Ecology, 86(10), 2005, pp. 2805–2814

TESTING ECOLOGICAL THEORY USING THE INFORMATION-THEORETIC APPROACH: EXAMPLES AND CAUTIONARY RESULTS

SHANE A. RICHARDS[1]

Department of Biological Sciences, University of Calgary, Calgary, Alberta T2N 1N4 Canada

Abstract. Ecologists are increasingly applying model selection to their data analyses, primarily to compare regression models. Model selection can also be used to compare mechanistic models derived from ecological theory, thereby providing a formal framework for testing the theory. The Akaike Information Criterion (AIC) is the most commonly adopted criterion used to compare models; however, its performance in general is not very well known. The best model according to AIC has the smallest expected Kullback-Leibler (K-L) distance, which is an information-theoretic measure of the difference between a model and the truth. I review the theory behind AIC and demonstrate how it can be used to test ecological theory by considering two example studies of foraging, motivated by simple foraging theory. I present plausible truths for the two studies, and models that can be fit to the foraging data. K-L distances are calculated for simulated studies, which provide an appropriate test of AIC. Results support the use of a commonly adopted rule of thumb for selecting models based on AIC differences. However, AIC_c, a corrected version of AIC commonly used to reduce model selection bias, showed no clear improvement, and model averaging, a technique to reduce model prediction bias, gave mixed results.

Key words: Akaike Information Criterion (AIC); foraging; model averaging; model selection; theoretical ecology.

INTRODUCTION

A key problem when analyzing ecological data is identifying which model, or models, best describe the data from a set of models proposed a priori, and is referred to as model selection. Often, the objective of an ecological analysis is to identify factors that most affect some response variable of interest. For example, the objective may be to identify habitat variables that influence species presence (e.g., Fernández et al. 2003, Godínez-Domínguez and Freire 2003, Westphal et al. 2003). In these types of problems, model selection generally involves the comparison of regression models.

Another common objective of a data analysis is to infer key ecological processes that generated the data. Regression models may describe patterns in the data, but often do not provide much insight about processes. However, models that explicitly incorporate ecological processes and dynamics can be developed from ecological theory (Gurney and Nisbet 1998). Model selection is increasingly being applied to mechanistic models to infer processes affecting population dynamics (Hilborn and Mangel 1997, Leirs et al. 1997, Fujiwara and Caswell 2001) and evolutionary dynamics (Posada and Buckley 2004), to interpret results from capture–recapture studies (Anderson et al. 1998), and to address problems involving fisheries management (Hilborn and Walters 1992, Helu et al. 2000). The ability to quantify and compare mechanistic models means that ecological theory can be rigorously tested. Model selection is useful in this case because it can help identify systems that are difficult to predict, thereby suggesting important directions for research. For example, if no single model is found to be clearly best, then new ecological hypotheses may need to be developed, or more data collected.

Model selection involves quantifying model performance based on some criterion. Information theory (IT) can be used to derive such a criterion and its application is becoming increasingly common in ecology (Burnham and Anderson 2002). The Akaike Information Criterion (AIC) is the most commonly adopted measure used to quantify and compare models under the IT approach. This measure combines the goodness of fit of a model to data and the number of estimated model parameters. AIC reflects model parsimony, which is a trade-off between prediction bias and parameter uncertainty. AIC is simple to implement and many statistical packages report the AIC value of a model. Burnham and Anderson (2002) present the theory behind the IT approach and AIC, and is primarily aimed at an ecological audience. A concise summary of the approach is presented in Burnham and Anderson (2001).

A key assumption of AIC is that model performance is measured by its expected Kullback-Leibler (K-L) distance (Kullback and Leibler 1951). The K-L distance quantifies the discrepancy between the distribution describing the true probability of observing outcomes from a study and the distribution predicted by the model. Despite an increase in the use of AIC, its

Manuscript received 19 January 2005; accepted 29 March 2005. Corresponding Editor: B. Shipley.

[1] E-mail: richas@ucalgary.ca

performance in general is not very well known (Burnham and Anderson 2002). Two reasons for this uncertainty are that K-L distances can only be evaluated if the true probability distribution of study outcomes is known, which is not possible for real studies, and the expected K-L distance is often numerically expensive to calculate. In this paper, I briefly review the theory behind AIC, and then demonstrate how it can be used to test simple foraging theory. I present a critical assessment of AIC performance by performing computationally intensive calculations of K-L distances based on plausible truths for two simple foraging studies.

By explicitly calculating K-L distances, I am able to clarify through examples many non-trivial concepts relevant to an AIC analysis. For example, I illustrate the definition of best model according to AIC, and demonstrate how the best model may change depending on the amount of data collected. I also demonstrate why the fact that AIC is only an estimate of the relative, expected K-L distance is important for model selection. Two example studies are presented because they help illustrate the generality of the results. I found that a commonly adopted rule of thumb for selecting models based on differences in their AIC values (Burnham and Anderson 2002) was supported by both studies. The addition of a correction factor to AIC that is commonly adopted and referred to as corrected AIC (AIC_c; Burnham and Anderson 2002), was not found to improve the likelihood of selecting the best model for both studies. On the other hand, the technique of model averaging, whereby a new unconditional model is derived using model weights, was only found to reduce model prediction bias in one of the two studies.

The Information-Theoretic Approach

If an ecological study were repeated, it is unlikely that the data collected would be the same. Variation in study outcomes might be partly due to differences in study subjects if they are chosen at random, but mostly due to the difficulty in controlling the enormous number of processes and factors affecting the outcome. In this paper, the term ''truth'' is defined as the probability distribution of outcomes that would be observed if a study were repeated an infinite number of times and the processes generating the outcome did not change. For example, suppose the outcomes from a study only comprise discrete variables (e.g., when study subjects are counted) and the probability of observing outcome indexed by i is p_i. Truth is the set (or vector) of these probabilities, denoted $\mathbf{p}$, and it clearly depends on the design of the study (e.g., number of study subjects). When outcomes are continuous variables (e.g., when study subjects are measured) truth is defined by a probability density function $f(x)$ where x is an outcome of the study. In this paper, I focus on the discrete case.

Suppose a stochastic mathematical model, based on some ecological theory, predicts that the study should produce outcome i with probability π_i. Let $\boldsymbol{\pi}$ denote the set of predicted probabilities associated with all possible study outcomes. Kullback and Leibler (1951) used IT to derive a metric describing the information lost when distribution $\boldsymbol{\pi}$ is used to approximate the true distribution $\mathbf{p}$; namely,

$$I(\mathbf{p}, \boldsymbol{\pi}) = \sum_i p_i \ln\left(\frac{p_i}{\pi_i}\right). \quad (1)$$

I is referred to as the Kullback-Leibler distance. When outcomes are continuous the K-L distance is calculated in the same manner except that $\mathbf{p}$ and $\boldsymbol{\pi}$ are replaced by $f(x)$ and $\pi(x)$, respectively, and the summation is replaced by integration with respect to x.

For models to make quantitative predictions, their parameters, denoted $\boldsymbol{\theta}$, need to be assigned values. The best parameter values for a model, according to the IT approach, are those that generate predictions $\boldsymbol{\pi}(\boldsymbol{\theta})$ that minimize the K-L distance. The K-L distance cannot be used directly to determine the best parameters for a model because it requires knowledge of the truth $\mathbf{p}$, which in practice is unknowable. However, parameters of a model can be estimated readily by fitting the model to data using maximum likelihood. Akaike (1973) suggested that the expected K-L distance of a model when fit to data using maximum likelihood could provide a basis for comparing models. The model with the lowest expectation might be considered the best because it most consistently has a low K-L distance when fit to data.

If the study were repeated many times and model parameters re-estimated each time, then the expected K-L distance for a model would be

$$E_{\mathbf{p}}[I(\mathbf{p}, \boldsymbol{\pi})] = \sum_j p_j I[\mathbf{p}, \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_j)], \quad (2)$$

where $\hat{\boldsymbol{\theta}}_j$ denotes the maximum-likelihood estimates for the parameters of a model when fit to outcome j. The model having the lowest expected K-L distance is referred to as the best K-L model, and is the model we wish to identify.

Substituting Eq. 1 into Eq. 2 gives

$$E_{\mathbf{p}}[I(\mathbf{p}, \boldsymbol{\pi})] = \sum_i p_i \ln p_i - E_{\mathbf{p}}\left[\sum_i p_i \ln \pi_i(\hat{\boldsymbol{\theta}}_j)\right]. \quad (3)$$

The first term on the right side of Eq. 3, $\Sigma\ p_i \ln p_i$, depends only on the fixed truth $\mathbf{p}$, and hence, is included in the calculation of $E_{\mathbf{p}}\{I\}$ for all models considered. Let this constant term be denoted c. Akaike (1973) showed that the last term could be estimated by

$$E_{\mathbf{p}}\left[\sum_i p_i \ln \pi_i(\hat{\boldsymbol{\theta}}_j)\right] \approx \ln L(\hat{\boldsymbol{\theta}}_j) - K \quad (4)$$

where $\ln L(\hat{\boldsymbol{\theta}}_j)$ is the maximum log-likelihood of the model, given outcome j, and K is the number of estimated model parameters. Note that $L(\hat{\boldsymbol{\theta}}_j) = \pi_j(\hat{\boldsymbol{\theta}}_j)$. Given Eq. 4, Akaike (1973) proposed the Akaike Infor-

mation Criterion (AIC) for model M when outcome j is observed:

$$\text{AIC}(M) = -2 \ln L(\hat{\boldsymbol{\theta}}_j) + 2K. \quad (5)$$

Substituting Eqs. 3 and 4 into Eq. 5, AIC can be shown to estimate

$$\text{AIC}(M) \approx 2\{E_{\mathbf{p}}[I(\mathbf{p}, \boldsymbol{\pi})] - c\}. \quad (6)$$

Hence, the lower the AIC value associated with a model, the more likely it has the lowest expected K-L distance (i.e., is the best K-L model).

The simplest model selection approach is to calculate the AIC value for each competing model and then choose the model with the lowest value as being the best K-L model. However, this approach ignores the fact that AIC is an estimate; because of sampling error, a model not having the lowest AIC value could be the model that has the lowest expected K-L distance. The problem now is to identify which models are likely to have the lowest expected K-L distance, and thus, should be kept and used for inference.

AIC values by themselves are relatively uninformative, what is important is the differences in AIC values among competing models. It is convenient to calculate a Δ value for each model, which indicates the difference in the AIC value of the model from the minimum AIC value of all models considered, i.e.,

$$\Delta(M) = \text{AIC}(M) - \text{AIC}_{\text{min}}. \quad (7)$$

Burnham and Anderson (2002) suggested the following rule of thumb for selecting among models. Models with a Δ value <2 are all likely to be the best K-L model, and hence, they should all be used when making further inferences about a system. Models with a Δ value in the range 4–7 are less likely to be best, but probably should not be discounted. Models with a Δ value >10 are extremely unlikely to be the best K-L model and can be ignored.

Akaike (1983) proposed that the likelihood of model M being the best K-L model, given the data, is proportional to the quantity $\exp(-\Delta(M)/2)$, which leads to the notion of model weights. The Akaike weight of model M_m is

$$w_m = \frac{\exp[-\Delta(M_m)/2]}{\sum_i \exp[-\Delta(M_i)/2]}. \quad (8)$$

These weights sum to 1 across models and have been interpreted as an estimate of the proportion of times that model M_m will be chosen as having the lowest AIC value if the study were repeated (Burnham and Anderson 2002). These weights are also often interpreted as the probability that model M_m is the best K-L model.

In some cases, quantitative predictions are of interest. Predictions could be made from the fitted model having the lowest AIC value; however, as AIC is an estimate, it seems reasonable to make predictions that take into account model uncertainty. Burnham and Anderson (2002) suggested that model bias could be reduced by model averaging, whereby predictions from each model are weighted by their Akaike weight. This new prediction is often termed unconditional because it is not conditional on a single model. Akaike weights are often used to calculate unconditional estimates of regression coefficients (Buckland et al. 1997).

The AIC estimate given by Eq. (5) assumes asymptotic properties that are well approximated when there are a large number of independent observations. Burnham and Anderson (2002) suggested that, when the ratio of observations to model parameters is low (e.g., $N/K < 40$), then a corrected version of AIC should be used:

$$\text{AIC}_c(M) = \text{AIC}(M) + \frac{2K(K+1)}{N-K-1}. \quad (9)$$

AIC_c assumes that the data were generated by a fixed-effect linear model with homogeneous, normally distributed errors, and the models analyzed also have this form (Hurvich and Tsai 1989). Despite these assumptions, the above correction factor is suggested as being useful in other contexts. Corrected Δ values and Akaike weights can also be calculated with AIC replaced by AIC_c.

Foraging Examples

Consider a study of pollen dispersal among flowers of the perennial plant, *Aquilegia brevistyla*. Bumble bees are a major pollinator of this species and their foraging behavior, in general, has been well studied (e.g., Pyke 1982, Cresswell 1990). If bumble bees maximize their long-term net rate of energy intake from collecting nectar (Pyke 1982), they should forage among plants so that the expected nectar reward received from each plant visited is equal. If flowers on *A. brevistyla* all produce nectar at a similar rate, then the expected reward when visiting an *A. brevistyla* plant will be equalized among plants if bumblebees visit plants at a rate proportional to the number of open flowers on the plant (Possingham 1992). This simple theory also suggests that if microhabitat affects nectar production, then it may influence plant visitation rates (Dreisig 1995). Two examples of microhabitat variables that affect nectar production are soil moisture and sunlight (Zimmerman 1983). Numerous observational studies and experiments have tested these ideas (e.g., Klinkhamer and de Jong 1990, Kadmon 1992, Mitchell et al. 2004).

To illustrate how the above foraging theory could be tested, I now consider two simple hypothetical observational studies, and models motivated by the theory that could be fit to the resulting data. For each study I present a plausible truth, $\mathbf{p}$. The IT approach was tested by comparing the theory presented in the previous section with results from simulated study trials. Expectations with respect to $\mathbf{p}$ were accurately estimated by averaging over 2000 trials. These simulations

were used to assess (1) AIC and AIC_c estimates of relative, expected K-L distance, (2) the rule of thumb for selecting models and retaining the best K-L model, (3) Akaike weight as the probability of being the best K-L model, and (4) reduction of model bias by model averaging.

Study 1: methods

The following study could be conducted to test whether flower number and microhabitat affect plant visitation rates by bumble bees. Suppose the study was performed in an open forest containing numerous *A. brevistyla* plants, and plant microhabitat was defined by the level of shade, with the expectation that shaded plants produce nectar at a lower rate. Plants were placed into one of eight categories according to the number of open flowers (one to four) and whether or not they were shaded. For each category, r plants were chosen randomly, giving $8r$ plants in total. During a 30-min period when bumble bees were active the randomly chosen plants were observed and noted whether they received at least one bumble bee visit.

Study 1: truth and approximating models

First, I present a possible truth for this study. Let s denote a plant's shading status, where $s = 0$ for unshaded plants, and $s = 1$ for shaded plants. Suppose bumble bee arrivals to a plant is a constant Poisson process. Let $\alpha(s, f)$ be the true long-term rate that *A. brevistyla* plants received bumble bee visits if they had f open flowers and shading status s. The true probability that a plant received one or more bumble bee visits during a time period of duration Δt is

$$q(s, f) = 1 - \exp[-\Delta t\, \alpha(s, f)]. \qquad (10)$$

Fig. 1 shows a possible example of the true probabilities that a plant received at least one visit. In this case, the truth was generated by a model with eight parameters (i.e., an α for each of the eight plant categories) and describes the case where arrivals are near linearly related to flower number and negatively related to shading. In reality, the truth could be thought of as being generated by a model with an infinite number of parameters (Burnham and Anderson 2002).

Outcomes from this study can be summarized by eight integers, $n(s, f)$, denoting the number of plants with f flowers and shading status s that received at least one bumble bee visit ($0 \leq n(s, f) \leq r$). The number of outcomes that could be observed with this study, denoted Z, rises very quickly as plant replication r is increased, $Z = (r + 1)^8$. Even with just one plant per category ($r = 1$, eight plants in total), there are 256 possible outcomes. For $r = 5$ (40 plants) there are 1 679 616 possible outcomes. The p_i ($i = 1, \ldots Z$) can be calculated using the following:

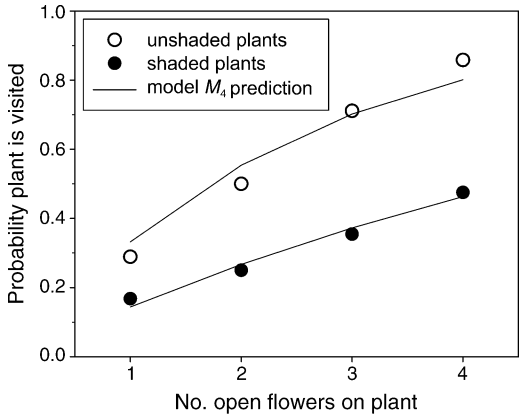

FIG. 1. Study 1, level of plant replication, $r = 1$. Circles represent the probabilities that a plant will be visited by at least one bumble bee during a 30-min observation period, q, depending on whether the plant is shaded and the number of open flowers it has. Lines represent predictions from model M_4 when its parameter values minimize the Kullback-Leibler distance.

$$p_i = \prod_{s=0}^{1} \prod_{f=1}^{4} \mathrm{B}[n_i(s, f); r, q(s, f)] \qquad (11)$$

where $n_i(s, f)$ is the number of plants with trait (s, f) visited according to outcome i, and $\mathrm{B}[n; r, q]$ is the binomial function describing the probability of observing n successes from r trials, given that successes occur with probability q and trials are independent. Fig. 2 shows the true distribution $\mathbf{p}$ when $r = 1$ and truth is parameterized by the probabilities in Fig. 1. For simplicity, I assumed that the truth did not describe a distribution of outcomes where the data were overdispersed. Examining the effect of overdispersed data on AIC estimates will be the focus of future work.

To test the foraging theory presented, I compared four models, denoted M_m ($m = 1, \ldots 4$), which differed in whether or not shading status or flower number affected bumble bee arrival rate α. The models can be summarized as follows.

M_1: $\alpha(s, f) = a$ (arrival rate is constant across plants, regardless of shading status and flower number, $\boldsymbol{\theta} = \{a\}$).

M_2: $\alpha(s, f) = a_s$ (arrival rate differs among shade environments, but not with flower number, $\boldsymbol{\theta} = \{a_0, a_1\}$).

M_3: $\alpha(s, f) = fb$ (arrival rate varies linearly with flower number, but not among shade environments, $\boldsymbol{\theta} = \{b\}$).

M_4: $\alpha(s, f) = fb_s$ (arrival rate differs among shade environments and varies linearly with flower number, $\boldsymbol{\theta} = \{b_0, b_1\}$).

Models M_3 and M_4 assume the rate of arrival is linearly related to flower number, in accordance with theory. If model M_1 fits the data well, then it might be that bumble bees forage randomly among plants, or some other unmeasured factors are important, in which

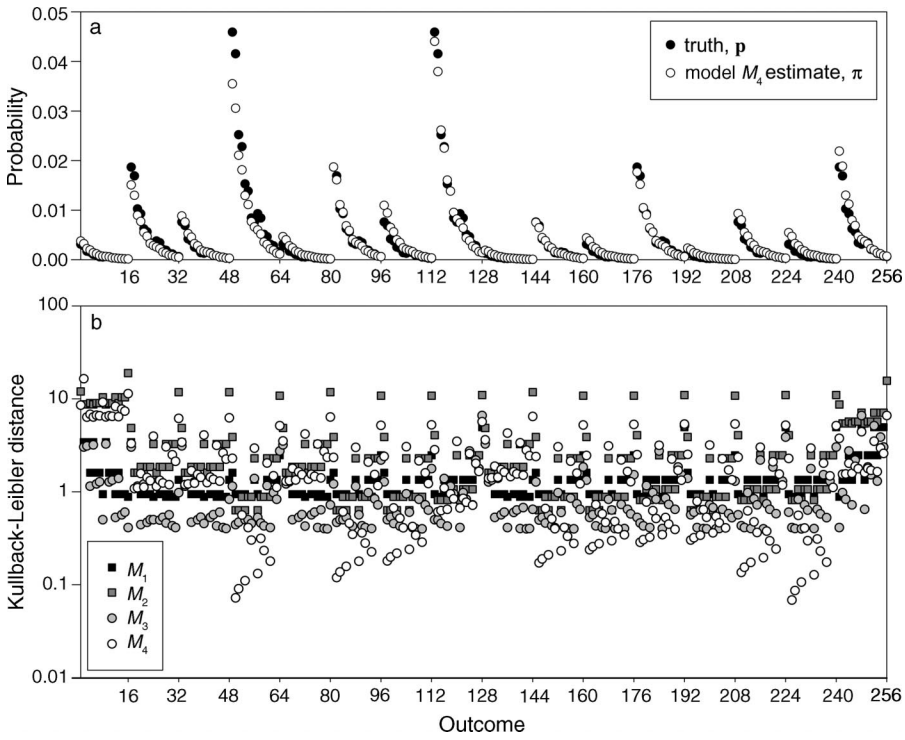

FIG. 2. Study 1, level of plant replication, $r = 1$. (a) True probabilities (black circles) and the best-fitting model M_4 probabilities (open circles), of observing all 256 possible outcomes. (b) The Kullback-Leibler distance associated with each of the four models for all outcomes, when their parameters are estimated using maximum likelihood. Model parameters were constrained so that they bounded the probability of visitation per plant to be between 0.01 and 0.99. Outcomes of the study are described by an eight-digit binary number. For example, the observation {00110001} describes the case when only the three- and four-flowered unshaded plant and the four-flowered shaded plant were visited during the study. Study outcomes on the *x*-axis represent the binary equivalent of the data collected + 1 (e.g., the data {00110001} is binary for 49, and corresponds to outcome 50).

case alternative hypotheses should be considered. An unconditional model can be derived by considering the following weighted arrival rate:

$$\bar{\hat{\alpha}} = \sum_m w_m \hat{\alpha}_m \tag{12}$$

where $\hat{\alpha}_m$ is the maximum-likelihood arrival rate estimated by model M_m.

Study 1: results

The $\boldsymbol{\pi}(\boldsymbol{\theta})$ and K-L distance for each model can be calculated by substituting the model's estimate of α into Eqs. 10 and 11. In this example, none of the four models describe the truth presented in Fig. 2 exactly. Model M_4 best approximates the truth, and this occurs when its two parameters are set to $b_0 = 0.81$ and $b_1 = 0.31$ visits per flower per hour (Figs. 1 and 2). Model M_4 can approximate the truth well because it incorporates two aspects of the truth; namely, variation among microhabitats and the positive effect of flowers. For most ecological examples, no model should be expected to be able to exactly describe the truth (i.e., $I(\mathbf{p}, \boldsymbol{\pi}(\boldsymbol{\theta})) > 0$ for all $\boldsymbol{\theta}$).

To illustrate the AIC definition of best model, I fit the four models to all 256 possible outcomes when $r = 1$ using maximum likelihood. The model having the lowest K-L distance depended on the outcome observed (Fig. 2). Even though model M_4 had the smallest K-L distance for a number of outcomes it also often performed poorly. Model M_4 is not a good general model in this example because it overfits the data; that is, it reproduces the data well when fit to some outcomes but performs poorly when fit to others (Draper 1995). At the same time, models M_1 and M_2 underfit the data because they simplify the system too much, which results in biased predictions. Models M_1 and M_2 typically underestimate visits to many-flowered plants and overestimate visits to few-flowered plants. In this example, Model M_3 is the most parsimonious model as it often describes the system relatively well with only one parameter.

Even though model M_3 was the best K-L model when $r = 1$, for larger r the best model was the more complex model M_4 (Table 1). M_3 and M_4 both included the effect of flower number, indicating that flower number affected arrival rates more than microhabitat (Fig. 1). For all three cases of r considered, the model most likely to be chosen by AIC had the lowest expected K-L distance. This result was also true for AIC_c, except when $r = 3$ (Table 1). For $r = 5$, both AIC and AIC_c provided

TABLE 1. Statistics from 2000 trials illustrating the performance of the models for both foraging studies.

r	Model	$E_{\mathbf{p}}\{I\}$	Lowest I (%)	Lowest AIC (%)	Lowest AIC_c (%)	AIC SD	$E_{\mathbf{p}}\{w\}$ (%)	w SD (%)
Study 1								
1	M_1	1.35	1.9	14.5	21.7	1.3	19.7	12.6
	M_2	2.34	3.0	13.2	6.2	2.6	18.2	13.9
	M_3	**0.95**	50.7	45.7	66.0	2.0	33.5	17.4
	M_4	1.40	44.4	26.7	6.2	2.6	28.6	14.0
3	M_1	2.99	0.0	7.6	7.6	3.0	10.6	12.6
	M_2	2.71	0.2	7.2	7.2	3.3	14.0	15.3
	M_3	1.69	19.2	37.6	43.1	3.1	31.1	21.6
	M_4	**1.27**	80.7	47.8	42.2	2.9	44.3	23.2
5	M_1	4.61	0.0	2.9	2.9	4.5	5.9	10.2
	M_2	3.56	0.0	6.0	5.7	4.0	10.2	14.7
	M_3	2.40	7.5	26.1	29.5	4.0	26.6	22.2
	M_4	**1.15**	92.5	65.0	61.8	2.9	57.3	26.0
Study 2								
5	M_5	1.04	0.2	33.2	33.8	2.9	30.7	18.2
	M_6	**0.57**	78.6	60.3	63.4	2.9	43.0	16.4
	M_7	1.23	21.3	6.6	2.8	2.5	26.3	12.6
10	M_5	1.55	0.0	26.4	26.5	3.6	24.8	20.8
	M_6	**0.58**	71.7	69.1	69.9	3.2	48.1	18.8
	M_7	1.04	28.4	4.6	3.7	2.8	27.1	9.8

Notes: Measures include level of plant replication (r), expected Kullback-Leibler (K-L) distance ($E_{\mathbf{p}}\{I\}$), percentage of times that the model had the lowest K-L distance (lowest I), percentage of times that the model had the lowest AIC and AIC_c value, the standard deviation (SD) of AIC values, expected Akaike weights ($E_{\mathbf{p}}\{w\}$), and the standard deviation of the Akaike weights (w). Bold values of $E_{\mathbf{p}}\{I\}$ indicate the best K-L model.

relatively unbiased estimates of expected, relative K-L distances (Fig. 3). Variation in AIC and AIC_c values among trials increased slowly with replication (Table 1). The rule of thumb for retaining the best K-L model performed well in all cases (Table 2). Δ values calculated from AIC_c values performed slightly better than AIC values when $r = 1$; however, AIC was better for $r = 3$ and 5. In all cases, the best K-L model was retained for at least 77% of the trials when the threshold was 2, and for at least 97% of the trials when the threshold was 7.

Expected Akaike weights and the probability a model would be chosen by AIC were positively correlated; however, there was large variation in the weights across the 2000 trials (Table 1). Expected Akaike weights and the expected K-L distance were also positively correlated. On average, predictions from the unconditional model had a lower K-L distance than the lowest AIC model; however, contrary to expectation, model averaging was less effective at reducing model bias (i.e., reducing $E_{\mathbf{p}}\{I\}$) as r was increased (Table 2).

Study 2: methods

This study focused on the effect of flower number on bumble bee arrival rates. During this study r one-flowered and r two-flowered plants were chosen randomly from an area within the forest where microhabitat was consistent (e.g., all plants were unshaded, plant density was similar, and soil moisture varied little). Plants were again monitored for 30 min but this time the number of pollinator visits to each plant was recorded.

Study 2: truth and approximating models

To keep the number of outcomes for this study tractable, the number of visits to a plant was recorded as either $y = 0, 1, 2, 3$, or 4^+, the latter indicating that four or more visits were observed. Data for this study are a set of 10 integers, denoted $n(y, f)$, which indicate the number of f-flowered plants receiving y visits. These data satisfy $\Sigma_y n(y, f) = r$ for $f = 1, 2$. The number of possible outcomes given r replicates is $Z = [(r + 1)(r + 2)(r + 3)(r + 4)/24]^2$. For this study, I examined two levels of plant replication: $r = 5$ ($Z = 15\,876$), and $r = 10$ ($Z = 1\,002\,001$).

Bumble bee arrivals to plants was modeled as a constant Poisson process. Plant visits were independent and bees visited one-and two-flowered plants at rate $\alpha(1) = 2.2$ and $\alpha(2) = 3.8$ visits/h, respectively. Thus, two-flowered plants received visits at a higher rate, which was less than twice the rate for one-flowered plants. The true probability distribution of outcomes $\mathbf{p}$ was calculated in a similar manner to study 1, except that the binomial distribution was replaced by a multinomial distribution because there were five possible observations for each plant.

I compared three models for this study, all of which assumed that bumble bee arrivals were a constant Poisson process. The models can be summarized as follows.

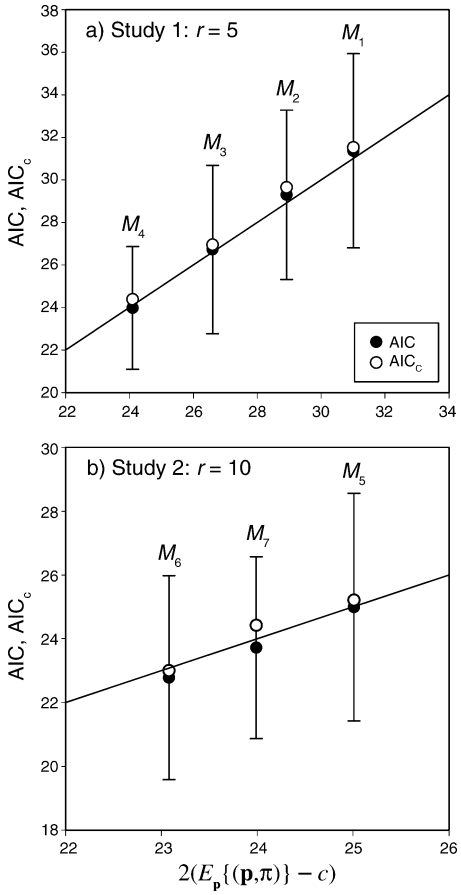

FIG. 3. AIC and AIC_c estimates of the relative, expected Kullback-Leibler distance for (a) study 1 and (b) study 2. The solid line indicates a 1:1 relationship. Circles indicate means, and vertical bars indicate the variation observed (SD) among AIC calculations across 2000 trials. Standard deviations for AIC_c values are the same as those for AIC and have been omitted for clarity.

M_5: $\alpha(f) = a$ (arrival rate is constant across plants, regardless of flower number, $\boldsymbol{\theta} = \{a\}$).

M_6: $\alpha(f) = fb$ (arrival rate varies linearly with flower number, $\boldsymbol{\theta} = \{b\}$).

M_7: $\alpha(1) = a_1$, $\alpha(2) = a_2$ (arrival rate varies with flower number, $\boldsymbol{\theta} = \{a_1, a_2\}$).

Models M_5 and M_6 make the same assumptions as M_1 and M_3 of the previous study, respectively. Models M_6 and M_7 both assume that flower number affects visitation rate; however, M_7 has no restriction on the form of the relationship. Model M_7 is so flexible it can match the truth exactly; however, it may not be particularly informative because it does not suggest any mechanism. If M_7 fits the data well it would suggest the need for alternative hypotheses.

Study 2: results

Analyses for this study are presented for two levels of plant replication ($r = 5$ and 10). In both cases, model M_6 was the best K-L model, indicating that the data include too little information (r was too low) to estimate accurately model parameters for M_7, even though it was used to generate the data. AIC and AIC_c were both most likely to choose the best K-L model in all cases (Table 1). As in the previous study, AIC and AIC_c values both provided a relatively unbiased estimate of the relative, expected K-L distance (Fig. 3). Variation in AIC and AIC_c values was comparable to that of the previous study (Table 1). Again, the rule of thumb for retaining the best K-L model was supported; however, Δ values calculated from AIC values were more likely to lead to correct retention than AIC_c (Table 2).

In contrast to the previous study, Akaike weights provided relatively poor estimates of the probability of a model being selected by AIC, and of a model's probability of being the best K-L model (Table 1). Variation in Akaike weights among trials was again very high (Table 1). For this study, the unconditional model had on average a higher K-L distance than the best AIC model, indicating that model averaging did not reduce model bias. As with the previous study, the unconditional model performed worse as r was increased (Table 2).

DISCUSSION

AIC allows models to be compared, thereby providing a formal way to compare theories. AIC estimates the expected, relative K-L distance, and the model having the lowest value for this metric is termed the best K-L model. AIC attempts to reward models that are robust to parameterization. For example, M_7 was not the best K-L model for study 2, even though it generated the data for the simulations (Table 1). In this case, model M_7 suffers from overfitting because, when it is fit to common outcomes of the study, its parameter estimates are often so poor that it then predicts other outcomes that do not reflect outcomes that are truly likely. The results presented here illustrate how AIC reflects model parsimony (i.e., a trade-off between model prediction bias and parameter uncertainty).

AIC can be applied to data combined from multiple studies that do not necessarily have the same design. For example, suppose study 1 and study 2 were performed with different plant or pollinator species, or in different locations. By combining data from both studies and carefully constructing appropriate probabilistic models, AIC could be used to compare competing theories regarding how flower number affects arrivals in a more general context. Hence, AIC can provide a formal framework for constructing an ecological synthesis.

How AIC ranks models depends on the amount and type of data. Simpler models often rank highly when data are scarce; however, more complex models typically improve their ranking as more data is collected, provided they incorporate important processes (Table 1, study 1). This data dependence is important when analyzing results from an AIC analysis. For example,

TABLE 2. Statistics from 2000 trials that assess the simple rule of thumb for selecting models, and the ability for model averaging to reduce model bias.

		Threshold for Δ				$E_{\mathbf{p}}\{I\}$	
r	Best K-L model	2	4	7	10	AIC best	UM
Study 1							
1	M_3	82.3 (85.4)	88.3 (97.0)	97.8 (98.1)	100.0 (100.0)	1.73	0.94
3	M_4	92.3 (77.2)	97.9 (97.6)	99.7 (99.6)	100.0 (100.0)	1.78	1.41
5	M_4	95.9 (90.3)	98.3 (98.0)	99.6 (99.6)	100.0 (100.0)	1.71	1.52
Study 2							
5	M_6	86.3 (77.4)	95.6 (86.9)	99.3 (95.8)	99.8 (99.6)	1.02	1.36
10	M_6	87.5 (79.5)	96.0 (96.0)	99.1 (96.0)	100.0 (99.2)	0.98	2.14

Notes: Values are the percentage of times that the best K-L model is retained if models were selected based on their Δ value being less than a threshold (values not in parentheses were calculated using AIC values and those in parentheses used AIC_c values). Also presented are the expected K-L distance ($E_{\mathbf{p}}\{I\}$) when the AIC best model and the unconditional model (UM) are used for prediction.

if plant replication were increased for study 2, then eventually model M_7 would be judged a better K-L model than M_6. This result would suggest that flower number affects bumble bee arrival rates, but this effect is not simply proportional to flower number. This result would not be unexpected as it is likely that the combination of many factors will make the relationship nonlinear. However, the fact that much data were required to recognize M_7 as a better K-L model reflects the effectiveness of the simpler model, M_6, at capturing much of the truth. Results from study 2 demonstrate the need for care when interpreting results from an AIC analysis; simply ignoring models with a Δ value greater than some predetermined threshold (e.g., 2) may not reflect the value of the associated hypothesis. The more data collected, the less likely a useful simple model will be judged best according to AIC.

Unavoidable sampling error means the model with the lowest AIC value is not necessarily the best K-L model. In order to incorporate this uncertainty a model selection criterion needs to be adopted. Burnham and Anderson (2002) suggest a very simple selection criterion based on model Δ values, which is supported by the results presented here. It is common for AIC analyses to retain models only if their Δ value is < 2 (e.g., Westphal et al. 2003); however, the results presented here suggest that a larger threshold may be appropriate if a probability of 0.95 or more of retaining the best model is desired. Fig. 3 and Table 1 suggest that this rule of thumb works well because the variation in AIC values among trials (not models) is relatively invariant among models and the amount of data collected. I found the standard deviation for AIC (and AIC_c) values for the best K-L model to be approximately 3 in all cases (Table 1). A very crude estimate of the probability of not correctly retaining the best K-L model can be made if we assume that AIC values among trials are normally distributed, and AIC values for each model are independent. In fact, simulations show that neither of these assumptions are quite true, as the AIC values among trials tend to have fatter upper tails and AIC values among models are slightly positively correlated. However, if the best- and second-best K-L models both have independent, normally distributed AIC values with standard deviation σ, then the Δ value of the best K-L model should also be normally distributed with a standard deviation of $\sqrt{2}\sigma$. If $\sigma = 3$ and the mean AIC of the second-best model is two units greater than that of the best K-L model, then there is a 31.9% chance that the Δ value for the best model is greater than 0. For mean AIC differences of 4, 7, and 10, the percentages of incorrectly choosing the best K-L model are 17.3, 4.9, and 0.9%, respectively. Simulations indicate that the probability the best K-L model is not retained, is about a third as often than suggested by this approximation (Table 2), mostly because the AIC values are positively correlated among models. Consistency in the variation in AIC values for the better models across all cases investigated, provides a potential clue to why the rule of thumb might work well in general. The robustness of this result needs to be investigated further.

Results from both studies showed no appreciable advantage to using the corrected version of AIC, despite the ratio of independent data to estimated parameters being low. A reason for this lack of improvement may be due to the correction factor being based on assumptions that do not match well with the foraging studies (e.g., homogeneous, normally distributed residuals). When AIC_c values are used to calculate Δ values, the rule of thumb generally performed worse than AIC (Table 2). These results suggest that data characteristics should be taken into account when considering the use of correction factors. It has been suggested that AIC might suffer from overfitting and AIC_c may alleviate this because it is more conservative (Burnham and Anderson 2002); however, overfitting was not a problem here (Table 1).

Due to computational constraints, analyses were restricted to studies involving relatively small samples.

However, the sample sizes considered here might be representative of many ecological systems where obtaining data is difficult or expensive. Relatively few data resulted in high variation among Akaike weights among trials, which limited their utility. Expected Akaike weights provided reasonable estimates of choosing the best AIC model among trials for study 1, but were poor for study 2 (Table 1). When Akaike weights were used for model averaging their ability to reduce model bias (i.e., $E_p\{I\}$) was mixed. Model averaging produced slightly better model predictions for study 1, but always resulted in worse predictions for study 2 (Table 2). In all cases, model averaging performed worse as plant replication r was increased. Bootstrapping could be used to estimate the variation in Akaike weights among repeated trials (Buckland et al. 1997), which would indicate when caution needs to be employed when using Akaike weights. These results support calls for continued research on the effectiveness of model averaging (Burnham and Anderson 2004).

Typically, attempts to evaluate the AIC approach have involved generating data from a known model and then determining the probability that the generating model is selected when it is included in a set of competing models (Anderson et al. 1998, Burnham and Anderson 2002). The advantage of this approach is that K-L distances do not need to be estimated; however, this is strictly not a true test of AIC because the best K-L model, which is what AIC seeks, may not be the model that generated the data (e.g., Table 1, study 2). It is unclear how general results regarding AIC performance are that come from studies where AIC is assessed according to criterion other than selecting the best K-L model.

The results presented here are representative of many other variations of the truth that I investigated. It is difficult to generalize too much from only two simple studies; however, results from both studies support a commonly adopted rule of thumb for retaining models based on AIC differences. On the other hand, the results suggest caution when applying correction factors, and model averaging with Akaike weights. These results highlight the need for further research on identifying conditions where AIC analyses are likely to be reliable (e.g., study design and characteristics of the data) and how analyses involving AIC should be interpreted.

Acknowledgments

The author gratefully thanks L. D. Harder, W. A. Nelson, E. A. Johnson, and reviewers for their helpful comments. A G8 legacy chair in wildlife ecology, The University of Calgary, provided funding.

Literature Cited

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. Pages 267–281 *in* B. N. Petrov and F. Csaki, editors. Second International Symposium on Information Theory. Akademiai Kiado, Budapest, Hungary.

Akaike, H. 1983. Information measures and model selection. International Statistical Institute **44**:277–291.

Anderson, D. R., K. P. Burnham, and G. C. White. 1998. Comparison of Akaike information criterion and consistent Akaike information criterion for model selection and statistical inference from capture–recapture studies. Journal of Applied Statistics **25**:263–282.

Buckland, S. T., K. P. Burnham, and N. H. Augustin. 1997. Model selection: an integral part of inference. Biometrics **53**:603–618.

Burnham, K. P., and D. R. Anderson. 2001. Kullback-Leibler information as a basis for strong inference in ecological studies. Wildlife Research **28**:111–119.

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

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

Cresswell, J. E. 1990. How and why do nectar-foraging bumblebees initiate movements between inflorescences of wild bergamot *Monarda fistulosa* (Lamiaceae)? Oecologia **82**: 450–460.

Draper, D. 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society, Series B **57**:45–97.

Dreisig, H. 1995. Ideal free distributions of nectar foraging bumblebees. Oikos **72**:161–172.

Fernández, N., M. Delibes, F. Palomares, and D. J. Mladenoff. 2003. Identifying breeding habitat for the Iberian lynx: inferences from a fine-scale spatial analysis. Ecological Applications **13**:1310–1324.

Fujiwara, M., and H. Caswell. 2001. Demography of the endangered North Atlantic right whale. Nature **414**:537–541.

Godínez-Domínguez, E., and J. Freire. 2003. Information-theoretic approach for selection of spatial and temporal models of community organization. Marine Ecology Progress Series **253**:17–24.

Gurney, W. S. C., and R. M. Nisbet. 1998. Ecological dynamics. Oxford University Press, New York, New York, USA.

Helu, S. L., D. B. Sampson, and Y. Yin. 2000. Application of statistical model selection criteria to the Stock Synthesis assessment program. Canadian Journal of Fisheries and Aquatic Sciences **57**:1784–1793.

Hilborn, R., and M. Mangel. 1997. The ecological detective: confronting models with data. Princeton University Press, Princeton, New Jersey, USA.

Hilborn, R., and C. J. Walters. 1992. Quantitative fisheries stock assessment: choice, dynamics, and uncertainty. Chapman and Hall, New York, New York, USA.

Hurvich, C. M., and C-L. Tsai. 1989. Regression and time series model selection in small samples. Biometrika **76**: 297–307.

Kadmon, R. 1992. Dynamics of forager arrivals and nectar renewal in flowers of *Anchusa strigosa*. Oecologia **92**:552–555.

Klinkhamer, P. G., and T. J. de Jong. 1990. Effects of plant size, plant density, and sex differential nectar reward on pollinator visitation in the protandrous *Echium vulgare* (Boraginacae). Oikos **57**:399–405.

Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. Annals of Mathematical Statistics **22**:79–86.

Leirs, H., N. C. Stenseth, J. D. Nichols, J. E. Hines, R. Verhagen, and W. Verheyen. 1997. Stochastic seasonality and non-linear density-dependent factors regulate population size in an African rodent. Nature **389**:176–180.

Mitchell, R., J. D. Karron, K. G. Holmquist, and J. M. Bell. 2004. The influence of *Mimulus ringens* floral display size on pollinator visitation patterns. Functional Ecology **18**: 116–124.

Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criteria and Bayesian approaches over likelihood ratio tests. Systematic Biology **53**:793–808.

Possingham, H. P. 1992. Habitat selection by two species of nectarivore: habitat quality isolines. Ecology **73**:1903–1912.

Pyke, G. H. 1982. Foraging in bumblebees: rule of departure from an inflorescence. Canadian Journal of Zoology **60**: 417–428.

Westphal, M. I., S. A. Field, A. J. Tyre, D. Paton, and H. P. Possingham. 2003. Effects of landscape pattern on bird species distribution in the Mt. Lofty Ranges, South Australia. Landscape Ecology **18**:413–426.

Zimmerman, M. 1983. Plant reproduction and optimal foraging: environmental nectar manipulations in *Delphinium nelsonii*. Oikos **41**:57–63.