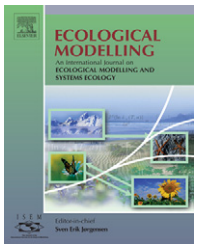
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Review

A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools

Eric J. Ward

School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195-5020, USA

ARTICLE INFO

Article history:
Received 9 January 2007
Received in revised form
24 September 2007
Accepted 11 October 2007
Published on line 26 November 2007

Keywords:
Model selection
AIC
BIC
DIC
Bayes factor

ABSTRACT

Many tools have become available for biologists for evaluating competing ecological models – models may be judged based on the fit to data alone (e.g. likelihood), or more formal statistical criteria may be used. Because of the implied assumptions of each tool, model selection criteria should be chosen *a priori* for the problem at hand, – a model that is considered 'good' in its explanatory power may not be the best choice for a problem that requires prediction. In this paper, I review the behavior and assumptions of the four most commonly used statistical criteria (Akaike's Information Criterion, AIC; Schwarz or Bayesian Information Criterion, BIC; Deviance Information Criterion, DIC; Bayes factors). Second, I illustrate differences in these model selection tools by applying the four criteria to thousands of simulated abundance trajectories. With the simulation model known, I examine whether each of the criteria are useful in selecting models to evaluate simple questions, such as whether time series support evidence of density dependent population growth. Across simulations, the maximum likelihood criteria consistently favored simpler population models when compared to Bayesian criteria. Among the Bayesian criteria, the Bayes factor favored the correct simulation model more frequently than the Deviance Information Criterion. There was considerable uncertainty in the ability of the Bayes factor to discriminate between models, this tool selected the simulation model slightly more frequently than other approaches.

Contents

E-mail address: warde@u.washington.edu.
0304-3800/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2007.10.030

1. Introduction

The purpose of mathematical modeling in biology is to provide a mechanism for evaluating scientific and statistical hypotheses using observational data (Lewin-Koh et al., 2004). Models may serve as both explanatory and predictive tools; if a model appears to realistically track variation in abundance through time, it may be used to predict future population sizes, set harvest quotas, or assess extinction risk. Population models are often complex, involving age- and sex-structured dynamics, predator–prey interactions, effects of both time and space, and potentially time lags. In developing statistical models for inference, researchers must decide *a priori* which, if any, of these factors may be important in the population being modeled. A second decision researchers must make is how model performance should be assessed. Model performance is defined broadly as the ability of a model to meet some specified objective (e.g. how well a model explains observed data, makes predictions, or minimizes risk). Statistical tools allow models to be compared relative to one another, but each criterion differs in what is considered a 'good' model. Differences between model selection criteria are rooted both in the philosophy of science, as well as statistics (Maurer, 2004). Criteria may view good models to be those that that minimize Type I and Type II error rates (Weiss, 1997) or those that involve the simplest explanations possible, minimizing the tradeoff between bias and variance (the principle of parsimony or Occam's razor; Forster, 2000; Burnham and Anderson, 2002). Because these tools are so widely used in the biological sciences (Johnson and Omland, 2004), it is important to understand the differences in the assumptions and performance of each criterion.

Over the last 20 years, the most widely used model selection tool in ecology has been Akaike's Information Criterion (AIC; Akaike, 1973). AIC began to see widespread use by biologists in the 1990s after several papers illustrated applications to capture–recapture analyses (Anderson et al., 1994; Burnham et al., 1995). AIC is computed as $\mathrm{AIC} = D(\hat{\theta}_{\mathrm{MLE}}) + 2K$, where the function $D(\cdot)$ represents the deviance function (twice the negative log-likelihood), $\hat{\theta}_{\mathrm{MLE}}$ the vector of maximum likelihood parameter estimates (MLEs) and K represents the number of model parameters. The second term in the AIC calculation has been interpreted as a measure of model order or complexity (Burnham and Anderson, 2002). A popular variant of AIC is the small sample AICc, where the complexity term $2K$ is replaced by $(2Kn/n - k - 1)$ (Hurvich and Tsai, 1995), where n represents the sample size. With small sample size AIC favors models with fewer parameters compared to AICc, and as the sample size becomes relatively large, the behavior of AICc and AIC converges (Fig. 1). AIC has inspired several additional variants to account for overdispersion (QAIC), correlations between model parameters (CAICF), and situations where models might not be good approximations to truth (TIC; Burnham and Anderson, 2002). These criteria are beyond the scope of this review, as there are not as commonly used as AIC or AICc.

A second important model selection criterion that is frequently compared to AIC is the Bayesian Information Criterion (BIC or Schwarz Information Criterion; Schwarz, 1978). While the two criteria may appear similar, BIC has a completely

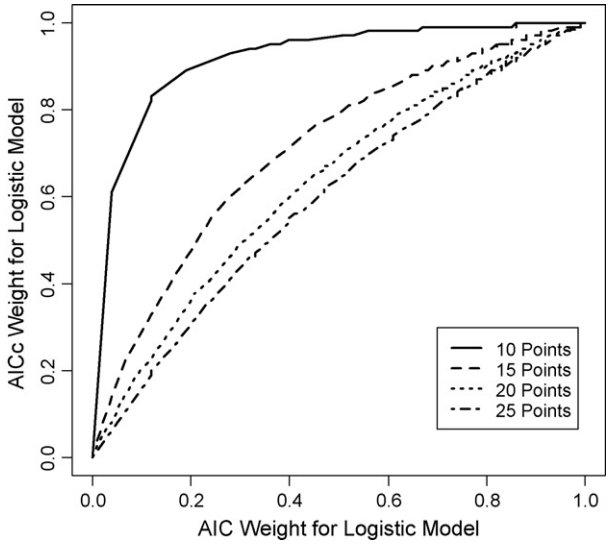


Fig. 1 – The asymptotic relationship between AIC weights and the small sample AICc weights as a function of the number of data points. Data were simulated from a theta-logistic process, and two models were considered: logistic and theta-logistic. When the sample size is large (~30) the methods perform similarly, but when less data are available, AIC tends to favor more complex models.

independent derivation (Burnham and Anderson, 2004). Like AIC, the calculation of BIC involves two terms: $\mathrm{BIC} = D(\hat{\theta}_{\mathrm{MLE}}) + K\ln(n)$. The first term is identical to that used in the calculation of AIC (representing the model fit to data), however the complexity term is slightly different, being a function of both the number of parameters (K) and sample size (n). When n is <7.4, BIC assigns more weight to complex models than does AIC, but as the sample size increases, BIC assigns more weight to simpler models when compared to AIC (Fig. 2; Raftery, 1995; Forster, 2000; Burnham and Anderson, 2002). Generally, BIC is interpreted as a rough approximation to the logarithm of the Bayes factor (Kass and Raftery, 1995). BIC exists in somewhat of a grey area—while the computation of the criterion is not Bayesian, there are some situations where the BIC model weights may be interpreted as posterior model probabilities (Raftery, 1999; Link and Barker, 2005). Although BIC does not require the explicit specification of a prior distribution on model parameters, the implicit prior assumed by BIC is a multivariate normal distribution centered on the MLE (with a covariance matrix equal to the inverse of the Fisher information matrix). Another name for this prior is the unit information prior, because has approximately the same contribution as one additional data point (Raftery, 1995). BIC assumes equal priors on each model, so that if M models are considered, the prior on each is 1/M. In cases where the implicit prior is similar to the prior used in the Bayes factor calculation, BIC can be used to calculate posterior probabilities for each model being considered (posterior probabilities could be computed using normalized BIC weights; Burnham and Anderson, 2002). BIC is known to be poorly suited for some problems (Stone, 1979), particularly when the multivariate normal distribution is not a reasonable prior and few

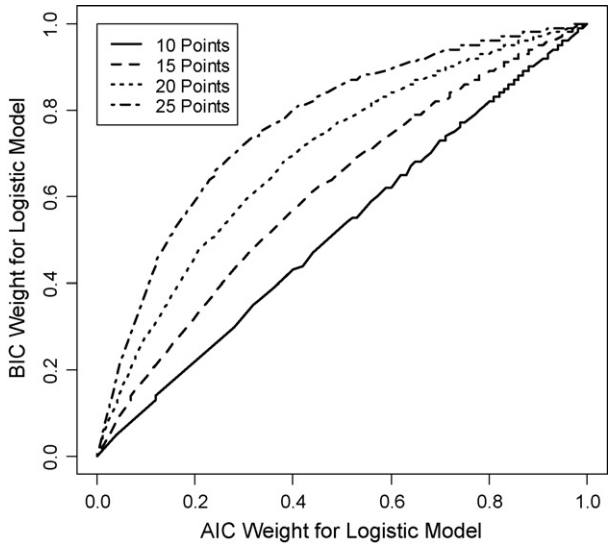


Fig. 2 – The relationship between AIC weights and BIC weights as a function of the number of data points. Data was simulated from a theta-logistic model, and two competing models were considered as estimation models: logistic and theta-logistic. When the sample size is small the methods perform similarly, but as more data is collected, BIC favors the model with fewer parameters.

data points exist (skewed parameters may include variance parameters or ratios, non-linear density dependence or scaling parameters). It should be noted that AIC weights have recently been interpreted in a Bayesian setting, using the logic of BIC (Burnham and Anderson 2004).

One downside to the conventional forms of AIC and BIC is that they are not well suited for complicated ecological models that include hidden states with non-Gaussian errors, or hierarchical parameters (Burnham and Anderson, 2002; Vaida and Blanchard, 2005). Spiegelhalter et al. (2002) proposed using the Deviance Information Criterion (DIC) as a Bayesian equivalent of AIC for hierarchical models. Like BIC and AIC, the calculation of DIC involves a measure of model fit, and a measure of model complexity: $\mathrm{DIC} = D(\hat{\theta}) + 2p_D$, where $D(\hat{\theta})$ represents the deviance evaluated at some point estimate of the joint posterior, and p_D represents the effective number of model parameters. The DIC equation closely resembles the calculation of AIC, and in the absence of informative priors on model parameters, the two criteria are expected to be equal (Ellison, 2004). The term p_D can also be expressed as the deviance function evaluated at the expected posterior values of the parameters subtracted from the mean deviance across all possible parameter vector values, $p_D = \overline{D(\theta)} - D(\hat{\theta})$. The mean deviance and mean parameter values were originally proposed by Spiegelhalter et al. (2002) as plug-in estimates (and are currently implemented as the default option in WinBugs; Spiegelhalter et al., 2003), however the posterior median or mode may also be appropriate (e.g. Celeux et al., 2006). Proponents of DIC argue that it is a Bayesian equivalent of AIC (Spiegelhalter et al., 2002), however the similarity between the two methods is still being investigated.

A final model selection criterion that has seen increased use by ecologists in the last decade is the Bayes factor (Jeffreys, 1935; Kass, 1993; Kass and Raftery, 1995). The Bayes factor for two models (M_1 and M_2) is the ratio of posterior odds to prior odds, $\mathrm{BF} = (P(x|M_1)/P(x|M_2)) = (P(M_1|x)/P(M_2|x)/P(M_1)/P(M_2))$ (Good, 1958). As the Bayes factor updates the prior distribution upon observing the data, it is essentially a measure of how much a researcher learns by observing data (Bernardo and Smith, 2000). If more than two models are being compared, and all models are given equal prior weight, the Bayes factor in favor of model j becomes $\mathrm{BF} = P(\underline{x}|M_j)/\sum_i P(\underline{x}|M_i)$. There are several approaches for approximating the marginal likelihood of the data, $P(\underline{x}|M_j)$, all of which are computationally intensive. For this analysis, I adopted the approach proposed by Gelfand and Dey (1994): $P(\underline{x}|M_j) = [(1/n)\sum_{i=1}^{n} h(\theta_i)/P(\theta_i|\underline{x})]^{-1}$, where n is the number of Markov Chain Monte Carlo (MCMC) samples, $P(\theta_i|x)$ is the posterior probability of parameter vector i for model j, and $h(\theta_i)$ represents an importance function evaluated at parameter vector i (a multivariate normal density centered at the posterior mode, with a covariance matrix estimated from the MCMC chain). Aside from computational challenges, the primary problem with implementing Bayes factors is the specification of prior distributions. Bayes factors are known to be unstable and sensitive to the choice of priors (Kadane and Lazar, 2004), and only proper priors may be considered (valid probability distributions that integrate to 1.0; Kass and Raftery, 1995). If no information about model parameters is known, a default option is that non-informative prior distributions such as Jeffrey's prior may be constructed (Kass and Wasserman, 1996). What makes the Bayes factor unique relative to the other criteria discussed here is that it does not explicitly include a term that quantifies model complexity—this still exists, however, because overly complex models are penalized in the marginal likelihood calculation (e.g. Dawid's discussion in Spiegelhalter et al., 2002).

There are several important differences between AIC, BIC, DIC, and Bayes factors (summarized in Table 1). All maximum likelihood and Bayesian criteria are alike in that model selection is linked to parameter estimation. In the calculation of AIC and BIC, parameter estimation is done by maximizing the likelihood in an attempt to find the single best point estimate. Parameters for DIC and Bayes factors are estimated using Bayesian methods, which integrate rather than maximize over the parameter space (Hobbs and Hilborn, 2006). While both DIC and Bayes factors incorporate parameter uncertainty and correlation in the sampling of the joint posterior distribution, one criticism of AIC and BIC is that neither consider parameter uncertainty in their calculations. This is not a maximum likelihood versus Bayesian issue; several maximum likelihood criteria, including the Information Complexity Criterion (ICOMP, Table 1; Bozdogan, 2000), do include parameter correlation and uncertainty, however these methods have seen little use in biology. In principle, the calculation of ICOMP is similar to AIC, $\mathrm{ICOMP} = D(\hat{\underline{\theta}}_{\mathrm{MLE}}) + \ln(n)C(\Sigma)$, where $C(\cdot)$ represents a complexity function based on the parameter variance–covariance matrix (Σ).

From a philosophical point of view, there are also differences in the objectives of these model selection criteria. Myung (2000) divided model selection tools into two groups: generalization-based criteria, and explanation-based criteria.

Table 1 – Summary of various properties of the model selection criteria used in this analysis

Property	Model selection criteria				
	AIC	BIC	DIC	BF	ICOMP
Estimation	MLE	MLE	Bayesian	Bayesian	MLE
Model performance	Best	Best	Average	Integrated	Best
Complex models	N	N	Y	Y	N
Non-nested models	Y	Y	Y	Y	Y
Prior	N	Implicit	Explicit	Explicit	N
Parameter uncertainty	N	N	Y	Y	Y
Complexity term	Explicit	Explicit	Explicit	Implicit	Explicit
Complex likelihoods	N	N	N	Y	N

In addition to AIC, BIC, DIC, and Bayes factors (BF), the ICOMP criterion has been included to illustrate an example of a maximum likelihood criterion that incorporates relationships between parameters. The property "complex models" is meant to include state-space models, or models with hierarchical parameters; "parameter uncertainty" is meant to indicate which criteria integrate over parameter space or include variance–covariance matrices in calculations of model complexity; "complex likelihoods" is meant to designate multimodal densities, likelihoods that are composed of mixture distributions, and missing data problems.

Generalization-based criteria (AIC and ICOMP) seek to find the best model that fits both current data in addition to hypothetical future data that may be observed from the same process that generated the original sample. Explanation-based criteria (BIC and Bayes factors) are concerned with identifying the process from which the data arose, and are not influenced by hypothetical future observations. Many biologists, particularly those working with natural resource management, may be more interested in predicting future quantities, rather than explanation (e.g. Fried and Hilborn, 1988). Prediction-based Bayesian model selection tools exist for these types of problems (Madigan and Raftery, 1995; Gelman et al., 1995; Bernardo and Smith, 2000), but will not be discussed here because the current analysis is only concerned with explanation.

A final difference between model selection tools is that other than Bayes factors, none of the criteria presented were ever intended to be used as decision tools, but they are still used to provide advice to decision making advice to managers. An argument could be made that BIC or AIC are compatible with decision making because they may be viewed as an approximation to the Bayes factor (Raftery, 1999; Link and Barker, 2005). Proponents of both AIC and DIC strongly encourage the use of these tools for making inferences, not decisions (Burnham and Anderson, 2002; Spiegelhalter et al., 2002). Richards (2005) illustrated potential pitfalls of using AIC weights, and other authors have noted statistical flaws in such approaches (see discussion of Forster and Sober, 2004). Another reason not to use these tools to make decisions about 'significance' is that there may be implicit costs that are not obvious. For example, assume that two nested modes are to be compared. One approach for evaluating these models might be to use a likelihood ratio test (LRT) as a decision rule. The LRT statistic rejects model M_1 when $L(x|M_1)/L(x|M_2) \leq C$, C representing the critical region of the test (this can also be expressed as the ratio of the costs of a Type I error to a Type II error; Bernardo and Smith, 2000). The significance level of the LRT (α, the probability of rejecting M_1 when M_1 is actually true) is usually set *a priori* at 0.05. Suppose that instead of using a LRT, however, a researcher uses AIC to make a decision about competing models. The significance level of AIC is not transparent, because it varies as a function of the complexity between the models considered (with 1 degree of freedom, the significance level is fixed at 0.157 rather than 0.05; Forster, 2000; Burnham and Anderson, 2002; Kuha, 2004). The implication of this difference is that AIC will give more weight to complex models relative to LRTs (Forster, 2000). As a third option, suppose the researcher uses BIC as a decision tool—the significance level of BIC becomes even more complex, because it is a function of both the difference in model complexity and sample size. With small sample sizes (~7.4), BIC gives similar results to AIC, but the significance level decreases rapidly with increasing sample size or an increasing difference in complexity (BIC eventually favoring simpler models than would be chosen by LRTs). If the researcher in this scenario chooses a model selection tool arbitrarily, simply based on which is easiest to calculate, he will be ignorant of the implied Type I and Type II errors.

The purpose of this analysis is to address the performance of four model selection criteria (AICc, BIC, DIC and Bayes factors) applied to simulated population abundance data. The majority of previous studies evaluating the performance of model selection criteria may not be completely applicable to ecological data, either because of the criteria compared, or the sample sizes involved. For example, Myung (2000) conducted numerous simulations for polynomial models, but did not consider the small sample variant of AIC (AICc), which is more applicable to ecological problems (Burnham and Anderson, 2002). Other simulations have focused on sample sizes of more than 100 (Kuha, 2004) which is often unrealistic for the natural sciences—many biologists work with sample sizes an order of magnitude smaller. Research with biological models utilizing few samples has been explored in a maximum likelihood framework (Shono, 2000), however few studies to date have bridged the gap between maximum likelihood and Bayesian model selection criteria (e.g. A'mar, 2004; Ellison, 2004).

All four criteria will be evaluated across multiple scenarios, and will be evaluated in their ability to (1) detect density dependent processes, (2) detect non-linear density dependence and (3) detect Allee effects (or inverse density dependence; Courchamp et al., 1999). In Monte Carlo comparisons of model selection performance, the true model is almost always considered among the candidate models (McQuarrie and Tsai, 1998; Forster, 2000; Myung, 2000). While the goal of Monte Carlo comparisons is to estimate the fre-

quency of selecting the true model, the goal of model selection applied to real data is to find a model that best approximates truth. If the simulation-based rules were applied to inference in the real world, all criteria would be wrong 100% of the time. Despite these issues, understanding the performance of model selection over many data sets is extremely valuable—not only can the performance of alternative criteria be compared, but it is also possible to examine the influence of different kinds of error on model selection.

2. Methods

In this analysis, I considered the ability of four model selection tools to select among four discrete time population models. The first four population models represent single stage models: (1) the geometric model $N_{t+1} = N_t(1+r)$; (2) logistic model $N_{t+1} = N_t + rN_t(1 - N_t/K)$; (3) theta-logistic model $N_{t+1} = N_t + rN_t(1 - (N_t/K)^{\phi})$ (Gilpin and Ayala, 1973); and (4) model with decreased growth rate at low density, $N_{t+1} = N_t + rN_t(N_t - a)(K - N_t)/K^2)$ (Lewis and Kareiva, 1993). In these models, the parameter r represents the growth rate, K represents the carrying capacity, ϕ represents the strength of density dependence, and a represents the critical depensation point, below which per capita growth becomes negative.

In turn, each of the four models was considered as the simulation model and 2500 time series were generated. Time series length was allowed to be random ($n = 10$, 15, 20 and 25). To make the data as realistic as possible, I included random lognormal observation error ($CV_{obs} = 0.1$, 0.3 and 0.5), and random lognormal process error ($CV_{pro} = 0.0$, 0.1 and 0.2). Each data set was assumed to have the same carrying capacity ($K = 1.0E5$), and each was initialized from some random fraction of K ($N_0 = 0.1$, 0.3, 0.5, 0.7 and 0.9). When generating logistic data, I varied the growth rate over the range 0.1—1.1 (in steps of 0.2) and the density dependence parameter ϕ over the range 0.7—1.3 (in steps of 0.2). For the Allee model, the Allee threshold (a) was also treated as a random fraction of K ($a = 0.1$, 0.3, 0.5, 0.7 and 0.9), subject to the constraint that $N_0 > a$.

For each estimation model, I assumed that observation error was the only type of error present, and it was assumed to be log-normally distributed (multiplicative). Prior distributions were assigned based on similar models in the literature (e.g. Punt and Hilborn, 1997; McAllister and Kirkwood, 1998). For the single-stage models, uniform priors were placed on K (500, 1.0E6), r (−0.5, 2.5), and the critical Allee point as a fraction of K (0, 0.9). These priors were chosen because they are known to not contain much information with respect to model parameters. The prior on the initial population size relative to K was assumed to be uniform (0, 1.0) and the CV of the observation error was assumed to be log-uniform (0, 2). For the theta-logistic model, the prior on ϕ was chosen to be ~lognormal (−0.5, 1). This prior has an expected value of 1.0 (linear density dependence), and is favored over the log-uniform prior because the log-uniform prior assigns more weight to small values of ϕ.

After each of the new data sets was generated, the four models were in turn treated as the estimation model and fit to the simulated data set. Maximum likelihood parameter estimation was done in AD Model Builder (Fournier, 1996), and the output from each estimation procedure was used to calculate AICc and BIC. AD Model Builder was also used to do Metropolis-Hastings MCMC sampling (Gelman et al., 1995). A small number of data sets were used to determine whether MCMC chains converged (Best et al., 1995). Results from this analysis were used to choose a burn-in length of 5.0E5 samples, followed by a chain length of 1.0E6 samples. Every 200th sample was retained from the MCMC chain, resulting in 5000 posterior vectors for each estimation model. Following MCMC sampling, the DIC value and marginal likelihood for the Bayes factor were calculated.

For each comparison between estimation models, the normalized model weights for AICc, BIC and DIC were computed following the general method for computing AIC weights (Burnham and Anderson, 2002). After the Δ values are computed for each model relative to the model with the lowest score, the normalized weight for the ith model can be calculated as $w_i = \exp(-0.5\Delta_i)/\sum_j \exp(-0.5\Delta_j)$. Although some studies have presented these model weights as Bayesian posterior probabilities (e.g. Wintle et al., 2003), and others have suggested that AIC may be justified in a Bayesian framework (Burnham and Anderson, 2004), I treated these quantities as weights. After computing the posterior model probability via the Bayes factor for each candidate model, the posterior probabilities were normalized to allow comparison to the model weights.

3. Results

In all comparisons, the true simulation model was not considered as an estimation model (simulation models included both process error and observation error). The first comparison I examined was the ability of each model selection criteria to detect density dependence, and the rate with which each was subject to Type II errors. In this comparison, Type II error represents instances where more weight is assigned to an estimation model that includes density dependence when the simulation model is geometric. Across all simulated data sets, I computed the frequency with which each criterion assigned more than 0.5 of the model weight (or posterior probability) to models that include density dependence. Using a cutoff value of 0.5 is somewhat arbitrary, however it was chosen because it represents the point at which one model is favored over all others. The Bayes factor and DIC were able to detect density dependence in almost all data sets (99.7, 99.3%, respectively), while AICc and BIC did slightly worse (80.3, 85.3%, respectively). One potential reason for the difference between AICc and BIC may be the assumption that each makes about whether the simulation model is included among the candidate models. McQuarrie and Tsai (1998) observed that BIC outperformed AICc when the simulation model was among those considered. Even though the growth rates in data generated from the geometric model were small (0.01–0.1), both the maximum likelihood and Bayesian criteria had low Type II error rates (all <4%).

As a second comparison, I examined the ability of each model selection criteria to detect non-linear density dependence. For simplicity, this involved only the theta-logistic and logistic population models. Before comparing the perfor-

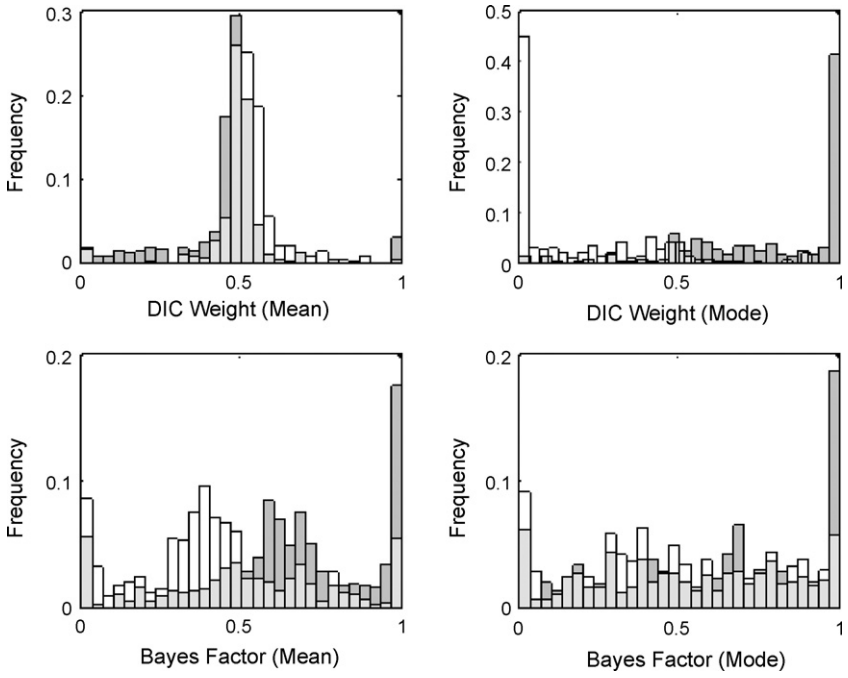


Fig. 3 – Distribution of DIC weights and posterior model probabilities (Bayes factor) evaluated at both the mean and mode for data generated from a logistic model (dark grey bars) and theta-logistic model (white bars). As the dark grey histogram is transparent, the light grey region represents the overlapping area. Each summary statistic (mean, mode) is used in both terms of the DIC calculation ($\overline{D(\theta)}$, $D(\hat{\theta})$), and each is used to center the importance function in the calculation of the multivariate normal importance function ($h(\theta)$). For both histograms, values close to 1.0 indicate that the true model is favored, while values close to 0.0 indicate that the wrong model is favored. A model selection criterion with low Type I and Type II error rates will have both densities concentrated on the right side of the plot.

mance of the maximum likelihood and Bayesian criteria, I examined the inferential consequences of choosing a particular summary statistic (mean or mode) in the calculation of DIC and the Bayes factor. For each summary statistic, I calculated both components of the DIC calculation ($\overline{D(\theta)}$, $D(\hat{\theta})$) as a function of that statistic (Spiegelhalter et al., 2002). Two versions of the Bayes factor were also calculated—one with the multivariate normal importance function centered on the posterior mode, and the other with the importance function centered on the posterior mean. When the mean likelihood and parameter means were used in the calculations of DIC, more weight was given to the complex model (Fig. 3). If the mode of the deviance is used in place of the mean deviance, and parameter modes are used instead of the parameter means, DIC tends to favor the simpler model. One problem with using the posterior mode in the calculation of $\overline{D(\theta)}$, however, is that more than 80% of the time series resulted in negative estimates of model complexity (p_D). Using the parameter means in the importance function of the Bayes factor tends to favor the more complex model, while there appears to be much more uncertainty in which model is favored when the posterior mode is used.

To understand why DIC may favor complex models, I examined the distribution of bias ($\hat{\theta} - \theta$) for each parameter in the theta-logistic model—the largest bias was observed in the density dependence parameter, ϕ. The posterior of ϕ is often skewed, with the posterior mean greater than the mode (Fig. 4). Although the prior on ϕ is centered at 1.0 (the point

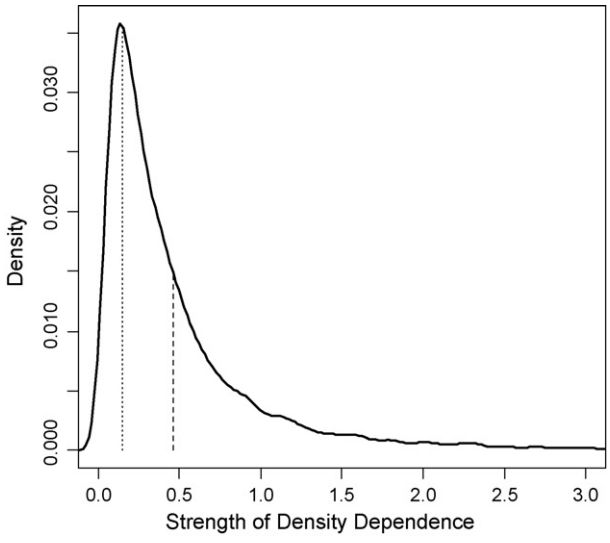


Fig. 4 – Posterior distribution of the density dependence parameter ϕ, estimated for one of the 2500 theta-logistic data sets in this analysis (the true value of ϕ is 1.3). In this example, the mean (0.463, dashed line) is much greater than the median (0.295) and the mode (0.147, dotted line).

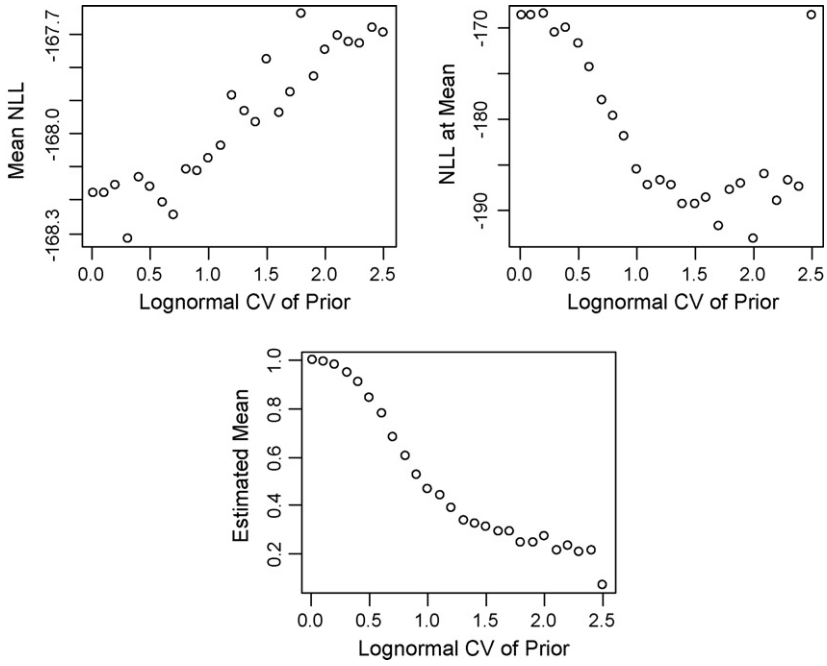


Fig. 5 – Sensitivity of the prior for the density dependence parameter θ on the components of the DIC criterion. The prior on ϕ is assumed to be log-normally distributed, with an expected value of 1.0. As the prior standard deviation (CV) increases, the expected value of the posterior of ϕ decreases, increasing the parameter bias (the true parameter value is 1.3) and a large difference in the negative log-likelihood (NLL) at the posterior means.

where density dependence is linear), increasing the standard deviation has a large effect on DIC, decreasing the expected value of the posterior of ϕ, and influencing both the mean deviance and deviance at the mean (Fig. 5). Increasing the prior standard deviation also has impacts on stability because of the negative correlation between ϕ and the logistic growth rate, r (large values of ϕ only allow for small values of r). It is possible that the lognormal prior may be a poor choice for ϕ, or that the assumption about its location may be wrong (it may be better for the prior to have an expectation that is less than or greater than 1.0).

When the maximum likelihood model selection tools are compared to their Bayesian counterparts in their ability to detect non-linear density dependence, it appears that both AICc and BIC tend to favor the simpler model, even when data are generated from the more complicated model (AICc favors the simpler model slightly more; Fig. 6). Although there is considerable uncertainty associated with the posterior probabilities, the Bayes factor had a slightly higher rate of successfully identifying the simulation model. The behavior of DIC is similar to that of AIC in that it tends to favor the simpler model, favoring linear density dependence over non-linear density dependence. A third comparison investigated here was the ability of each criteria to detect negative growth rates at low densities. For this comparison, I examined the performance of the logistic, theta-logistic, and Allee models on data generated from those three processes. The performance of AIC, BIC, and DIC was similar in that when the logistic and theta-logistic models are considered together against the Allee model, all three criteria tend to give less weight to the Allee model (Fig. 7). This same result held when the theta-logistic model was compared alone against the Allee model, even though both models have the same number of parameters ($n = 5$). The posterior model probabilities estimated by the Bayes factor leads to different results (Fig. 7). First, there is less uncertainty in whether or not Allee dynamics are present. Second, the Bayes factor gives slightly more weight to the Allee model, which is the opposite of the other criteria.

4. Discussion

Because each model selection tool has different intrinsic assumptions and behavior, it is crucial to understand the differences between each. Before using any of the model selection criteria discussed in this paper, ecologists should consider the following questions. First, is the purpose of the analysis to make predictions, or to decide which model best represents reality (Ghosh and Samanta, 2001)? While AICc may have better predictive ability than BIC, order-consistent criteria (Taper, 2004) or bootstrap approaches may have more desirable properties. Secondly, does the model contain hierarchical parameters, random effects, or highly correlated parameters? Expressing complexity for these models as a single number is often difficult, and few maximum likelihood criteria are well understood (e.g. Vaida and Blanchard (2005) proposed conditional AIC; others have proposed hybrid approaches between AIC and DIC). An alternative expression of model order is the parameter correlation matrix (Bozdogan, 2000). Third, is anything known about biases in model parameters? While this might be impractical in reality, parameter bias has a large impact on model selection (Forster, 2000). Fourth, if using

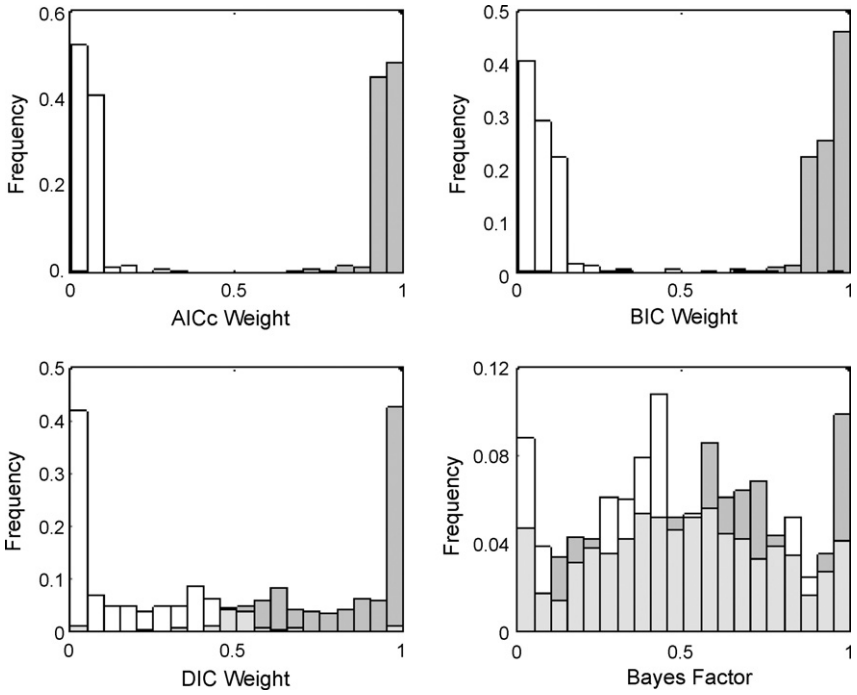


Fig. 6 – Model weights for evaluating the evidence of non-linear density dependence. In each case the model weight (or in the case of the Bayes factor, the posterior probability) represents the frequency of selecting the correct model that generated the data. 2500 data sets have been generated from a logistic model (grey bars) and 2500 data sets have been generated from a theta-logistic model (white bars). For both histograms, values close to 1.0 indicate that the true model is favored, while values close to 0.0 indicate that the wrong model is favored. A model selection criterion with low Type I and Type II error rates will have both densities concentrated on the right side of the plot.

DIC, is the posterior mean a reasonable estimator? Posterior modes or medians may be much more appropriate for some parameters—variances in particular (e.g. Celeux et al., 2006). Last, how sensitive is the chosen criterion to sample size and the choice of prior distributions?

The first important result from this analysis is that despite different origins and computation of model complexity, AICc and BIC have remarkably similar behavior. AICc tends to favor simpler models to a slightly greater extent than BIC, however both criteria strongly favor simpler models in all comparisons considered here. One reason for this is that the sample sizes used in the simulated data sets are relatively small (as sample size approaches infinity, the BIC weight for the 'best' model approaches 1.0), To illustrate the effect of sample size, consider a simple comparison between the logistic and theta-logistic models. The logistic model has one less parameter (linear density dependence), but because there is a strong negative correlation between the parameters r and θ, a logistic model can fit theta-logistic data almost as well as the theta-logistic model by simply changing the growth rate (although this may no longer be biologically reasonable). When the difference in log likelihood values between the two models is smaller than the difference in their complexity (AIC = 2, BIC = ln (K)), the simpler logistic model is assigned more weight. Even for a simulated time series with precise abundance estimates (CV ~0.01), a population exhibiting non-linear density dependence ($N_0 = 200$, $r = 0.2$, $K = 1000$, $\phi = 0.7$) would have to be continuously observed for 25 years – to more than 90% of K – before the theta-logistic model would receive more AIC or BIC weight than the logistic model.

A second point is that this paper raises more questions than provides answers considering the use of DIC. Kadane and Lazar (2004) illustrated that one concern about the use of Bayes factors is that they are sensitive to the choice of priors on model parameters. Like the Bayes factor, DIC is strongly sensitive to the choice of priors. Highly informative priors for parameters that are difficult to estimate, such as the theta-logistic parameter ϕ, tend to influence both the posterior and the DIC statistic (Fig. 5). In the comparison between logistic and theta-logistic models, decreasing the variance of the prior on ϕ has the effect of making DIC smaller, giving more model weight to the theta-logistic model. A second point about DIC is that when posterior mode is used in the calculation the average model performance ($\overline{D(\theta)}$), most of the estimates of model complexity tend to be negative. As negative values of p_D generally indicate poor model fit and are not interpretable (Spiegelhalter et al., 2002), using summary statistics other than the mean or median to calculate DIC is not advised. A third potential issue is that while DIC is supposed to be a Bayesian analogue of AIC, it may behave quite differently. Recent studies have indicated that DIC may favor complex models to a greater extent than AIC.

While no one model selection tool is superior to all others under all conditions, the Bayes factor appeared to do slightly better than other model selection criteria in minimizing the total error (both Type I and Type II errors). For example, exam-

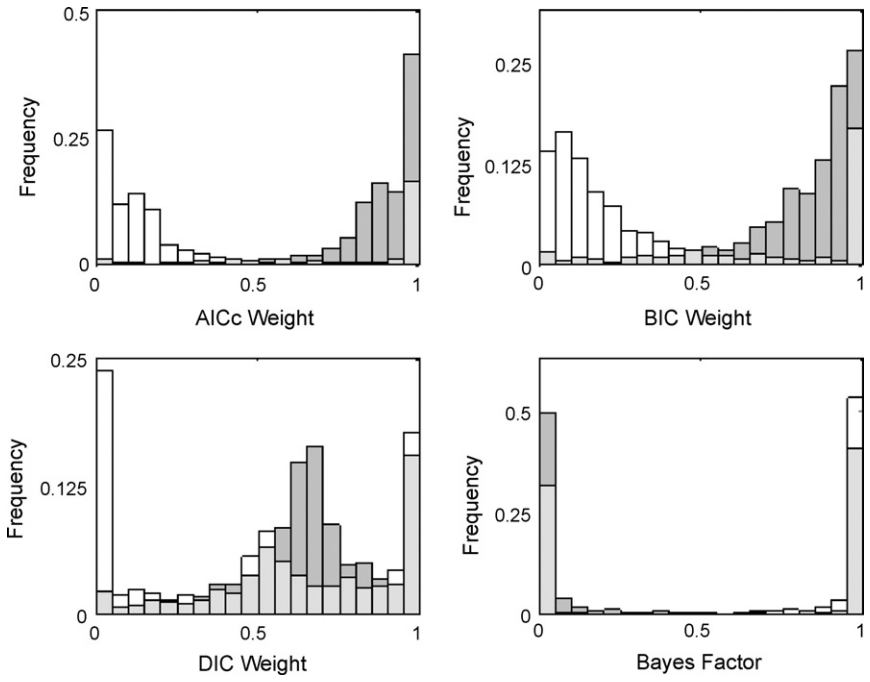


Fig. 7 – Model weights for evaluating the evidence of Allee dynamics (depensation). In each case the model weight (or in the case of the Bayes factor, the posterior probability) represents the frequency of selecting the correct model that generated the data. 5000 data sets have been generated from logistic and theta-logistic models (grey bars) and 2500 data sets have been generated from an Allee model (white bars). For both histograms, values close to 1.0 indicate that the true model is favored, while values close to 0.0 indicate that the wrong model is favored. A model selection criterion with low Type I and Type II error rates will have both densities concentrated on the right side of the plot.

ining the distribution of weights greater than 0.5 illustrates that AICc selected the simulation model only 50% of the time, BIC selected the simulation model 49.3% of the time, DIC selected the simulation model 47.6% of the time, and the Bayes factor selected the simulation model 53.1% of the time. This result is promising, because of the four criteria considered here, Bayes factors are most compatible with model averaging and decision making. Although Bayes factors have been studied and used in practice to a greater degree than DIC, more research needs to be done on several components of the Bayes factor, including sensitivity of prior distributions (specifically applied to non-linear population models) and the importance function used in the marginal likelihood calculation. A much larger question that needs to be studied is how maximum likelihood criteria (AICc, BIC, ICOMP) perform when used in the context of Bayesian model averaging (Hoeting et al., 1999). Preliminary results have indicated that AICc gives mixed results (Richards, 2005; Richards, 2008), but it is unknown how other criteria perform.

REFERENCES

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle. In: Petrov, B.N., Csaksi, F. (Eds.), Proceedings of the 2nd International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary, pp. 267–281.

A'mar, Z.T., 2004. Quantifying error and uncertainty in fishery stock assessment models. M.Sc. Thesis, School of Aquatic and Fishery Sciences, University of Washington.

Anderson, D.R., Burnham, K.P., White, G.C., 1994. AIC model selection in overdispersed capture-recapture data. Ecology 75, 1780–1793.

Bernardo, J.M., Smith, A.F.M., 2000. Bayesian Theory. Wiley, New York, NY.

Best, N.G., Cowles, M.K., Vines, S.K., 1995. CODA Manual version 0.30. MRC Biostatistics Unit, Cambridge.

Bozdogan, H., 2000. Akaike's Information Criterion and recent developments in model complexity. J. Math. Psychol. 44, 62–91.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection. Sociol. Methods Res. 33, 261–304.

Burnham, K.P., White, G.C., Anderson, D.R., 1995. Model selection strategy in the analysis of capture–recapture data. Biometrics 51, 888–898.

Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M., 2006. Deviance information criteria for missing data models. Bayesian Analysis 4, 651–674.

Courchamp, F., Clutton-Brock, T., Grenfell, B., 1999. Inverse density dependence and the Allee effect. Trends Ecol. Evol. 14, 405–410.

Ellison, A.E., 2004. Bayesian inference in ecology. Ecol. Lett. 7, 509–520.

Forster, M.R., 2000. Key concepts in model selection: performance and generalizability. J. Math. Psychol. 44, 205–231.

Forster, M.R., Sober, E., 2004. Why likelihood? In: The Nature of Scientific Evidence. University of Chicago Press.
Fournier, D.A., 1996. AUTODIFF. A C++ array language extension with automatic differentiation for use in nonlinear modeling and statistics. Otter Research Limited, Nanaimo, British Columbia, Canada.
Fried, S.M., Hilborn, R., 1988. In-season forecasting of Bristol Bay, Alaska, sockeye salmon (*Oncorhynchus nerka*) abundance using Bayesian probability theory. Can. J. Fish. Aquat. Sci. 45, 850–855.
Gelfand, A.E., Dey, D.K., 1994. Bayesian model choice: asymptotics and exact calculations. J. Roy. Stat. Soc. Ser. B 56, 501–514.
Gelman, A.B., Carlin, J.S., Stern, H.S., Rubin, D.B., 1995. Bayesian Data Analysis. Chapman and Hall, New York, NY.
Ghosh, J.K., Samanta, T., 2001. Model selection—an overview. Curr. Sci. 80, 1135–1144.
Gilpin, M.E., Ayala, F.J., 1973. Global models of growth and competition. Proc. Natl. Acad. Sci. 70, 3590–3593.
Good, I.J., 1958. Significance tests in parallel series. J. Am. Stat. Assoc. 53, 799–813.
Hobbs, N.T., Hilborn, R., 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecol. Appl. 16, 5–19.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Stat. Sci. 14, 382–401.
Hurvich, C.M., Tsai, C.L., 1995. Model selection for extended quasi-likelihood models in small samples. Biometrics 51, 1077–1084.
Jeffreys, H., 1935. Some tests of significance, treated by the theory of probability. Proc. Cambridge Philos. Soc. 31, 203–222.
Johnson, J.B., Omland, K.S., 2004. Model selection in ecology and evolution. Trends Ecol. Evol. 19, 101–108.
Kadane, J.B., Lazar, N.A., 2004. Methods and criteria for model selection. J. Am. Stat. Assoc. 99, 279–290.
Kass, R., 1993. Bayes factors in practice. Statistician 42, 551–560.
Kass, R., Raftery, A.E., 1995. Bayes factors. J. Am. Stat. Assoc. 90, 773–795.
Kass, R., Wasserman, L., 1996. The selection of prior distributions by formal rules. J. Am. Stat. Assoc. 91, 1343–1370.
Kuha, J., 2004. AIC and BIC: comparisons of assumptions and performance. Sociol. Methods Res. 33, 188–229.
Lewin-Koh, N., Taper, M.L., Lele, S.R., 2004. A brief tour of statistical concepts. In: The Nature of Scientific Evidence. University of Chicago Press.
Lewis, M.A., Kareiva, P., 1993. Allee dynamics and the spread of invading organisms. Theor. Popul. Biol. 43, 141–158.
Link, W.A., Barker, R.J., 2005. Model weights and the foundations of multimodel inference. Ecology 87, 2626–2635.
Madigan, D., Raftery, A.E., 1995. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Am. Stat. Assoc. 89, 1535–1546.
Maurer, B.A., 2004. Models of scientific inquiry and statistical practice: implications for the structure of scientific knowledge. In: The Nature of Scientific Evidence. University of Chicago Press.
McAllister, M.K., Kirkwood, G.P., 1998. Bayesian stock assessment: a review and example application using the logistic model. ICES J. Mar. Sci. 55, 1031–1060.
McQuarrie, A.D.R., Tsai, C.-L., 1998. Regression and Time Series Model Selection. World Scientific Publishing Company, London.
Myung, I.J., 2000. The importance of complexity in model selection. J. Math. Psychol. 44, 190–204.
Punt, A.E., Hilborn, R., 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Rev. Fish Biol. Fish. 7, 35–63.
Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. 25, 111–169.
Raftery, A.E., 1999. Bayes factors and BIC: comment on Weakliem. Sociol. Methods Res. 27, 411–427.
Richards, S.A., 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology 86, 2805–2814.
Richards, S., 2008. Dealing with overdispersed count data in applied ecology. J. Appl. Ecol., in press.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6, 461–464.
Shono, Hiroshi, 2000. Efficiency of the finite correction of Akaike's information criteria. Fish. Sci. 66, 608–610.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of complexity and fit (with discussion). J. Roy. Stat. Soc. B 64, 583-540.
Spiegelhalter, D., Thomas, A., Best, N.G., 2003. WinBUGS version 1.4 User Manual. MRC and Imperial College of Science, Technology, and Medicine. Available from http://www.mrc-bsu.cam.ac.uk/bugs (accessed June 2006).
Stone, M., 1979. Comments on model selection criteria of Akaike and Schwartz. J. Roy. Stat. Soc. B 41, 276–278.
Taper, M.L., 2004. Model identification from many candidates. In: The Nature of Scientific Evidence. University of Chicago Press.
Vaida, F., Blanchard, S., 2005. Conditional Akaike information for mixed-effects models. Biometrika 92, 351–370.
Weiss, R., 1997. Bayesian sample size calculations for hypothesis testing. Statistician 46, 185–191.
Wintle, B.A., McCarthy, M.A., Volinsky, C.T., Kavanagh, R.P., 2003. The use of Bayesian model averaging to better represent uncertainty in ecological models. Conserv. Biol. 17, 1579–1590.