist approaches to inference when using hierarchical models. Next I provide a list of problems where Bayesian methods of analysis may arguably be preferred over frequentist methods. These problems are usually encountered in analyses based on hierarchical models of data. I describe the essentials required for applying modern methods of Bayesian computation, and I use real-world examples to illustrate these methods. I conclude by summarizing what I perceive to be the main strengths and weaknesses of using Bayesian methods to solve ecological inference problems.

This manuscript was submitted for the special feature based on a symposium in Tsukuba, Japan, held on 11 October 2014.

✉ Robert M. Dorazio
bdorazio@usgs.gov

[1] U.S. Geological Survey, Southeast Ecological Science Center, 7920 NW 71st Street, Gainesville, FL, USA

Introduction

During the 20th century scientists in many fields of study (including ecology) largely relied on the frequentist system of inference for the analysis of their data. This approach was appealing for at least two reasons. First, frequentist inference could be used to solve many of the problems routinely encountered by scientists (i.e., parameter estimation, prediction, hypothesis testing, and model criticism and selection). Second, the approach was practical in the sense that it could be implemented using commonly available computing resources and algorithms.

In contrast, the Bayesian approach to inference did not enjoy these advantages during most of the 20th century. Though Bayesian inference was much older (Bayes 1763; Laplace 1774a, b) and known to be remarkably simple and versatile (at least conceptually), relatively few classes of models could be fitted to data using commonly available computing resources (Jeffreys 1961; Box and Tiao 1973); thus Bayesian inference was not widely used by scientists. The main impediment to Bayesian analysis was the inability to compute high-dimensional integrations required for models that contained many parameters.

This computational barrier was essentially eliminated during the last two decades of the 20th century when the widespread availability of high-speed computing catalyzed the development of a set of generic algorithms for fitting Bayesian models to data. These algorithms, known collectively as *Markov chain Monte Carlo* (MCMC) algorithms, can be used to fit *any* Bayesian model, including models whose posterior density cannot be evaluated analytically or approximated numerically (Geyer 2011; Robert and Casella 2011).

Armed with MCMC algorithms, 21st-century scientists can easily conduct Bayesian analyses of their data. In some

cases software, such as WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml), JAGS (mcmc-jags.sourceforge.net), or Stan (mc-stan.org/), may be used to facilitate analysis. More importantly, however, is that scientists are free now to formulate entirely new classes of models for the analysis of their data, models that allow sampling processes, which lead to observations (data), to be linked with scientific processes, whose effects may not be directly observable but are usually the subjects of interest. Various adjectives have been used to describe these models, including "hierarchical", "state-space", or "multi-level"; however, a common feature of these models is that separate—and often different—assumptions are used to describe the variation in observed quantities and the variation in latent (unobserved) quantities of scientific interest.

The rapid development of new classes of hierarchical models has fueled, to some extent, the use of Bayesian inference. In the field of ecology, for example, Bayesian methods of analysis have been described in several recently published books (Clark 2007; McCarthy 2007; Royle and Dorazio 2008; King et al. 2010; Link and Barker 2010; Kéry and Schaub 2012; Parent and Rivot 2013). While the analyses in these books range from partially to fully Bayesian, all of these books describe various kinds of hierarchical models and illustrate the inferential benefits of using Bayesian methods to fit these models.

The present article is not intended as a review of the entirety of Bayesian models and topics described in these books; nor is this article intended to resurrect any ideological or philosophical comparisons of Bayesian and frequentist systems of inference. Instead, my goal is entirely utilitarian: I attempt in the present article to provide guidance to ecologists, in particular, to population ecologists, who would like to decide whether Bayesian methods can be used to improve their conclusions and predictions. I begin by providing a (hopefully) concise summary of Bayesian methods of analysis, including a comparison of differences between Bayesian and frequentist approaches to inference when using hierarchical models. Next I provide a list of problems where Bayesian methods of analysis may arguably be preferred over frequentist methods. These problems are usually encountered in analyses based on hierarchical models of data. In the third and fourth sections of this article, I describe the essentials required for applying modern methods of Bayesian computation (i.e., MCMC) and I use real-world examples to illustrate these methods. I conclude the article by summarizing what I perceive to be the main strengths and weaknesses of using Bayesian methods to solve ecological inference problems.

The Bayesian approach to learning

Suppose we have a model that describes the processes assumed to have produced an observed set of data. For the purposes of learning, this model is used to approximate the complexities of Nature's true data-generating "model." Let θ denote the parameters (a vector of unknown values) of the model, and let $\boldsymbol{y} = (y_1, \ldots, y_m)'$ denote the data (a vector of m observations). We wish to make inferences (that is, to learn) about θ given the information in $\boldsymbol{y}$.

Recall that under the frequentist approach to inference, the model and its unknown parameters θ are viewed as generating a hypothetical sequence of data sets, say

$$\theta \rightarrow \boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \boldsymbol{y}^{(3)}, \ldots$$

via repeated sampling or repeated experiments. Any member $\boldsymbol{y}^{(i)}$ of this sequence and its associated parameter estimate $\hat{\theta}^{(i)}$ are viewed as random outcomes when making inferences about θ.

Under the Bayesian approach to inference, the view that θ has an unknown fixed value which generates data $\boldsymbol{y}$ randomly under the model's assumptions remains valid. A distinguishing characteristic of the Bayesian approach is that uncertainty about the magnitude of θ *prior to observing the data* is specified formally using a *prior distribution*. In addition, once the data from a sample or experiment have been collected, $\boldsymbol{y}$ is regarded as fixed (not random) and all inferences about θ are made conditional on the observed value of $\boldsymbol{y}$ and the prior distribution. The Bayesian approach does *not* assume that θ is random (Laplace 1774b; Lindley 2000). Probability is simply used to express uncertainty in the magnitude of θ both before and after the data $\boldsymbol{y}$ have been observed. The Bayesian approach *does* assume that the path to updating prior uncertainty in θ must follow the axioms of probability theory and its calculus. Unlike frequentist inference, the Bayesian approach to inference does not rely on the idea of a hypothetical sequence of repeated or replicated data sets or on the asymptotic properties of estimators of θ. In fact, Bayesian probability statements (inferences) about θ are valid for any sample size, as will be shown shortly.

Conceptually, computing Bayesian inferences is quite straightforward. The calculus of probabilities is used for everything, including estimation of parameters, prediction of unknown quantities, and quantification of uncertainties. To be specific, let $[\boldsymbol{y}|\theta]$ denote the probability density of the data[1] given the assumptions of the model and the value of its parameters θ, and let $[\theta]$ denote the probability density

---

[1] I use bracket notation (Gelfand and Smith 1990) to specify probability density functions; thus, $[x, y]$ denotes the joint density of random variables X and Y, $[x \mid y]$ denotes the conditional density of X given $Y = y$, and $[x]$ denotes the unconditional (marginal) density of X.

that specifies prior beliefs about the magnitude of $\boldsymbol{\theta}$. Using the definitions of conditional and joint probabilities yields Bayes' rule:

$$[\boldsymbol{\theta}|\boldsymbol{y}] = \frac{[\boldsymbol{y}, \boldsymbol{\theta}]}{[\boldsymbol{y}]} = \frac{[\boldsymbol{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{[\boldsymbol{y}]}$$

where the marginal density of the data, $[\boldsymbol{y}] = \int [\boldsymbol{y}|\boldsymbol{\theta}][\boldsymbol{\theta}]\mathrm{d}\boldsymbol{\theta}$, is the normalizing constant for $[\boldsymbol{\theta}|\boldsymbol{y}]$. This density function is often called the *posterior density* because it expresses quantitatively how prior beliefs about the magnitude of $\boldsymbol{\theta}$ are changed given a model of the data $[\boldsymbol{y}|\boldsymbol{\theta}]$ and the information in $\boldsymbol{y}$. The posterior distribution of $\boldsymbol{\theta}$ is expressed as $\boldsymbol{\theta}|\boldsymbol{y}$ to indicate the explicit conditioning on the fixed value $\boldsymbol{y}$. The posterior provides the basis for *every* aspect of Bayesian inference (estimation, prediction, etc.). For example, results of Bayesian analyses are often reported using summaries of the posterior distribution of $\boldsymbol{\theta}$, such as the posterior mean $\mathrm{E}(\boldsymbol{\theta}|\boldsymbol{y})$ or variance $\mathrm{Var}(\boldsymbol{\theta}|\boldsymbol{y})$. Alternatively, quantiles of the posterior distribution (such as $\Pr(\boldsymbol{\theta}_\alpha \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_{1-\alpha}|\boldsymbol{y}) = 1 - \alpha$) may be used to express uncertainty in the magnitude of $\boldsymbol{\theta}$. Posterior quantiles are called *credible limits* to distinguish them from frequentist confidence limits, which are based on the uncertainty of estimates $\hat{\boldsymbol{\theta}}$ in hypothetical repeated samples or experiments.

A simple hierarchical model

In the previous section I described Bayesian inference in general terms. For some ecologists this description may be too abstract to be fully understood or appreciated. Therefore, to illustrate the simplicity of Bayesian learning, I describe in this section an analysis of data based on a hierarchical model that is widely used in ecology. This model was developed to estimate the occurrence of a species (plant or animal) from data collected in presence-absence surveys. Importantly, these surveys use a sampling protocol that allows species occurrences to be estimated while accounting for the effects of errors in detection.

Suppose m sites were selected at random from a region where a particular species was thought to be present. In addition, suppose J independent surveys were conducted at each site to determine whether this species was present or "absent" (more correctly, not detected). Let $y_i \in \{0, \ldots, J\}$ denote the number of surveys where the species was detected at the ith site. The vector $\boldsymbol{y} = (y_1, \ldots, y_m)'$ summarizes the observed data. The hierarchical model first proposed for the analysis of these data is often called a model of "site-occupancy" (MacKenzie et al. 2002; Tyre et al. 2003) or simply "occupancy" (MacKenzie et al. 2006; Royle and Dorazio 2008). Let Z_i denote a binary-valued random variable that indicates whether the species

was present ($Z_i = 1$) or absent ($Z_i = 0$) at site i. The simplest occupancy model makes two distributional assumptions:

$$Z_i \sim \mathrm{Bernoulli}(\psi)$$
$$Y_i | Z_i = z_i \sim \mathrm{Binomial}(J, pz_i)$$

where ψ denotes the probability of species occurrence (i.e., presence of one or more individuals of the species) and where p denotes the conditional probability of detecting the species given that it was present. We wish to learn about the values of the parameters ψ and p given the information in the observed data $\boldsymbol{y}$. In some cases we also want to learn about the value of $\bar{Z} = (1/m) \sum_{i=1}^{m} Z_i$, which equals the proportion of surveyed sites in which the species was present. This quantity is defined unambiguously regardless of the model actually used to analyze the data. Shortly we will see that $\bar{Z}$ can be viewed as a derived parameter of the occupancy model.

To fit this occupancy model using Bayesian methods, assume that $[\psi, p] = [\psi][p]$ specifies the prior density of parameters ψ and p. In other words, prior to observing the data, ψ and p are assumed to vary independently. The posterior density of the parameters $\boldsymbol{\theta} = (\psi, p, Z_1, \ldots, Z_m)'$ is

$$[\psi, p, \boldsymbol{z}|\boldsymbol{y}] = \frac{[\psi][p] \prod_{i=1}^{m} [z_i|\psi][y_i|J, pz_i]}{[\boldsymbol{y}]} \quad (1)$$

$$= \frac{[\psi][p]}{[\boldsymbol{y}]} \prod_{i=1}^{m} \psi^{z_i}(1-\psi)^{1-z_i} \binom{J}{y_i} (pz_i)^{y_i} (1-pz_i)^{J-y_i} \quad (2)$$

where the normalizing constant $[\boldsymbol{y}]$ cannot be expressed in closed form. I could have removed $\boldsymbol{z}$ from this density (by integration) to obtain the posterior density of just ψ and p:

$$[\psi, p|\boldsymbol{y}] = \frac{[\psi][p]}{[\boldsymbol{y}]} \prod_{i=1}^{m} \psi \binom{J}{y_i} p^{y_i}(1-p)^{J-y_i} + (1-\psi)I(y_i = 0) \quad (3)$$

$$= \frac{[\psi][p]C}{[\boldsymbol{y}]} \{\psi^n p^{y.}(1-p)^{nJ-y.}\} \times \{\psi(1-p)^J + 1 - \psi\}^{m-n} \quad (4)$$

$$= \frac{[\psi][p]}{[\boldsymbol{y}]} [\boldsymbol{y}|\psi, p] \quad (5)$$

where $n = \sum_{i=1}^{m} I(y_i > 0)$ is the number of sites where the species was detected, $y. = \sum_{i=1}^{m} y_i$ is the total number of detections in the sample of m sites, and $C = \prod_{i=1}^{m} \binom{J}{y_i}$. Note that $[\boldsymbol{y}|\psi, p]$ is the marginal likelihood function for the parameters $\boldsymbol{\theta} = (\psi, p)'$. The normalizing constants in Eqs. 2 and 5 are actually different, but neither can be expressed in closed form; therefore, a Bayesian analysis of the data gains little by integrating away the parameter values $\boldsymbol{z}$. In fact, doing so is actually counterproductive

when inferences are required for $\bar{Z}$, the proportion of surveyed sites in which the species was present. For example, it is easily proved that $\Pr(Z_i = 1 | y_i > 0) = 1$, that is, a species must have been present at site i if it was detected at that site. If the species was not detected at site i, we can show also that $\Pr(Z_i = 1 | y_i = 0) = \psi(1-p)^J / \{\psi(1-p)^J + 1 - \psi\}$. These two results imply that $\bar{Z}$ can be estimated as follows:

$$\bar{Z} = (n + N_0)/m \tag{6}$$

where n is the number of sites where the species was detected and where N_0 is a random variable for the number of sites where the species was present but not detected. Given the model parameters, N_0 has the following distribution:

$$N_0 | n, \psi, p \sim \text{Binomial}(m - n, \ \psi(1-p)^J / \{\psi(1-p)^J + 1 - \psi\})$$

Therefore, if the values of ψ and p were known, we could estimate the expected value of $\bar{Z}$ using

$$\mathrm{E}(\bar{Z} | n, \psi, p) = \frac{n}{m} + \frac{m-n}{m} \times \frac{\psi(1-p)^J}{\psi(1-p)^J + 1 - \psi} \tag{7}$$

Of course, the parameters ψ and p are unknown and must be estimated from the data. In addition, the estimator of $\bar{Z}$ should account for posterior uncertainty in these parameters. In a Bayesian analysis of the data, probability calculations are used to infer the values of $\bar{Z}$. Specifically, the posterior probability of a particular value of $\bar{Z}$ is equivalent to the posterior probability of a particular value of N_0 (owing to Eq. 6):

$$\Pr(\bar{Z} = (n + n_0)/m \,|\, \mathbf{y})$$
$$= \int_0^1 \int_0^1 \Pr(N_0 = n_0 | n, \psi, p) \, [\psi, p | \mathbf{y}] \, \mathrm{d}\psi \mathrm{d}p \tag{8}$$

where $\bar{Z} \in \{n/m, (n+1)/m, \ldots, 1\}$ and where the probability in the integrand corresponds to the binomial mass function of the conditional distribution of N_0. Equation 8 shows that the posterior probability of a particular value of $\bar{Z}$ averages (integrates) the binomial probabilities of n_0 over the posterior distribution of ψ and p values. A similar calculation yields the posterior expectation of $\bar{Z}$:

$$\mathrm{E}(\bar{Z} | \mathbf{y}) = \int_0^1 \int_0^1 \mathrm{E}(\bar{Z} | n, \psi, p) \, [\psi, p | \mathbf{y}] \, \mathrm{d}\psi \mathrm{d}p \tag{9}$$

$$= \frac{n}{m} + \frac{m-n}{m} \int_0^1 \int_0^1 \frac{\psi(1-p)^J}{\psi(1-p)^J + 1 - \psi} \, [\psi, p | \mathbf{y}] \, \mathrm{d}\psi \mathrm{d}p \tag{10}$$

To illustrate a Bayesian analysis using real data, I fitted the occupancy model to a set of salamander data that have been analyzed previously using frequentist methods

(MacKenzie et al. 2006, p. 99–101). The data were obtained by conducting $J = 5$ biweekly presence-absence surveys of Blue-Ridge two-lined salamanders (*Eurycea wilderae*) at each of $m = 39$ sample locations (sites). During the period of sampling (April to mid-June) the occupancy status (presence or absence) of these salamanders was assumed to have been constant at each site. Salamanders were detected at $n = 18$ of the sites that were surveyed. The number of surveys where salamanders were detected equaled one (12 sites), two (1 site), three (4 sites), or four (1 site).

In the analysis of these data, I assumed uniform prior distributions for ψ and p to specify prior ignorance about the values of these parameters. For now, let's ignore the technical details of the analysis (MCMC algorithm provided in Electronic Supplementary Material (ESM) S1) and focus on computing inferences for the model's parameters and for the proportion of sites occupied by this species. The estimated posterior distributions of ψ, p, and $\bar{Z}$ are plotted in Fig. 1. The probability of detecting this species during a single survey appears to have been much less than one (posterior mean = 0.26; 95 % credible interval = 0.16–0.38), so the observed proportion of sites where salamanders were detected ($n/m = 0.46$) is likely to
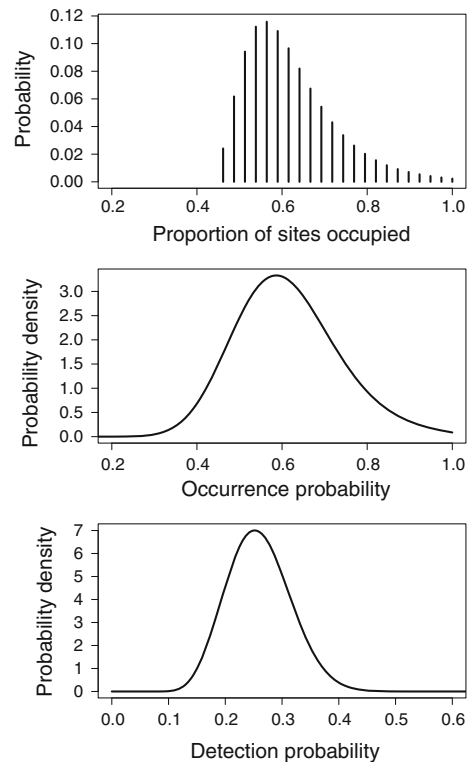
Fig. 1 Estimated posterior distribution of the proportion of sites occupied $\bar{Z}$ and of the probabilities of occurrence ψ and detection p obtained by fitting a Bayesian occupancy model to data collected during presence-absence surveys of the Blue-Ridge two-lined salamander

be an underestimate of the true occurrence probability of salamanders. In fact, the estimated posterior mean of ψ equals 0.61 (95 % credible interval = 0.40–0.88), which is about 30 % higher than the naive estimate provided by n/m. The estimated posterior mean for the proportion of sites occupied by salamanders equals 0.62, only slightly higher than that of ψ. However, the estimated 95 % credible interval for $\bar{Z}$ ranges from 0.46 to 0.87, indicating that uncertainty about $\bar{Z}$ is lower than that of ψ.

To compare the Bayesian analysis of the salamander data with a frequentist analysis, I computed maximum likelihood estimates (MLEs) of ψ and p using $[\mathbf{y}|\psi, p]$ as the marginal likelihood function. The MLEs and 95 % confidence intervals ($\hat{\psi} = 0.59$ (0.35–0.80); $\hat{p} = 0.26$ (0.16–0.39)) are similar to the Bayesian estimates, which is not surprising given the sample size and my choice of prior. In a frequentist analysis $\bar{Z}$ is estimated using the so-called "best unbiased predictor" (Laird and Ware 1982), which equals $\hat{\bar{Z}} = \mathrm{E}(\bar{Z}|n, \hat{\psi}, \hat{p}) = 0.59$ (using Eq. 7). However, constructing a confidence interval for $\bar{Z}$ is somewhat problematic. The delta method can be used to approximate $\mathrm{Var}(\hat{\bar{Z}})$ from the estimated covariance matrix of the MLEs, but what distribution should be used to compute the confidence interval? As an alternative Laird and Louis (1987) proposed that confidence intervals be computed using parametric bootstrapping. In this approach the MLEs are used to simulate an arbitrarily large number of data sets and the best unbiased predictor is computed for each of these simulated data sets. A confidence interval for $\bar{Z}$ then would be estimated using quantiles of the simulated distribution of best unbiased predictions. Applying this approach to the salamander data yields a 95 % confidence interval of 0.38–0.92. The lower limit of this confidence interval presents a problem because it falls below $n/m = 0.46$, the proportion of sites where salamanders were observed. The problem occurs because salamanders were "observed" at fewer than n sites in some of the simulated data sets. Therefore, while the parametric bootstrapping approach of Laird and Louis (1987) may be useful in some cases, the approach also can produce undesirable or unexpected results, as with the salamander data.

Pros and cons of Bayesian inference

The Bayesian approach to learning is compelling because it is conceptually and mathematically coherent and it is entirely prescriptive. As noted by Little (2006, 2011), once a model and prior distribution are specified, probability theory provides a clear path to inference regardless of the target. The posterior distribution is used for *every* inference regardless of whether the target is a model parameter, a function of model parameters, or a model-based prediction. No other system of inference—frequentist or otherwise—is as clear and complete as the Bayesian approach to learning.

Another benefit is that Bayesian inferences are valid for *any* sample size. This benefit is particularly useful in analyses of ecological data sets that contain relatively few observations. In these cases the asymptotic assumptions that form the basis of frequentist confidence intervals can be violated, leading to incorrect inferences and conclusions. While Bayesian inferences are valid in these cases, information in the prior distribution can exert a substantial effect on the results. For this reason, having to specify a prior distribution can be viewed as the "price" paid for the validity of Bayesian inference.

Prior specification may be regarded as a benefit or as a disadvantage of Bayesian methods. In problems where prior information about a model's parameters exists or may be elicited (say, using expert opinion), Bayes' rule specifies precisely how such information should be used when computing inferences from data. On the other hand, in problems where little is known about a model's parameters prior to data collection, an analyst may want to specify a prior that is vague or non-informative. Unfortunately, the definition of non-informative has proven to be elusive. Considerable debate remains among statisticians that have given the issue of non-informative priors careful thought (Berger 2006). As a practical matter, I believe that specification of non-informative (or weakly-informative) priors generally requires scientific context because most prior distributions are not invariant to transformation of parameters. For example, a prior that is uniform for $\boldsymbol{\theta}$ can be quite non-uniform (and informative) for $g(\boldsymbol{\theta})$, a one-to-one transformation of $\boldsymbol{\theta}$. As an illustration, recall that I assumed a uniform prior distribution for ψ in the occupancy model described earlier. In more complicated models this parameter is often transformed to the logit scale (say, $\beta = \log\{\psi/(1 - \psi)\}$). It is easily proved that the prior density of β induced by assuming a uniform prior for ψ is $p(\beta) = \exp(\beta)/\{1 + \exp(\beta)\}^2$, which is a symmetric, unimodal function centered at zero with non-negligible probability mass in the vicinity of zero (e.g., $\Pr(-3 \le \beta \le 3) \doteq 0.905$). Therefore, assuming a uniform prior for ψ implies an informative prior for its logit.

An important limitation of Bayesian (and frequentist!) methods of analysis is that they fail to include a unified set of procedures for comparing alternative models of data and for assessing the adequacy of an individual model. In practice, model building often iterates between proposing changes in model complexity and evaluating the consequences of those proposals. Therefore, frequentist ideas would seem to have an important role to play in model development and assessment. Models that

are selected to have good frequentist properties should be less prone to errors in specification. For example, if a model is selected so that its 95 % credible interval covers an unknown parameter or an out-of-sample prediction 95 % of the time in (hypothetical) repeated surveys or experiments, then the model would seem to be acceptable. A wide variety of approaches have been developed for Bayesian model comparison (Royle and Dorazio 2008; Hooten and Hobbs 2015), but none of these approaches has produced a clear winner among statisticians. Instead, there appears to be a growing appreciation that while inferences and predictions are best conducted using Bayesian methods, the evaluation and comparison of models are best accomplished from a frequentist perspective (Box 1980; Rubin 1984; Draper 1996; Gelman et al. 1996; Little 2006; Gelman 2011; Little 2011). The idea here is that if a model's inferences and predictions are to be well-calibrated, they should have good operating characteristics in hypothetical repeated samples (Royle and Dorazio 2008).

Motivations for Bayesian analysis

In the analysis of real data, I am generally pragmatic when deciding between Bayesian and frequentist methods of analysis. I am content to use frequentist methods in circumstances where those methods have the potential to produce sensible and useful results. These circumstances usually involve relatively simple models and large sample sizes (i.e., sufficiently large for the application of frequentist asymptotics). As a corollary, however, I recognize that not all inference problems can be solved with frequentist methods, and I view Bayesian methods as essential for the solution of these problems.

The following list contains problems or circumstances where I believe Bayesian methods of analysis can arguably be preferred over frequentist methods. These problems are often encountered when using hierarchical models of data that link a submodel of sampling processes with a submodel of ecological processes, whose latent (unobserved) effects are often the subject of interest.

Inference for latent state variables In many problems the target of inference is a set of parameters or state variables that lies at an intermediate level of a hierarchical model. The set of occupancy state variables $\{Z_1, \ldots, Z_m\}$ defined in hierarchical models of presence-absence data (MacKenzie et al. 2006; Royle and Dorazio 2008; Dorazio and Taylor Rodríguez 2012; Johnson et al. 2013) is an example. The set of site-specific abundance parameters defined in metapopulation models of spatially referenced counts (Royle and Dorazio 2006, 2008 chapter 8) is another example. In other problems interest may be

focused on a quantity whose definition transcends a particular model but can be computed as a function of a model's latent state variables (e.g., the proportion of sites occupied in presence-absence surveys). In still other problems, predictions of latent state variables at unsampled locations, such as predictions of spread of invasive species (Wikle 2003; Hooten and Wikle 2007), may be the target of inference. A Bayesian analysis is straightforward to apply in all of these inference problems. In contrast, the frequentist solution to these problems, often referred to as "best unbiased prediction" (Laird and Ware 1982) or "empirical Bayes estimation" (Morris 1983; Laird and Louis 1987), does not provide an automatic procedure that can account fully for the uncertainty involved in estimating a hierarchical model's parameters.

Missing data problems Ideally surveys or experiments are conducted such that responses and predictors are measured for every unit of observation. Statistical analysis then may be conducted on a complete set of data. Unfortunately, missing observations are commonplace in real surveys and experiments. The sources of missingness may be random (e.g., equipment failure preventing sampling), or they may be induced by design (e.g., non-detection of animals in capture-recapture surveys). Missing observations are also common in restrospective analyses of observational surveys, where the same sample locations were not necessarily surveyed each year during some longer period of time. Whatever the cause, accounting for missingness with frequentist methods generally requires difficult marginalizations to evaluate the likelihood function for the observed, but incomplete, data (Little and Rubin 2002). In contrast, Bayesian methods—especially those based on MCMC sampling—provide a straightforward solution to inference problems with missing data. Kéry et al. (2009), for example, showed how Bayesian methods could be used to estimate temporal trends in abundance even though survey data were missing from different locations in different years.

Intractable likelihood functions In some classes of statistical models the likelihood function of the observed data cannot even be evaluated, much less maximized. The number of parameters in these models typically increases with the sample size and some parameters must be removed (by integration) to obtain the observed data likelihood. A problem occurs when the integrations cannot be done analytically and cannot be accurately approximated using numerical methods or simulation. The likelihood function for models of non-Gaussian responses with spatial dependence (Diggle et al. 1998) provides a good example. Bayesian methods based on MCMC algorithms are essential for fitting these models. For example, Bayesian analyses have been used to predict maps of spatially correlated, latent abundances and

occurrences of wildlife species (Royle and Wikle 2005; Johnson et al. 2013).

Complex models of different sources and types of data Using modern methods of Bayesian computation, ecologists may formulate and fit complicated models that allow inferences to be drawn from different sources and types of data. An example is the class of integrated population models (Buckland et al. 2007; Schaub and Abadi 2011) that allow a population's dynamics to be estimated using a joint analysis of demographic data (e.g., capture-recapture surveys of individuals) and time-series data (e.g., annual counts of the population). In some cases the different types of data can be analyzed separately with or without using Bayesian methods. In other cases information shared among different data types allows inferences to be computed for parameters that otherwise would not be estimable. While parameter identifiability is naturally a concern in these cases and must be considered, the Bayesian approach to inference provides a straightforward way to fit these models.

Note that my list does not include two items that are often cited as motivations for conducting a Bayesian analysis: use of prior information and small sample size. In analyses of data from ecological surveys or experiments, prior information about the targets of inference is often unavailable or ignored, so it is hard to argue that methods of ecological analysis are driven by access to prior information. On the other hand, when prior information does exist, there is no doubt that inferences can benefit from this information (e.g., see McCarthy and Masters 2005; Morris et al. 2013). Regarding sample size, Bayesian methods should definitely be preferred for valid inference when there is little data; however, an analyst must be willing to accept that the prior can influence the results of the analysis in these cases. In my experience when ecologists are confronted with this reality, they often become motivated to collect additional data so that the effects of the prior become less important to their conclusions.

MCMC methods of fitting Bayesian models

There are many texts that describe Bayesian computional methods and MCMC methods in particular. One of the earliest book-length treatments by Tanner (1996) is still highly relevant; however, a more comprehensive and up-to-date treatment of MCMC methods is provided in the handbook edited by Brooks et al. (2011). In this section I describe the essentials required for applying modern methods of Bayesian computation. My intention is to provide an introduction to the subject, not an exhaustive description of MCMC methods.

Recall that for many models the normalizing constant $[y] = \int [y|\theta][\theta] d\theta$ of the posterior density function $[\theta|y]$ cannot be calculated, especially when models contain many parameters. In these cases summaries of the posterior distribution can be estimated using MCMC algorithms. The idea is to construct a Markov chain for the model's parameters whose stationary (invariant) distribution is equivalent to the posterior $\theta|y$ provided a set of technical conditions are satisfied (irreducibility, Harris recurrence, reversibility). By simulating an arbitrarily long chain (say, $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$), the posterior expectation of a function $g(\theta)$,

$$\mathrm{E}(g(\theta)|y) = \int g(\theta)[\theta|y] d\theta,$$

can be approximated by the arithmetic average

$$\bar{g}_M = \frac{1}{M} \sum_{i=1}^{M} g(\theta^{(i)}).$$

This estimator is justified through the Markov chain strong law of large numbers: If $\mathrm{E}(|g(\theta)| \,|y) < \infty$, then $\bar{g}_M \to \mathrm{E}(g(\theta)|y)$ almost surely as $M \to \infty$ (Flegal and Jones 2011; Geyer 2011). Therefore, by using an appropriate MCMC algorithm, summaries of the posterior distribution can be estimated using ergodic averages without having to evaluate the posterior's normalizing constant.

An important practical consideration for using these theoretical results involves the choice of M. A Markov chain must be run sufficiently long to have converged to its stationary distribution; however, in practice the assessment of convergence is usually subjective and based on experience. In special cases M can be determined a priori if one establishes a drift and minorization condition for the chain (Hobert 2011). More commonly, however, Bayesian analyses are conducted using black box MCMC (as coined by Geyer 2011) wherein a Markov chain is constructed with the appropriate stationary distribution (the posterior) but nothing is known about the chain's dynamics or its stationary distribution except what may be learned from running the chain. In this situation the large of amount of theory about convergence of Markov chains is not really applicable (Geyer 2011).

This situation has generated a persistent folklore about the practice of MCMC methods. This folklore includes the idea of computing convergence diagnostics from multiple, independent Markov chains constructed to have the same stationary distribution. One such diagnostic, the "potential scale reduction factor" (denoted by $\hat{R}$) (Gelman and Shirley 2011) is based on a comparison of variances within and between different Markov chains. Although this diagnostic seems sensible, it cannot be used to determine with certainty that a Markov chain has been run sufficiently long to

guarantee convergence. Again, the convergence theory of Markov chains does not suggest that $\hat{R}$ or any other diagnostic provides a foolproof measure for assessing convergence.

Given that posterior summaries must be estimated from a finite-length Markov chain containing unknown amounts of autocorrelation, every Bayesian analyst using MCMC methods should estimate and report the magnitude of Monte Carlo error associated with each estimate of a posterior summary. Methods for estimating Monte Carlo error are described by Flegal and Jones (2010, 2011). Although Monte Carlo error generally declines with the length of a Markov chain, an estimate of this error indicates the reliability of an estimate of the posterior summary. In practice, the estimated Monte Carlo error can be used to construct an asymptotically valid confidence interval for any posterior summary (Flegal and Jones 2011), thereby allowing the reliability of the estimate of the posterior summary to be assessed.

Several MCMC algorithms exist for constructing a Markov chain whose stationary distribution is equivalent to the posterior. I conclude this section by describing two of these algorithms.

The Gibbs sampling algorithm

One of the most widely used MCMC algorithms is called the *Gibbs sampler*. This algorithm is a special case of the *Metropolis-Hastings* algorithm, and the two are often used together to produce efficient hybrid algorithms that are relatively easy to implement. To illustrate the Gibbs sampler, consider a model that includes three parameters (that is, $\theta = (\theta_1, \theta_2, \theta_3)'$) and assume that the full conditional distributions of each parameter are relatively easy to sample. These distributions are defined by conditioning on the data and on the value of all other parameters as follows:

- $\theta_1 | \theta_2, \theta_3, \boldsymbol{y}$
- $\theta_2 | \theta_1, \theta_3, \boldsymbol{y}$
- $\theta_3 | \theta_1, \theta_2, \boldsymbol{y}$

To begin the Gibbs sampler, an arbitrary set of initial values is assigned to each parameter, i.e., $\theta_1 = \theta_1^{(0)}$, $\theta_2 = \theta_2^{(0)}$, $\theta_3 = \theta_3^{(0)}$, where the superscript in parentheses denotes the iteration of the Gibbs sampler. The Gibbs sampling algorithm proceeds by taking a random draw from each full conditional distribution in sequence and using the updated value of each parameter in subsequent draws as follows:

Step 1:　Draw $\theta_1^{(1)} \sim \theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \boldsymbol{y}$
Step 2:　Draw $\theta_2^{(1)} \sim \theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \boldsymbol{y}$
Step 3:　Draw $\theta_3^{(1)} \sim \theta_3 | \theta_1^{(1)}, \theta_2^{(1)}, \boldsymbol{y}$

This sequence of draws completes one iteration of the Gibbs sampler and generates a new set of parameter values, say $\theta = \theta^{(1)}$. Steps 1–3 are then repeated using the values of θ from the previous iteration to obtain the Markov chain $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots$. Note that the order in which parameters are updated does not matter, though it may affect how rapidly the chain converges to its stationary distribution.

The Metropolis-Hastings sampling algorithm

Suppose the full conditional distributions are familiar and easy to sample for some, but not all, parameters of a model. In this case the unfamilar full conditionals may be sampled using an alternative algorithm, the Metropolis-Hastings sampler, whereas the remaining full conditionals may be sampled using Gibbs. This approach leads to a hybrid MCMC algorithm, and many classes of Bayesian models may be fitted by adopting this approach.

To describe the Metropolis-Hastings (MH) algorithm, suppose the normalizing constant C for the full conditional distribution of a parameter θ is unknown. In other words suppose only the non-negative function $f(\theta)$ that satisfies the following equation is known:

$$f(\theta) = C \, p(\theta)$$

where $p(\theta)$ is the unknown density function and $C = \int_{-\infty}^{\infty} f(\theta) \mathrm{d}\theta$. The MH algorithm depends on a *proposal* distribution to provide a potential update of the state of the Markov chain: $\theta^{(t)} \to \theta^{(t+1)}$. The proposal distribution must be easy to sample and have known probability density function $g(\theta)$. Often—though not always—the proposal specifies an update that depends on the current state of the Markov chain $\theta^{(t)}$. For this reason, the density of the proposal distribution is often written conditionally as $g(\theta | \theta^{(t)})$. Using this notation, the MH sampling algorithm is:

Step 1:　Draw $\theta \sim g(\theta | \theta^{(t)})$.
Step 2:　Calculate the Hastings ratio $R = \{f(\theta) \, g(\theta^{(t)} | \theta)\} / \{f(\theta^{(t)}) \, g(\theta | \theta^{(t)})\}$
Step 3:　Accept θ as $\theta^{(t+1)}$ with probability $a = \min(R, 1)$
　　　　　(that is, assign $\theta^{(t+1)} = \theta$ with probability a, and assign $\theta^{(t+1)} = \theta^{(t)}$ with probability $1 - a$).

The efficiency of a MH algorithm depends, of course, on the choice of proposal distribution. If θ contains many parameters, a multivariate-t or multivariate normal distribution is often used as a proposal by selecting its parameters to approximate the target distribution. This can be accomplished by assigning the mean of the proposal

distribution to equal the value of $\boldsymbol{\theta}$ that maximizes $\log(f(\boldsymbol{\theta}))$, which may be found analytically or numerically depending on the form of $f(\boldsymbol{\theta})$. The covariance of the proposal distribution can be computed by inverting the negative of the hessian matrix of $\log(f(\boldsymbol{\theta}))$ evaluated at its maximum. The objective is to construct a proposal distribution that allows the MH algorithm to sample regions of high posterior density and not waste computing time by proposing values of $\boldsymbol{\theta}$ with low posterior density.

The Gibbs and MH samplers are two of the most widely used MCMC algorithms. Other algorithms exist (Brooks et al. 2011), but many Bayesian models can be fitted using either or both of these algorithms. All of the models described in the present article were fitted using Gibbs and MH algorithms (see ESM).

Examples of Bayesian analysis

In this section I illustrate how Bayesian methods of analysis can be used to solve two different inference problems. The first is a missing-data problem that requires inferences for the abundance of meadow voles (*Microtus pennsylvanicus*) surveyed by capture-recapture methods. The second problem involves an intractable likelihood function that is produced by modeling unobserved sources of spatial dependence among the abundances of seaside sparrows (*Ammodramus maritimus mirabilis*) surveyed by double-observer sampling.

Capture-recapture with individual-level covariates

The data in this example were collected using capture-recapture surveys of meadow voles at the Patuxent Wildlife Research Center in 1981 (Nichols et al. 1984). Each of the voles encountered in these surveys was weighed and examined to assess its sexual identity. The data have been analyzed in many journal articles and books (e.g., Williams et al. 2002, chapter 19). Here I analyze the recaptures of male voles that were trapped and released during each of five consecutive nights beginning June 27. Trapping yielded $n = 56$ male voles, and individuals with higher body masses were often captured more frequently (Fig. 2).

It seems clear from Fig. 2 that the effect of body mass on capture probability should be considered when formulating a model to estimate the unknown abundance of voles. The problem is that the body masses of uncaptured voles are not available; therefore, a model of these missing data is needed to solve the inference problem. Let X denote a random variable for the body mass of an individual vole, and let Y denote a random variable for the number of trapping periods (nights) in which an individual vole was
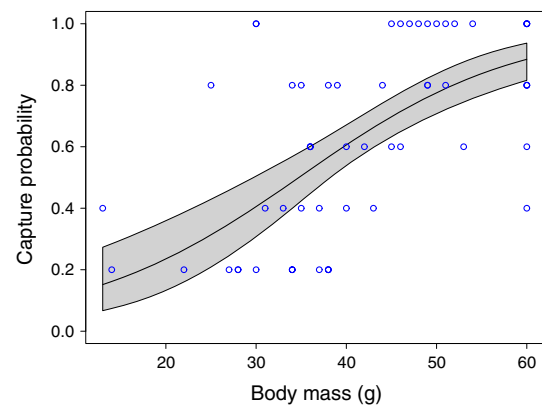

Fig. 2 Increase in predicted capture probability of meadow voles with individual body mass. Estimated mean of the posterior-predictive distribution is plotted using a *solid line*. Shading indicates 95 % credible interval. *Points* indicate the observed proportion of nights (y / J) that each vole was captured

captured out of $J = 5$ total periods. To relate these variables, assume

$$X \sim \text{Normal}(\mu, \sigma^2)$$
$$Y | X = x \sim \text{Binomial}(J, p(x))$$

where $\text{logit}(p(x)) = \alpha_0 + \alpha_1 x$ specifies an individual's capture probability $p(x)$ as a logit-linear function of its body mass x and two unknown parameters (α_0 and α_1).

Recall that X and Y were observed only for the n individual voles that were captured. Following the conventional notation used in capture-recapture models of demographically closed populations, let N denote the unknown abundance of voles that were present and available to be captured in the traps. For the $N - n$ unobserved voles the number of captures is known ($Y = 0$ for these individuals); however, their body masses are not known. A frequentist analysis of these data could be conducted using the following likelihood function of the observed, but incomplete, data:

$$[\boldsymbol{x}, \boldsymbol{y}, n | N, \boldsymbol{\alpha}, \mu, \sigma] = \binom{N}{n} \pi_0^{N-n} \prod_{i=1}^{n} \text{N}(x_i | \mu, \sigma^2) \, \text{Bin}(y_i | J, p(x_i))$$

(11)

where $\text{N}(x_i | \mu, \sigma^2)$ denotes the (normal) probability density of x_i, $\text{Bin}(y_i | J, p(x_i))$ denotes the (binomial) probability of y_i, and π_0 is the unconditional probability that an individual vole was not captured during the J trapping periods. Computing π_0 requires a marginalization over the unknown body mass of this individual as follows:

$$\pi_0 = (\sqrt{2\pi}\sigma)^{-1} \int_{-\infty}^{\infty} \exp\{-(x - \mu)^2/(2\sigma^2)\}$$
$$\{1 + \exp(\alpha_0 + \alpha_1 x)\}^{-J} \, dx$$

(12)

Unfortunately, this integral cannot be evaluated in closed form and must be approximated numerically. However, a frequentist analysis can be conducted by finding the value of the parameter vector $\boldsymbol{\theta} = (N, \boldsymbol{\alpha}, \mu, \sigma)'$ that maximizes $[\boldsymbol{x}, \boldsymbol{y}, n | N, \boldsymbol{\alpha}, \mu, \sigma]$ while evaluating the integral at each iteration of the maximization.

Suppose the model is expanded by assuming $N \sim \text{Binomial}(M, \psi)$, where M is a known upper limit for vole abundance and ψ is an unknown parameter to be estimated along with other parameters of the model. The likelihood function of this expanded model (after marginalizing N) is

$$[\boldsymbol{x}, \boldsymbol{y}, n | \psi, \boldsymbol{\alpha}, \mu, \sigma] = \binom{M}{n} \psi^n (\psi \pi_0 + 1 - \psi)^{M-n} \prod_{i=1}^{n}$$
$$\text{N}(x_i | \mu, \sigma^2) \; \text{Bin}(y_i | J, p(x_i))$$
$$(13)$$

A frequentist analysis could be conducted by maximizing this likelihood function, and an empirical Bayes estimator of N could be computed using the MLEs (indicated by circumflexes on parameters) as follows:

$$\hat{N} = \text{E}(N | n, \hat{\psi}, \hat{\boldsymbol{\alpha}}, \hat{\mu}, \hat{\sigma}) = n + (M - n) \times \frac{\hat{\psi} \hat{\pi}_0}{\hat{\psi} \hat{\pi}_0 + 1 - \hat{\psi}}$$

Unfortunately, this analysis still requires π_0 to be evaluated at each iteration of the maximization, and estimating a confidence interval for $\hat{N}$ that accounts for uncertainty in the MLEs is problematic.

I introduced this expanded model because the Binomial(M, ψ) model of N is implied when the capture-recapture data are analyzed using Bayesian inference and parameter-expanded data augmentation (Royle and Dorazio 2012). To apply this approach, the original vector $\boldsymbol{y}$ of n capture frequencies is augmented with $M - n$ zeros, where M is a known constant that specifies the maximum value of N. The value of M can be assigned using prior information about N or, in the absence of prior information, M can be assigned an arbitrarily high value (Royle and Dorazio 2012). The assumptions used to model the augmented data are similar to those of the original model:

$$Z \sim \text{Bernoulli}(\psi)$$
$$X \sim \text{Normal}(\mu, \sigma^2)$$
$$Y | X = x, Z = z \sim \text{Binomial}(J, z\, p(x))$$

where Z is a binary-valued random variable that denotes whether an individual in the augmented data set belongs ($Z = 1$) or does not belong ($Z = 0$) to the population of N voles that could have been captured in the traps. In this model the abundance of voles is a derived parameter, specifically $N = \sum_{i=1}^{M} Z_i$. The unnormalized posterior density for the augmented data is

$$[\psi, \boldsymbol{\alpha}, \mu, \sigma, \boldsymbol{z}, \boldsymbol{x}_{-n} | \boldsymbol{y}, \boldsymbol{x}_n] \propto [\psi, \boldsymbol{\alpha}, \mu, \sigma] \prod_{i=1}^{M} \text{Bern}(z_i | \psi)$$
$$\text{N}(x_i | \mu, \sigma^2) \; \text{Bin}(y_i | J, z_i\, p(x_i))$$

where $\boldsymbol{x}_n = (x_1, \ldots, x_n)'$ denotes the vector of body masses of captured individuals and where $\boldsymbol{x}_{-n} = (x_{n+1}, \ldots, x_M)'$ denotes the vector of unobserved body masses that must be estimated.

I completed a Bayesian analysis of the vole data using MCMC methods. In this analysis I centered and scaled the body mass measurements to have zero mean and unit variance. Details about the MCMC algorithm and prior distributions are described in ESM S2. The results of my analysis are similar to those reported by Royle and Dorazio (2008) table 6.4, who used different priors and the Win-BUGS software program. Estimates of posterior summaries of the model parameters are given in Table 1. The results corroborate the positive association between capture probability and body mass that was evident in the raw data and suggest that the posterior means of vole capture probabilities ranged from about 0.15 to 0.88 (Fig. 2). The estimated posterior mean of μ was slightly less than zero, the average of the standardized body masses of captured voles. This result is consistent with the positive association between capture probability and body mass—that is, on average the uncaptured voles must have had lower body masses than did voles in the sample. The estimated posterior distribution of vole abundance N is skewed (Fig. 3) but has most of its mass near the observed number of voles ($n = 56$). This result is consistent with the relatively high capture probabilities obtained by trapping these animals on five successive evenings.

An important benefit of the Bayesian analysis is that solves the inference problem for N completely. In contrast, suppose a frequentist analysis was conducted by maximizing the marginal likelihood in Eq. 13, computing the empirical Bayes estimator $\hat{N}$, and approximating the uncertainty in $\hat{N}$ by the delta method. What sampling distibution should be used to compute a confidence interval for N? The frequentist approach to this inference problem

Table 1 Estimates of posterior summary statistics for parameters of the capture-recapture model fitted to the meadow vole data

Parameter	Mean	2.5 %	97.5 %
N	59.534 (0.0462)	56.000 (0.0059)	67.000 (0.1441)
α_0	0.578 (0.0007)	0.300 (0.0016)	0.859 (0.0015)
α_1	0.980 (0.0015)	0.657 (0.0020)	1.327 (0.0026)
μ	−0.096 (0.0034)	−0.436 (0.0061)	0.209 (0.0043)
σ	1.064 (0.0027)	0.866 (0.0024)	1.330 (0.0053)

Limits of 95 % credible intervals are indicated by the columns labeled 2.5 and 97.5 %. Monte Carlo standard errors are given in parentheses
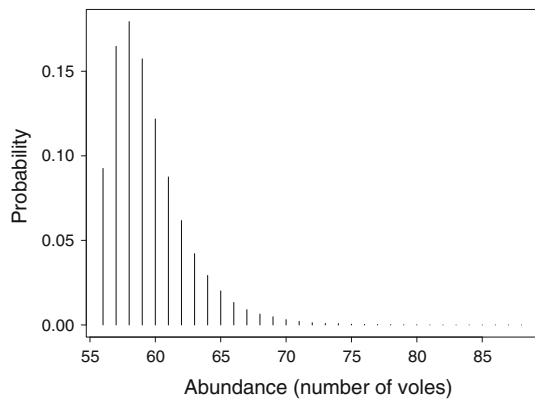
Fig. 3 Estimated posterior distribution of the abundance of meadow voles based on the analysis of capture–recapture data

runs into the same roadblock that was illustrated earlier with occupancy modeling.

Spatial dependence in abundance of sparrows surveyed by double-observer sampling

The data in this example were simulated for an endangered subspecies of seaside sparrows whose geographic range is limited to areas within and adjacent to Everlades National Park, Florida. This subspecies has been studied extensively (Walters et al. 2000), and much is known about its biology. However, the sparrow population has not been sampled using methods that can account for errors in detection of sparrows; therefore, for purposes of illustration I simulated the spatial distribution of sparrow abundances using a parametric model and maps of sparrow habitat (Fig. 4). From these abundances I simulated double-observer surveys at randomly selected locations within the sparrow's known range.

To be more specific, let s denote the spatial coordinates of a location within the sparrow's range. Assume that the abundance $N(s)$ of sparrows at location s has a Poisson distribution with mean $\mu(s) = \exp(u(s))$. Assume further that extra-Poisson variation in $N(s)$ is specified as follows:

$$u(s) \mid \eta(s) \sim \text{Normal}(x'(s)\boldsymbol{\beta} + \eta(s), \, \sigma^2)$$

where the conditional mean depends on a vector of covariate measurements $x(s)$, the effects of these covariates $\boldsymbol{\beta}$, and a parameter $\eta(s)$ that specifies the effect of latent sources of spatial variation in $u(s)$. The values of $\eta(s)$ and $\eta(r)$ for two locations s and r are assumed to be positively correlated. The correlation is assumed to decline exponentially with the Euclidean distance between the locations, that is, $\exp(-||s - r||/\phi)$, where ϕ is a strictly positive range parameter with units of Euclidean distance. The spatial effect parameters are assumed to have a zero-centered, multivariate normal distribution, that is,
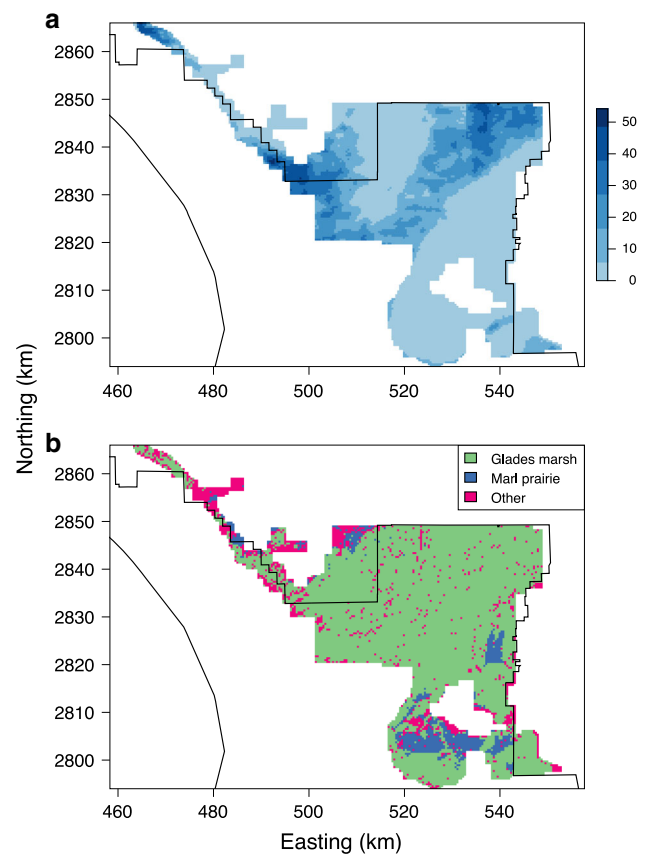

Fig. 4 Spatial distribution of seaside sparrow habitat: **a** average water level (cm) during the breeding season and **b** vegetation categories. *Black line* indicates the boundary of Everglades National Park, Florida

$\boldsymbol{\eta} \sim \text{Normal}(\mathbf{0}, \sigma_\eta^2 \boldsymbol{R})$, where $R_{sr} = \exp(-||s - r||/\phi)$ and σ_η^2 is a scale parameter that specifies the magnitude of the spatial variation.

The assumptions of the hierarchical model of sparrow abundances can be stated succinctly using vector notation:

$$N|\boldsymbol{u} \sim \text{Poisson}(\exp(\boldsymbol{u})) \tag{14}$$

$$\boldsymbol{u}|\boldsymbol{\eta} \sim \text{Normal}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta}, \, \sigma^2 \boldsymbol{I}) \tag{15}$$

$$\boldsymbol{\eta} \sim \text{Normal}(\mathbf{0}, \sigma_\eta^2 \boldsymbol{R}) \tag{16}$$

where $\boldsymbol{I}$ denotes the identity matrix. This model is often used to specify spatial dependence in counts observed at a finite number of locations (Diggle et al. 1998; Wikle 2010). Bayesian methods are essential for fitting this model because an evaluation of the observed-data likelihood requires the latent parameters $\boldsymbol{u}$ to be integrated away. Whereas $\boldsymbol{\eta}$ can be integrated out analytically (see ESM S3), numerical methods for removing $\boldsymbol{u}$ do not provide an accurate approximation of the observed-data likelihood.

The analysis of sparrow counts poses even greater computational problems because errors in detection of
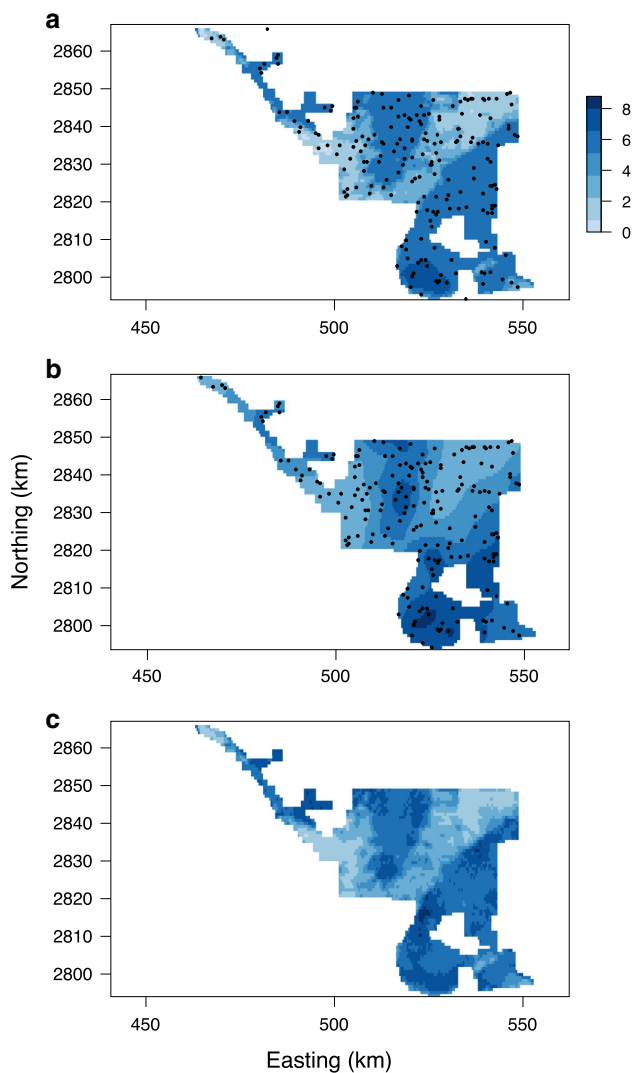
Fig. 5 Spatial distribution of predicted mean abundance of seaside sparrows. Locations of double-observer surveys are superimposed. **a** Predictions based on average water level during the breeding season as a spatial covariate of abundance. **b** Predictions based on spatial dependence alone and no covariates of abundance. **c** True mean abundances for the simulated population of seaside sparrows

sparrows prevents their abundances N from being directly observable. Suppose m locations were selected at random from the sparrow's known range. At each location assume that two observers (A and B) counted sparrows independently for a fixed period of time. If the observers were able to keep track of individual sparrows, their counts can be summarized as follows: $y_A(s_i)$, the number of sparrows at location s_i detected by observer A only; $y_B(s_i)$, the number of sparrows at location s_i detected by observer B only; and $y_{AB}(s_i)$, the number of sparrows at location s_i detected by both observers ($i = 1, \ldots, m$). I used a multinomial model to relate these counts to the unknown abundance $N(s_i)$ of sparrows at location s_i. Let $\mathbf{Y}(s_i) = (Y_A(s_i),$

$Y_B(s_i), Y_{AB}(s_i))'$ denote a random variable for the vector of counts, and assume

$$\mathbf{Y}(s_i) \mid N(s_i) = n(s_i) \sim \text{Multinomial}(n(s_i), \boldsymbol{\pi}(s_i))$$

where $\boldsymbol{\pi}(s_i) = (p_A(s_i)(1 - p_B(s_i)), p_B(s_i)(1 - p_A(s_i)), p_A(s_i) p_B(s_i))'$. These multinomial probabilities depend on the detection probabilities of both observers, which were specified as logit-linear functions of parameters $\boldsymbol{\alpha} = (\alpha_A, \alpha_B, \alpha_w)'$ and a covariate measurement $w(s_i)$ as follows:

$$\text{logit}(p_A(s_i)) = \alpha_A + \alpha_w w(s_i)$$
$$\text{logit}(p_B(s_i)) = \alpha_B + \alpha_w w(s_i)$$

Using the models of sparrow abundance and sampling described above, I simulated double-observer counts at each of $m = 200$ locations within the sparrow's range. Nest success of seaside sparrows is closely tied to water level in the Everglades because the nests built slightly above ground level are vulnerable to seasonal flooding events. To simulate the effects of flooding, I used average water level during the breeding season (Fig. 4a) as a spatially referenced covariate $x(s)$, and I assumed that the effects of this covariate on abundance were negative. In addition, I assumed that latent, spatially correlated sources of variation in sparrow abundances were present and that their magnitude greatly exceeded the effects of non-spatial sources of variation, which were parameterized by σ^2. I used vegetation categories (Fig. 4b) to simulate spatial heterogeneity in detection probabilities. Specifically, I assigned the detection covariate w a value of zero at sample locations with glades marsh (the dominant vegetation category) and a value of one at locations with other types of vegetation. The parameter values used to generate the simulated data are provided in ESM S3.

I used MCMC methods to fit two different Bayesian models to the simulated data. In one model mean sparrow abundance was formulated as a log-linear function of average water level and a spatially correlated random effect. This model is identical to the model that generated the data. In the second model mean sparrow abundance was formulated as a log-linear function of a spatially correlated random effect. This model was fitted to assess how well the effects of water level on sparrow abundance could be predicted by modeling spatial dependence without the benefit of an informative covariate. Details about the MCMC algorithm and prior distributions are described in ESM S3.

The spatial distributions of sparrow abundance predicted by these two models were quite similar (Fig. 5a, b). Moreover, the predictions appeared to capture the true spatial patterns in sparrow abundance (Fig. 5c), which were known, of course, since the abundances were simulated. Modeling spatial dependence alone appeared to

provide a reasonable approximation of the spatial distribution of sparrow abundance. This result was not entirely surprising because the hierarchial model of sparrow abundances (Eqs. 14–16) can be reparameterized as follows:

$$N|u \sim \text{Poisson}(\exp(u)) \tag{17}$$

$$u|\gamma \sim \text{Normal}(X\beta + H\gamma, \sigma^2 I) \tag{18}$$

$$\gamma \sim \text{Normal}(0, \sigma_\eta^2 \Lambda) \tag{19}$$

where H is a $m \times m$ matrix of latent covariates (eigenvectors of R scaled to provide an orthonormal basis), γ denotes the effects of these covariates, and Λ is a diagonal matrix containing the m eigenvalues of R (see ESM S3). In other words, the spatial effects η can equivalently be assumed to have been produced from a set of latent spatially varying covariates H. In the analysis of the sparrow data, H appears to have captured the effects of water level when the measurements of water level were not included in X. This suggests that a spatial analysis of double-observer counts could provide a useful approach for monitoring the population of seaside sparrows living in Everglades National Park, even if important covariates of sparrow abundance have not been measured or identified.

Summary

The results of many ecological studies (empirical and observational) are analyzed using hierarchical models. These models include at least two components: one that specifies the effects of the sampling or observational process (which leads to data), and another that specifies the scientific process of interest in terms of ecologically relevant state variables and parameters. Hierarchical models may include additional layers of complexity, but the separation of observational and ecological processes is fundamental to these models.

In this article I have shown that while some hierarchical models can be fitted using frequentist methods, Bayesian methods of analysis generally provide a more complete solution to the inference problem. For example, ecological problems that require inferences about latent state variables or missing data can be solved easily using Bayesian methods, whereas frequentist solutions to these problems often fail to account fully for errors in estimation when calculating the uncertainty of model-based predictions. For ecological problems that involve intractable likelihood functions (e.g., models of spatial dependence for non-Gaussian responses), Bayesian methods of analysis are preferred simply because they provide the only available solution.

Bayesian methods of analysis are perhaps most compelling because every aspect of inference follows the rules of probability theory. There is a clear path to inference regardless of whether interest is focused on the parameters of a model or on its predictions. In contrast, the Bayesian approach to model assessment and model comparison includes many options. Statisticians have developed a variety of procedures for assessing the adequacy of an individual model and for comparing alternative models, but no clear winner has emerged. There is a growing appreciation that while inferences and predictions are best conducted using Bayesian methods, the evaluation and comparison of models are best accomplished from a frequentist perspective. Therefore, statistical analyses in the 21st century will likely involve a combination Bayesian and frequentist ideas, as noted by Efron (2005) and Little (2006).

References

Bayes T (1763) An essay towards solving a problem in the doctrine of chances. Philos Trans R Soc 53:370–418

Berger JO (2006) The case for objective Bayesian analysis. Bayesian Anal 1:385–402

Box GEP (1980) Sampling and Bayes inference in scientific modelling and robustness (with discussion). J Royal Statist Soc Ser A 143:383–430

Box GEP, Tiao GC (1973) Bayesian inference in statistical analysis. Addison-Wesley, Reading

Brooks S, Gelman A, Jones GL, Meng XL (eds) (2011) Handbook of Markov chain Monte Carlo. Chapman & Hall / CRC, Boca Raton, Florida

Buckland ST, Newman KB, Fernández C, Thomas L, Harwood J (2007) Embedding population dynamics models in inference. Stat Sci 22:44–58

Clark JS (2007) Models for ecological data: an introduction. Princeton University Press, Princeton

Diggle PJ, Tawn JA, Moyeed RA (1998) Model-based geostatistics (with discussion). Journal of the Royal Statistical Society. C (Applied Statistics) 47:299–350

Dorazio RM, Taylor Rodríguez D (2012) A Gibbs sampler for Bayesian analysis of site-occupancy data. Methods Ecol Evol 3:1093–1098

Draper D (1996) Utility, sensitivity analysis, and cross-validation in Bayesian model-checking. Stat Sinica 6:760–767

Efron B (2005) Bayesians, frequentists, and scientists. J Am Stat Assoc 100:1–5

Flegal JM, Jones GL (2010) Batch means and spectral variance estimators in Markov chain Monte Carlo. Ann Stat 38:1034–1070

Flegal JM, Jones GL (2011) Implementing MCMC: estimating with confidence. In: Brooks S, Gelman A, Jones GL, Meng XL (eds) Handbook of Markov chain Monte Carlo, Chapman & Hall / CRC. Florida, Boca Raton, pp 175–197

Gelfand AE, Smith AFM (1990) Sampling-based approaches to calculating marginal densities. J Am Stat Assoc 85:398–409

Gelman A (2011) Induction and deduction in Bayesian data analysis. Rational Markets Morals 2:67–78

Gelman A, Shirley K (2011) Inference from simulations and monitoring convergence. In: Brooks S, Gelman A, Jones GL, Meng XL (eds) Handbook of Markov chain Monte Carlo, Chapman & Hall / CRC. Florida, Boca Raton, pp 163–174

Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Stat Sinica 6:733–807

Geyer CJ (2011) Introduction to Markov chain Monte Carlo. In: Brooks S, Gelman A, Jones GL, Meng XL (eds) Handbook of Markov chain Monte Carlo, Chapman & Hall / CRC. Florida, Boca Raton, pp 3–48

Hobert JP (2011) The data augmentation algorithm: theory and methodology. In: Brooks S, Gelman A, Jones GL, Meng XL (eds) Handbook of Markov chain Monte Carlo. Chapman & Hall / CRC, Boca Raton, Florida, pp 253–293

Hooten MB, Hobbs NT (2015) A guide to Bayesian model selection for ecologists. Ecol Monogr 85:3–28

Hooten MB, Wikle CK (2007) A hierarchical bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasion collared-dove. Environ Ecol Stat 15:59–70

Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, New York

Johnson DS, Conn PB, Hooten MB, Ray JC, Pond BA (2013) Spatial occupancy models for large data sets. Ecology 94:801–808

Kéry M, Schaub M (2012) Bayesian population analysis using WinBUGS. Academic Press, Waltham

Kéry M, Dorazio RM, Soldaat L, van Strien A, Zuiderwijk A, Royle JA (2009) Trend estimation in populations with imperfect detection. J Appl Ecol 46:1163–1172

King R, Morgan BJT, Gimenez O, Brooks SP (2010) Bayesian analysis for population ecology. Chapman and Hall/CRC, Boca Raton

Laird NM, Louis TA (1987) Empirical Bayes confidence intervals based on bootstrap samples (with discussion). J Am Stat Assoc 82:739–757

Laird NM, Ware JH (1982) Random-effects models for longitudinal data. Biometrics 38:963–974

Laplace PS (1774a) Memoir on the probability of the causes of events. Stat Sci 1:364–378 (**English translation of the French original by S. M. Stigler in 1986**)

Laplace PS (1774b) A philosophical essay on probabilities. John Wiley & Sons, New York (**English translation of the French original by F. W. Truscott and F. L. Emory in 1902**)

Lindley DV (2000) The philosophy of statistics (with discussion). Statist 49:293–337

Link WA, Barker RJ (2010) Bayesian inference. Academic Press, Amsterdam

Little R (2011) Calibrated Bayes, for statistics in general, and missing data in particular. Stat Sci 26:162–174

Little RJ (2006) Calibrated Bayes: a Bayes/frequentist roadmap. Am Stat 60:213–223

Little RJA, Rubin DB (2002) Statistical analysis with missing data, 2nd edn. Wiley, Hoboken

MacKenzie DI, Nichols JD, Lachman GB, Droege S, Royle JA, Langtimm CA (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology 83:2248–2255

MacKenzie DI, Nichols JD, Royle JA, Pollock KH, Bailey LL, Hines JE (2006) Occupancy estimation and modeling. Elsevier, Amsterdam

McCarthy M (2007) Bayesian methods for ecology. Cambridge University Press, Cambridge

McCarthy MA, Masters P (2005) Profiting from prior information in Bayesian analyses of ecological data. J Appl Ecol 42:1012–1019

Morris CN (1983) Parametric empirical Bayes inference: theory and applications (with discussion). J Am Stat Assoc 78:47–65

Morris WK, Vesk PA, McCarthy MA (2013) Profiting from prior studies: analysing mortality using Bayesian models with informative priors. Basic Appl Ecol 14:81–89

Nichols JD, Pollock KH, Hines JE (1984) The use of a robust capture-recapture design in small mammal population studies: A field example with *Microtus pennsylvanicus*. Acta Theriol 29:357–365

Parent E, Rivot E (2013) Introduction to hierarchical Bayesian modeling for ecological data. Chapman and Hall/CRC, Boca Raton

Robert C, Casella G (2011) A short history of MCMC: subjective recollections from incomplete data. In: Brooks S, Gelman A, Jones GL, Meng XL (eds) Handbook of Markov chain Monte Carlo. Chapman & Hall / CRC, Boca Raton, Florida, pp 49–66

Royle JA, Dorazio RM (2006) Hierarchical models of animal abundance and occurrence. J Agr Biol Environ Stat 11:249–263

Royle JA, Dorazio RM (2008) Hierarchical modeling and inference in ecology. Academic Press, Amsterdam

Royle JA, Dorazio RM (2012) Parameter-expanded data augmentation for Bayesian analysis of capture-recapture models. J Ornithol 152:S521–S537

Royle JA, Wikle CK (2005) Efficient statistical mapping of avian count data. Environ Ecol Stat 12:225–243

Rubin DB (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. Ann Stat 12:1151–1172

Schaub M, Abadi F (2011) Integrated population models: a novel analysis framework for deeper insights into population dynamics. J Ornithol 152:S227–S237

Tanner MA (1996) Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions, 3rd edn. Springer-Verlag, New York

Tyre AJ, Tenhumberg B, Field SA, Niejalke D, Parris K, Possingham HP (2003) Improving precision and reducing bias in biological surveys: estimating false-negative error rates. Ecol Appl 13:1790–1801

Walters JR, Beissinger SR, Fitzpatrick JW, Greenberg R, Nichols JD, Pulliam HR, Winkler DW (2000) The AOU conservation committee review of the biology, status, and management of Cape Sable seaside sparrows: final report. Auk 117:1093–1115

Wikle CK (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology 84:1382–1394

Wikle CK (2010) Hierarchical modeling with spatial data. In: Gelfand AE, Diggle PJ, Fuentes M, Guttorp P (eds) Handbook of spatial statistics. Chapman & Hall / CRC, Boca Raton, Florida, pp 89–106

Williams BK, Nichols JD, Conroy MJ (2002) Analysis and management of animal populations. Academic Press, San Diego, California