Behav Ecol Sociobiol (2011) 65:77–89
DOI 10.1007/s00265-010-1035-8

ORIGINAL PAPER

Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework

Shane A. Richards · Mark J. Whittingham · Philip A. Stephens

Received: 22 April 2009 / Revised: 14 July 2010 / Accepted: 29 July 2010 / Published online: 19 August 2010

Abstract Behavioural ecologists often study complex systems in which multiple hypotheses could be proposed to explain observed phenomena. For some systems, simple controlled experiments can be employed to reveal part of the complexity; often, however, observational studies that incorporate a multitude of causal factors may be the only (or preferred) avenue of study. We assess the value of recently advocated approaches to inference in both contexts. Specifically, we examine the use of information theoretic (IT) model selection using Akaike's information criterion (AIC). We find that, for simple analyses, the advantages of switching to an IT-AIC approach are likely to be slight, especially given recent emphasis on biological rather than statistical significance. By contrast, the model selection approach embodied by IT approaches offers significant advantages when applied to problems of more complex causality. Model averaging is an intuitively appealing extension to model selection. However, we were unable to demonstrate consistent improvements in prediction accuracy when using model averaging with IT-AIC; our equivocal results suggest that more research is needed on its utility. We illustrate our arguments with worked examples from behavioural experiments.

Keywords Effect size · Inference · Model weighting · Null hypotheses · Process-based models · Statistics

Communicated by: L. Garamszegi

This contribution is part of the Special Issue: "Model selection, multimodel inference and information-theoretic approaches in behavioural ecology" (see Garamszegi 2010).

S. A. Richards · P. A. Stephens (✉)
School of Biological and Biomedical Sciences,
Durham University,
South Road,
Durham DH1 3LE, UK
e-mail: philip.stephens@durham.ac.uk

M. J. Whittingham
School of Biology, Newcastle University,
Newcastle upon Tyne NE1 7RU, UK

Introduction

The behaviour of living organisms is the outcome of "intricate and sophisticated mechanisms involving sensory, neural, endocrine and cognitive structures, and active interactions with genes" (Ydenberg et al. 2007, p19). Unsurprisingly, therefore, the science of behaviour is complex; causal studies are hampered by multiple potential confounds within a vast web of cause and effect. One solution to this problem is to use carefully controlled laboratory experiments to isolate the causal factors and responses of interest. Such elaborate and ingenious experiments have long been a hallmark of the science of behavioural ecology. Nevertheless, the continuum from tightly controlled experiments to field observation is not a one-way descent from desirable to less desirable (e.g. see Ruxton and Colegrave 2006). Indeed, experimental designs are plagued by the possibility that behaviour in highly unnatural laboratory situations might be a poor reflection of reality (e.g. Heyes and Dawson 1990; Mitchell et al. 1999). Thus, behavioural ecology exploits the full range of data collection opportunities, from field observation through 'natural experiments' and field manipulations to laboratory experiments. Increasingly, ecologists are considering how to make the most effective and robust inferences from data collected using this wide range of methods (e.g. Nakagawa and Cuthill 2007; Stephens et al. 2007b).

One approach to inference that has been widely promoted in recent years is the use of information theoretic (IT) approaches for model selection (e.g. Hilborn and Mangel 1997; Burnham and Anderson 2002; Johnson and Omland 2004; Garamszegi 2010). IT methods represent a rather different approach to inference than that characterised by the classical, null hypothesis significant testing (NHST) paradigm that prevails throughout behavioural ecology. In particular, IT methods stem from a recognition that data seldom provide absolute support for a single hypothesis; rather, data can only influence the extent to which we feel that any given hypothesis is supported (relative to competing explanations). This more Bayesian viewpoint contrasts with NHST approaches, which focus exclusively on the plausibility of the null hypothesis and purport to use data either to falsify that null hypothesis (leaving its alternative open as a 'working explanation' for the phenomenon of interest), or to fail to reject the null (traditionally interpreted to imply that its alternative is not well supported as an explanation for the focal phenomenon).

In spite of the extent to which IT methods have been promoted and NHST methods criticised in recent years, there is limited evidence to suggest that behavioural ecologists are beginning to embrace IT at the expense of NHST. Stephens et al. (2007b) presented a coarse assessment of the use of different statistical approaches in various journals at the end of 2005. One of those journals was *Behavioral Ecology*; of the last 50 data-driven articles published in that journal in 2005, all but one had employed NHST, whilst only five (10%) had employed an IT approach. We repeated the assessment, looking at the first 50 data-driven articles published in *Behavioral Ecology* in 2008. Of those 50, all had used NHST, and only six (12%) had used IT analyses. In contrast, equivalent data are available for the *Journal of Applied Ecology*. During the same period (end of 2005 to beginning of 2008), the proportion of 50 sampled papers using NHST had reduced from 90% to 80%, whilst the proportion using IT approaches had increased from 6% to 20%. These data are obviously coarse and are themselves vulnerable to questions regarding statistical significance. Nevertheless, they are suggestive that behavioural ecologists currently make less use of some recently advocated statistical techniques.

The apparent reluctance of behavioural ecologists to abandon traditional approaches to statistical analysis was a major motivation for the editors of this special issue (see also Garamszegi 2010). We believe that two principal reasons underlie the lack of enthusiasm for IT among behavioural ecologists. First, behavioural ecologists are usually strong proponents of simple, controlled experiments. Whilst the advantages of IT in more complex settings are well recognised (e.g. Stephens et al. 2005; Whittingham et al. 2006), the advantages for simpler problems with single variable causation have received less attention (but see Lukacs et al. 2007). Second, adapting to a new statistical paradigm requires effort. Only if the advantages are made explicitly clear in a behavioural ecology context, is a substantial shift likely within the field..

In response to the perceived limitations on the uptake of IT that we have detailed, we briefly present what we see as the potential benefits that an IT-based analysis can provide for behavioural ecologists. For studies of single parameter causation, such as those conducted using simple experiments, IT represents a more appealing philosophical viewpoint than NHST; nevertheless, the interpretations will typically be similar, regardless of the approach used. Moreover, the appropriate emphasis in such studies should generally be on the size of an effect, rather than on the support for models that specify that an effect does or does not exist (Nakagawa and Cuthill 2007). As such, the benefits of IT for behavioural ecologists involved exclusively with simple experimentation are unlikely to be sufficient to encourage the shift to working within a new paradigm. By contrast, we believe that IT approaches really come into their own when more complex studies are considered. In particular, we focus on what we perceive to be the main benefit of IT over NHST: specifically, that IT-based model selection is well suited for inferring biological mechanisms. Here, we focus solely on Akaike's information criterion (AIC) as a metric of model performance. We note that Bayesian approaches can also be used to compare models, conduct model averaging and infer biological mechanism (McCarthy 2007). Bayesian techniques are fundamentally different in their approach to data and in their assumptions regarding the potential for researchers to propose 'true' models. Owing to these differences, covering Bayesian approaches to model selection is not within the scope of this treatment. In this paper, however, we do examine the utility of model averaging (within the IT-AIC framework) for improving inference. Model averaging has been advocated as an advantage of IT-AIC approaches (Burnham and Anderson 2002) despite wide acknowledgment that more work on model averaging is required (Buckland et al. 1997; Burnham and Anderson 2002, pp. 152–3; Richards 2005; Freckleton 2010; Nakagawa and Freckleton 2010). Examples are provided that illustrate key concepts and support our recommendations.

Analysing data using AIC

Often, when behavioural ecologists are interpreting their data, they consider a number of biological mechanisms (or hypotheses) to be plausible explanations of their data.

Typically, each hypothesis can be translated into a clear model with mathematical expression; the problem is to determine which model or models are the most parsimonious description of the data. There are many ways to define model parsimony, which quantifies the trade-off between fitting the current data well (including any outliers that may be present) and having confidence that the fitted model would accurately predict similarly collected data. In this paper, we focus on IT methods that use AIC as a metric of model parsimony. Burnham and Anderson (2002) and Richards (2005) provide details regarding parsimony in this context.

From here on, we assume that the reader has an understanding of how AIC values can be calculated (e.g. Anderson et al. 2000; Burnham et al. 2010). Key steps in the analytical process are to calculate an AIC value for each model proposed, to examine the differences between the AIC values of competing models, and to discard those models associated with very poor support relative to their competitors. On the basis of available data, the model having the lowest AIC value (often referred to as the AIC best model) may be deemed the most parsimonious model; however, as AIC is only an estimate of model parsimony, another model having a higher AIC value may in fact be more parsimonious. As a consequence, we cannot afford to focus interpretation exclusively on the AIC best model; rather, we should acknowledge that a number of models may share similar levels of support. These models are often termed the "confidence set", and various selection rules have been proposed to define which models should be included (Burnham and Anderson 2002, pp170–1). In practise, cut-off rules are often adopted to determine the confidence set. Cut-offs are usually based on model Δ values defined as the difference between the AIC values of the focal model and the AIC best model. Many published model selection studies use a cut-off of $\Delta \leq 2$, to include only those models with 'substantial support' from the data (Burnham and Anderson 2001, p114). However, simulations suggest that this rule is too stringent, and a cut-off threshold of $\Delta \leq 6$ is often necessary to be 95% sure that the truly most parsimonious model is retained within the confidence set (Richards 2005, 2008). However, even this second rule can often lead to retention of overly complex models (Richards 2008). An alternative model selection approach that reduces the retention of overly complex models is to exclude from the candidate set those models that are more complicated versions of any model with a lower AIC value (Burnham and Anderson 2002, p131; Richards 2008). In this context, models are described as "nested": for example, if model A contains all the parameters of model B, as well as one or more additional parameters, then model B can be said to be 'nested' within model A (and, conversely, model A can be described as a more complex version of model B).

A conservative approach is to base inference principally upon the simplest models in the confidence set, and to infer marginal support for the importance of factors present in the more complex models (but absent from the simpler models) in the confidence set. A more appealing approach is to associate with each model the probability that it is the best model (i.e. is the most parsimonious model; Anderson et al. 2000, p918). Burnham and Anderson (2002) and Lukacs et al. (2007) suggest that Akaike weights can be used to estimate these probabilities, where the weight (or probability) associated with model M_i is given by

$$w_i = \frac{\exp(-\frac{1}{2}\Delta_i)}{\sum_j \exp(-\frac{1}{2}\Delta_j)} \quad (1)$$

where Δ_i is the Δ value associated with model M_i. Support for one model over any other model considered can be estimated using evidence ratios, which are simply the ratio of the model weights of the two models being compared.

As we note in the Introduction section, AIC analyses seldom result in unequivocal support for a single model (i.e. rarely will a single model weight be approximately one). To account for model uncertainty, Burnham and Anderson (2002) and Lukacs et al. (2007) also suggest that within the IT-AIC framework, model weights can be used to make formal inferences from multiple models simultaneously. For example, model averaged predictions can be calculated by weighting each model prediction by its probability. The idea here is that explicitly accounting for model uncertainty will result in more reliable predictions. If all the models proposed are linear and nested, then it has been proposed that model parameters can also be model-averaged (Burnham and Anderson 2002). Lukacs et al. (2007) argue that model weights (and their ability to account explicitly for model uncertainty) are major reasons why IT-AIC approaches should be favoured highly over NHST.

Why behavioural ecologists should consider the IT approach

For ecology in general, the case for caution in the use of NHST and the potential of IT as an alternative has been made elsewhere (Cohen 1994; Johnson 1999; Anderson et al. 2000; Johnson and Omland 2004; Nakagawa and Cuthill 2007), in some cases, quite forcibly. We will not dwell further on the general problems with NHST. Rather, in this section, we consider three examples of areas in which it has been suggested that IT approaches (in preference to NHST) could be particularly useful to behavioural ecologists. Specifically, we discuss the utility of IT approaches in the case of tightly controlled experiments designed to test the effect of only one or two treatments, the utility of IT

approaches when data come from observational studies (where many factors are beyond the researcher's control), and the potential of IT approaches to reveal process and mechanism in behaviour.

IT and controlled experiments

As we have already observed, behavioural ecologists tend to rely heavily on controlled experiments to test hypotheses. This is borne out by the coarse assessment of journal articles presented in Stephens et al. (2007b). The proportions of 50 data-driven papers that incorporated some element of experimental (rather than sampling) design were: 62% (*Behavioral Ecology*), 52% (*Ecology Letters*), 44% (*Evolution*) and 24% (*Journal of Applied Ecology*). The equivalent figures in our more recent assessment (from early 2008) were 68% (*Behavioral Ecology*) and 22% (*Journal of Applied Ecology*). The type of experiments used are often designed to minimise the number of factors that can vary between treatments, often restricting required analyses to simple statistical tests, such as *t* tests, χ^2 tests or one- or two-factor ANOVAs.

Two of us (PAS & MJW) have, in the past, argued that NHST is adequate for analysing the outcomes of tightly controlled experiments (Stephens et al. 2005; Whittingham et al. 2006). However, others have countered that, whilst that might be true, IT methods are also appropriate and, moreover, may be more informative than their NHST equivalents (Lukacs et al. 2007). In many cases, the inferences drawn will be very similar, regardless of the approach used (Stephens et al. 2007b) but, undeniably, the information gained about the alternative hypothesis differs according to whether IT or NHST is used. In particular, Lukacs et al. (2007, Table 1) list several advantages of an IT approach over an NHST approach. Perhaps the most important of these is that the *p* value obtained using NHST gives only the probability that the data (or data yielding a more extreme test statistic) would have been obtained, *if the null hypothesis was true*. In that way, NHST says nothing about the extent to which the data support the alternative hypothesis (Cohen 1994). By contrast, the IT approach yields a direct estimate (the "model weight") of the relative support for each hypothesis (null and alternative). Thus, the IT approach asks only which of the two models (null or alternative) is better supported by the data. The latter approach seems sensible given that, in many cases, we would not expect a treatment to have absolutely no effect on the outcome of an experiment.

In addition, Lukacs et al. (2007) note that whilst NHST yields nothing more than the *p* value dependent on the null hypothesis, IT approaches also yield evidence ratios and, via model averaging, less biased parameter estimates and unconditional estimates of precision. Evidence ratios are an easily interpretable metric of the relative extent to which the data support two competing models; however, model averaging is less straightforward in the advantages that it offers, particularly in the case of simple experiments. We consider purported advantages of model averaging further in the examples that follow. Broadly, however, there are advantages to considering what the data tell us about the relative merits of both the null and its alternative, rather than considering only what the data tell us conditional on the null being true. In particular, this gives some support to the contention of Lukacs et al. (2007): that IT methods have value even in the analysis of tightly controlled experiments (see Behavioural questions and model selection: some examples section below).

IT and observational studies

Many of the most informative studies in behavioural ecology are field studies in which experimentation is impossible or is only limited in scope. For example, as outlined by Lind and Cresswell (2005), there are a multitude of ways for an animal to increase anti-predation behaviour (for example, by reducing foraging time, increasing vigilance, decreasing body mass, or foraging in a less risky habitat) and so looking at changes in any one behaviour in a laboratory study can only ever inform a small part of the overall picture.

Observational data are often collected in combination with large numbers of potentially explanatory factors. A common approach to reducing the complexity of the resulting data set is to analyse it with stepwise multiple regression. The problems inherent to stepwise regression are well known in the statistical literature and have recently been explored in an ecological context (Whittingham et al. 2006; Hegyi and Garamszegi 2010). In brief, the main problems of stepwise regression are: (1) a bias in parameter estimation (leading to either an over or underestimation of parameter values), (2) inconsistencies in model selection algorithms (e.g. forwards and backwards selection methods can yield different answers), (3) an inherent problem with multiple hypothesis testing within the framework of a single analysis, and (4) an inappropriate focus or reliance on a 'best' model. This latter point is examined in detail by Whittingham et al. (2006) who show that Minimum Adequate Models derived from four different years worth of observational data are each distinct, but, in no individual case, provide an accurate description of the data. Further problems of stepwise approaches are that they can only be used to choose among a set of completely nested models and that they seldom deal well with collinear variables (Freckleton 2010). IT-AIC approaches overcome several of the limitations of stepwise regression, but the latter technique remains in widespread use in behavioural

ecology. We contend that, for the reasons discussed in Whittingham et al. (2006), behavioural ecologists should pay more attention to IT as an alternative to classical stepwise techniques. It is important to note that stepwise procedures should not be used with AIC, and that relying on a single best model using IT-AIC is counter to the philosophy of the IT-AIC approach (as outlined by Burnham and Anderson 2002).

Process and mechanism in behaviour

Perhaps the greatest advantages of IT methodologies are those arising from their flexibility, wherever an appropriate likelihood function can be specified, to incorporate complex or unconventional model structures. The complex nature of animal behaviour has resulted in a tradition of mathematical models being developed to formalise proposed behavioural mechanisms in order to understand better and predict animal feeding, survival and mating strategies (e.g. Krebs and Davies 1978). Typically, these mechanistic models predict non-linear relations between variables which can pose problems for many NHSTs that require relations to be linear. In some cases, data transformations can make relations linear; however, often this is not the case and non-linear techniques must be considered (Bolker 2008). Furthermore, transformations that linearise data may not result in homoscedasticity, which may also affect the reliability of many NHSTs. Instead of massaging data to suit a NHST, it is preferable to fit the raw data directly to the proposed models.

In addition to predicting the mean relation between variables, well-posed mechanistic models also predict the nature of variation about the mean (Hobbs and Hilborn 2006). Such models often naturally lead to the formulation of model likelihoods (Hilborn and Mangel 1997; Richards 2005) which are directly used by IT approaches. Thus, IT approaches may offer a significant advantage to behavioural ecologists, as they can easily incorporate process-based models of animal behaviour (in a way that NHST approaches with frequently rigid data requirements often cannot).

Behavioural questions and model selection: some examples

We have argued that IT techniques may have value in a range of situations encountered by behavioural ecologists. For behavioural ecologists contemplating analyses of observational data, we suggest that Whittingham et al. (2006) will, at least, provide some points for consideration. By contrast, in the cases of tightly controlled experiments and process-based model fitting, our suggestions may be more easily understood with reference to some behaviourally motivated examples. In this section, we present two simple examples illustrating the use of AIC and enabling discussion of the potential advantages of the IT approach. In particular, due to its intuitive appeal, we address the utility of model averaging. Our discussion recognises that model averaging remains an open research area, with further work required to understand fully its properties and limitations (Burnham and Anderson 2002, pp152–155; Stephens et al. 2007a; Freckleton 2010; Nakagawa and Freckleton 2010).

A simple controlled experiment on fish behaviour

As an example of using IT to analyse the outcome of a controlled experiment, consider the simple case of data appropriate to a *t* test. A researcher wishes to see whether the time taken to emerge from cover when threespine sticklebacks (*Gasterosteus aculeatus*) are released into a tank is influenced by whether the tank has previously housed a predator (pike, *Esox lucius*). Of 50 sticklebacks, 25 are randomly assigned to the control group (to be placed in tanks that had not formerly housed a predator) and the remainder to the treatment group (to be placed in tanks formerly occupied by a pike). Experimental subjects are placed into an opaque chamber with a single exit within the main tank, and the tank is filmed from above. Video analysis enables the researcher to determine the exact time from placement of the fish in the chamber to the fish being entirely outside the chamber. The outcome of the experiment is shown in Fig. 1. The means are suggestive of a difference caused by the treatment, but the outcome of a 2-sample *t* test gives $t_{49}=-1.98$, $p=0.053$ (the same inference could also be reached by considering the confidence interval about the estimated effect size). Strictly, we have failed to reject the null, that the previous presence of a predator has no effect. Most practitioners, however, are likely to accept that the data demonstrate a ‘tendency’ for greater caution among sticklebacks placed in tanks that have previously contained a pike.

What do we infer from adopting the IT approach using AIC to the same problem? AIC scores for the two models (generated using the glm function in R; R Development Core Team 2005) and related information are given in Table 1. Note that the IT analysis shows that the data obtained in the experiment actually provide slightly more support for the alternative hypothesis than for the null. This is unsurprising as, in fact, we might have expected the potential perceived presence of a predator to make a difference to stickleback behaviour. The evidence ratio in this case is $E=0.72/0.28=2.6$. An evidence ratio of less than 10 is usually taken to imply that the data provide limited support for that model (Lukacs et al. 2007); hence, we conclude limited support for the alternate hypothesis.

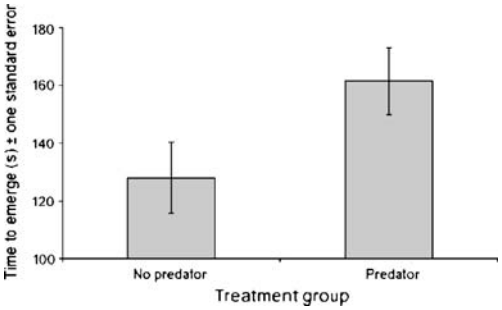

Fig. 1 Outcome of an experiment to determine whether the previous presence of a predator has any effect on the time taken for a stickleback to emerge from cover. See main text for further details

As we noted above, the take-home message from the two types of analysis can often be quite similar: in this case, that the data are suggestive of an effect, but that the effect is not strong relative to other causes of variance in the system. Importantly, however, the IT approach allows us to say something about the plausibility of the alternative hypothesis (namely that it is 2.6 times more plausible, given the data, than the null hypothesis); no such inference is possible using a NHST approach. Whether this in itself is sufficient incentive for researchers to adopt the IT approach for such analyses, remains to be seen. However, additional benefits have also been proposed, as we now discuss.

Lukacs et al. (2007, p.459) recommend the use of model averaging even in the case of investigations to assess the effect of a single parameter. For our stickleback example, we can calculate a model averaged estimate of the effect size (i.e. the mean increased time to emergence due to the previous presence of a predator). The means of the two groups were 128.0 s (for the control) and 161.4 s (for the treatment). The observed treatment effect was, thus, 33.4 s. The model weight for the alternate hypothesis was 0.72 (see Table 1) and, thus, the model averaged effect size was $33.4\times0.72+0\times0.28=24.08$ s. To determine whether model averaging provided a better estimate of the true effect size (compared with simply using the effect size predicted by the AIC best model, which was the alternate model), we would need to know that true effect size (which, of course, we cannot know). However, two approaches allow us to assess whether model averaging in these situations is likely to be useful, as follows.

Table 1 Applying an IT approach to data on the *t* test example

Model	K[a]	n[b]	AIC	Δ[c]	w[d]	R^2
Null ($\mu_1=\mu_2$)	2	50	556.7	1.9	0.28	–
Alternative ($\mu_1\neq\mu_2$)	3	50	554.8	0.0	0.72	0.076

Note that data typically given in addition to the AIC score are indicated by the table footnotes

[a] Number of estimated parameters

[b] Sample size

[c] The difference between the current model's AIC score and that of the model with the lowest AIC score

[d] Model weight

First, for the type of example we depict, it is generally acknowledged that there will unquestionably be a difference in the means of the two groups (Martinez-Abrain 2007) and assumed that data in the two samples are drawn from normal distributions. It follows that the probability with which sampling error leads to an overestimate of the true difference between the means of the two groups is equal to the probability of underestimating that difference. If, by chance, we underestimate the difference, the weight accorded to the null model will be high, and model averaging will result in us exacerbating that underestimate (by multiplying the observed difference by the relatively low model weight of the alternate hypothesis). If, by contrast, the data sample leads to an overestimate of the true difference between means, the model weight accorded to the alternate hypothesis will be very high; in this case, model averaging will do little to reduce the bias. Thus, underestimates will be exacerbated and overestimates will be relatively unaffected. The exception to this will be when model selection suggests that the null is the better model. In this case, if the alternate hypothesis has any weight of evidence, the error in accepting a zero effect size (as indicated by the null) will be mitigated for by model averaging. This thought experiment suggests that model averaging will, in general, reduce the accuracy with which we can estimate the true effect, unless the effect is sufficiently weak or the sample size sufficiently small that the null is often chosen as the better model.

The reasoning above is supported by a simple simulation study. Specifically, we randomly generated data drawn from two normal distributions with the same variance but a nominal difference between means of one (i.e. $\delta_{true}=1$). For selected sample sizes (ranging from 20 to 100) and standard deviations (of up to three times the nominal difference between means), we generated 1,000 Monte Carlo replicates and determined the AIC scores for the null and alternate models. From those, we determined model weights, as well as the estimated effect size using both the model with the lowest AIC score, δ_{best}, and using model averaging (a weighted average of the effect sizes suggested by the two models), δ_{av}. For each replicate, we calculated both $(\delta_{best}-\delta_{true})^2$ and $(\delta_{av}-\delta_{true})^2$. The means of these over all replicates gave the mean squared error arising from the two approaches (MSE_{best} and MSE_{av}). We used the ratio MSE_{av}/MSE_{best} to indicate which approach generated more error, on average. Clearly, a ratio of greater than 1 indicates that model averaging is less accurate, on average, than using the AIC best model for effect size estimation.

The results of our simulation are shown in Fig. 2. Panel A of the figure shows the proportion of Monte Carlo data sets that gave greater support to the alternate hypothesis than to the null. Clearly, for situations in which the standard deviation of the underlying distributions is of a similar (or lower) magnitude to the mean difference between them, the alternate is almost always chosen as the AIC best model. For greater standard deviations, this starts to decline and, unsurprisingly, the speed of the decline is related to the sample size. With a standard deviation of three times the effect size, sample sizes of 100 led to the alternate hypothesis receiving greater support over 80% of the time whilst, for sample sizes of 20, that figure was reduced to 36%. The consequences of model averaging for the accuracy with which the effect size is estimated are shown in panel B of the figure. It is clear that the suggestions arising from our thought experiment above are borne out by this analysis. Specifically, for the case where there is a true difference between the means of the distributions from which the two samples were drawn, estimating that difference on the basis of the AIC best model alone is likely to be more accurate, on average, than using model averaging. Only when (a) the standard deviation is substantially larger than the effect size and (b) the sample size is very low, is it likely that model averaging will reduce the errors resulting from effect size estimation. In such cases (i.e. those with a standard deviation of greater than 2.5 times the effect size), it is likely that pilot data would prompt the researcher to use sample sizes greatly in excess of 20, reducing the likelihood of finding themselves within the region of Fig. 2b within which model averaging is worthwhile.

In spite of this reasoning, there may be reasons to prefer the use of model averaging. In particular, as Burnham and Anderson (2002) note, model averaging provides a more stable inference than accepting the AIC best model. This can easily be envisaged using our simple example (Fig. 2c, d). The figures show that using the best model, inference regarding the effect size is highly contingent on which model is selected as best. Thus, inference is effectively bimodal, flipping from one mode to the other, depending on the sample to hand (a problem identical to that arising from the arbitrary use of NHST p values; Whittingham et al. 2006). By contrast, using model averaging removes the instability arising from model misspecification (Burnham and Anderson 2002). The single parameter example is relatively trivial in this regard, not least because the emphasis ought, in any case, to be on effect size estimation (Nakagawa and Cuthill 2007). Nevertheless, it enables us to illustrate some of the considerations involved with model averaging before discussing how the technique can be applied to more complex scenarios.

A more complex, controlled foraging experiment

We have suggested that an IT analysis can be used in a behavioural ecology context to infer important mechanisms

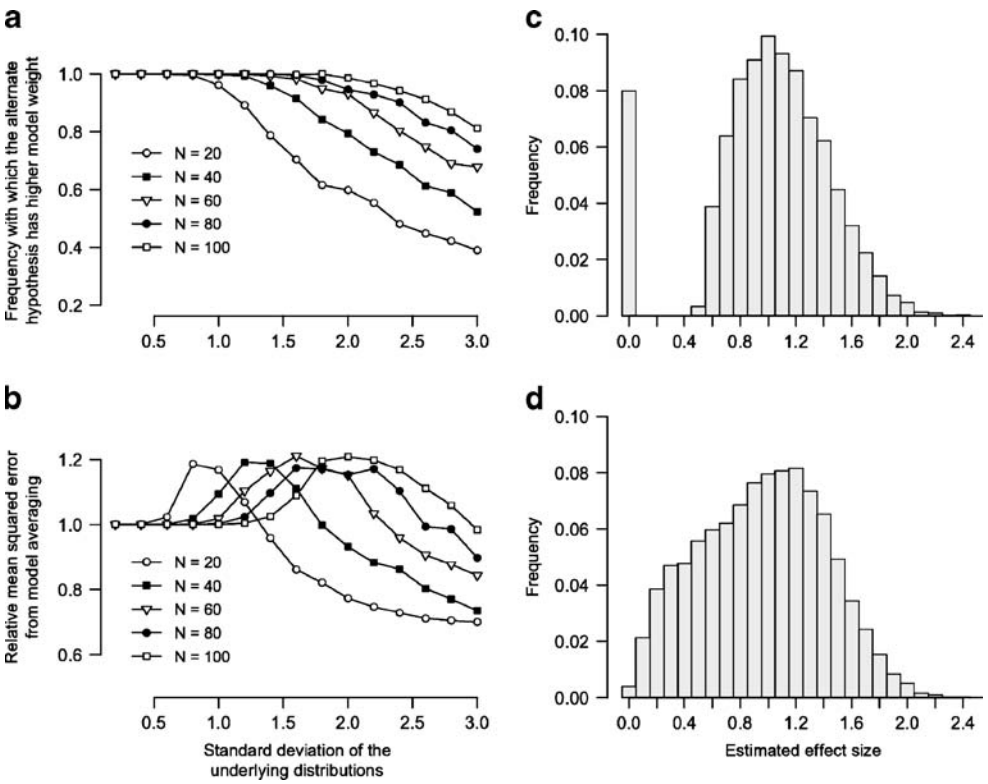

Fig. 2 Utility of model averaging in the case of an experiment on one-parameter causation. **a** Frequency with which the data supported a lower AIC score for the alternate hypothesis ($\mu_1 \neq \mu_2$) than for the null ($\mu_1 = \mu_2$). **b** Mean squared error of the model averaged effect size relative to the AIC best effect size. Legends show the size of each sample, n. Note that values greater than 1 indicate that model averaging tends, on average, to be more erroneous than using the AIC best model (see main text for further details). **c**, **d** Histograms showing the effect sizes estimated in 10,000 replicates when $\mu_1 - \mu_2 = 1$ (with standard deviation of samples = 2.0) and $n = 60$ using the AIC best model (**c**) and model averaging (**d**)

reliably and, here, we present a hypothetical study to illustrate that point. Specifically, we illustrate how IT-AIC can account for data which, based on biologically informed hypotheses, are not expected to exhibit either linear relations or normally distributed variation. Suppose an ecologist was interested in how environmental conditions affect foraging success; in particular, whether soil moisture affects either seed encounter rate or seed handling time for a small rodent. Soil moisture could increase the release of olfactory signals by seeds in soil, thereby increasing the chance they are encountered by seed foragers (Vander Wall 2000). Soil moisture could also change physical characteristics of the seeds and influence seed handling time. To test the effect of soil moisture, individual foragers were placed in an enclosure where the initial density of seeds was set at one of five equally spaced levels (seed density treatment), and the soil in the enclosure was either dry or wet (soil moisture treatment). Suppose $n=5$ randomly chosen individuals were allocated to each unique seed density and soil treatment pairing and, after time T, the number of seeds consumed (y) was noted for each forager. Thus, the data from the experiment consist of 50 independent triplets (x, i, y), where x is the initial density of seed, i = D or W (indicating dry or wet soil, respectively), and y is the number of seeds consumed. If a statistical test were performed on these data, then seed density would be a covariate, and soil moisture would be a fixed factor with two levels.

A mechanistic-based analysis of the above-mentioned data can be informed by the functional response which describes the relation between food density and food intake rate. Suppose a forager finds food items when searching for them at rate s and the time a forager takes to handle food items when they are found is h. Assuming no new food items can be searched for when food items are handled and consumed, the expected intake rate (items consumed per unit time) when the density of food is x, is given by

$$f(x|s,h) = \frac{sx}{1+hsx} \tag{2}$$

(Gurney and Nisbet 1998). If individuals are presented with food, initially at density x, and allowed to forage for time T, then provided depletion of food is relatively low, the expected number of items consumed by the forager will be approximately $Tf(x)$. If resource depletion is substantial, then mean consumption could be modelled by solving numerically the non-linear equation provided by Rogers (1972). Assuming resources are encountered randomly, the actual number of food items consumed may be expected to follow a Poisson distribution (PD). However, if seeds are non-randomly distributed in space or variation exists among foragers (e.g. forager size, experience, and hunger state), then the variation in foraging success may be greater than that predicted by a PD (i.e. the variance may be greater than the mean); such data are said to be overdispersed with respect to the PD. A flexible approach to describing variation of a continuous characteristic among individuals is the gamma distribution. If variation in mean foraging success among individuals can be described by a gamma distribution, then the variation in consumption among individuals will follow a negative binomial distribution (NBD) (Bolker 2008; Richards 2008). Thus, in this case, the NBD may be thought of as a model that implicitly encapsulates many unknown or unmeasured factors that contribute to an individual's foraging success.

In this example, it makes biological sense to propose that soil moisture could affect either s or h; hence, models can be developed based on one of four assumptions. Let (−), (s), (h) and (s + h) denote the assumptions that neither s nor h, only s, only h, and both s and h are affected by soil moisture, respectively. Combining these four assumptions with the two assumptions regarding the distribution of data about the mean (PD or NBD), gives eight possible models. For example, NBD(h) denotes the assumption that soil moisture only affects mean seed handling time, and there is considerable unknown variation among individuals in their foraging abilities.

We simulated 100 experimental data sets using model NBD(s) as the generator (see Appendix for more details). True parameter values were set to $s_D=0.5$, $s_W=1$, and $h_D=h_W=h=0.5$; hence, we considered the situation where drying the soil halved the rate at which seeds were found but did not affect seed handling time. We also assumed that the duration of the experiment was $T=20$ time units and $\phi=0.075$ which resulted in marginally overdispersed data [i.e. on average, the variation in the data was $(1+\phi T)=2.5$ times that expected according to a PD; see Appendix]. The density of resources among treatments was set to $x=1, 2,...5$ resource units. Note that although we generated data using a model proposed by the modeller, in reality, we would often not expect the true model to be one that is proposed (Burnham and Anderson 2002; Richards 2005). By setting the truth as a model that is not the most complex proposed, namely model NBD(s + h), this example allows us to investigate the propensity for different selection rules to select this overly complex model.

The 100 data sets produced by the model NBD(s) were characterised by (1) low to moderate counts, (2) counts that overlapped between the two soil treatments, (3) counts that suggested a slight asymptotic relation, and (4) wet soil counts tending to be higher across all seed densities (Fig. 3a). Model NBD(s) most often had the lowest AIC value when all eight models were included in the analysis (61%), followed by model NBD(h) (30%; Table 2). Hence, the AIC analysis was regularly able to detect overdispersion in the data and most often correctly concluded that the best

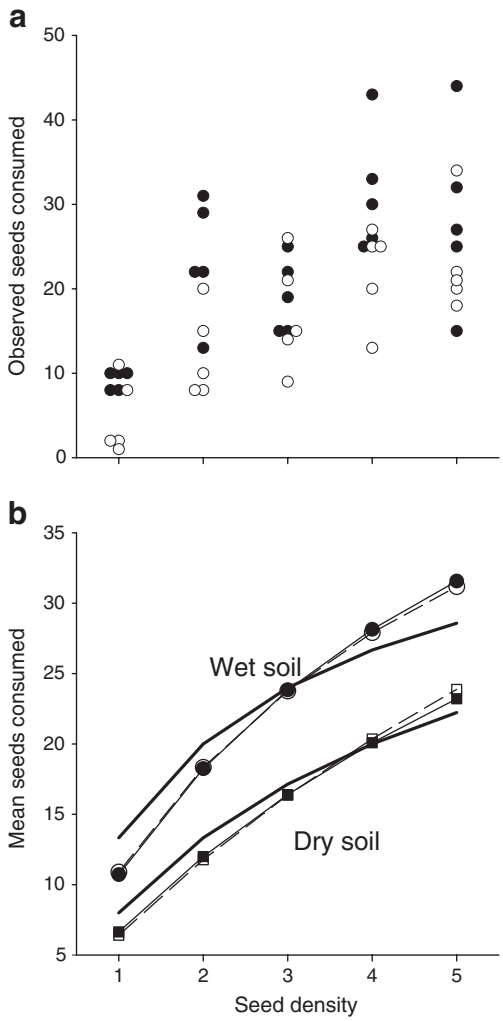

Fig. 3 An example of a typical data set for the hypothetical foraging experiment discussed in the main text. **a** Raw counts of seeds consumed (dry and wet soil replicates are indicated by *open* and *closed circles*, respectively). **b** The true mean seed encounter rates that generated the data (*solid lines*), the estimated mean encounter rates according to the model having the lowest AIC value (*closed symbols*) and model averaged expectations (*open symbols*). *Upper three curves* relate to the wet soil treatment and the *lower three curves* relate to the dry soil treatment

explanation was that soil treatment only affected seed encounter rate. However, nearly a third of analyses incorrectly concluded that the best explanation was that soil moisture only affected seed handling time. Rarely was the most complex model NBD(s + h) considered the AIC best model (7%).

Selecting all models with $\Delta_i \leq 2$ increased the chance of selecting model NBD(s) to 85% but also increased the chance of selecting the model NBD(s + h) to 99%. Increasing the threshold to $\Delta_i \leq 6$ increased these two probabilities to 97% and 100% (Table 2). However, adopting the nesting selection rule resulted in selecting the model NBD(s) 95% of the time and restricted the selection of the unnecessarily complex model NBD(s + h) to 7%. Hence, the nesting rule reduced the chance of selecting overly complex models without severely compromising the chance of selecting the most parsimonious model. The nesting rule also reduced the average number of models chosen from 3.30 to 2.33 (Table 2). These results indicate that, for this experiment, there is unlikely to be sufficient information in the data to distinguish strongly between models NBD(s) and NBD(h).

Analyses of the simulated data sets also highlighted the propensity for AIC to select overly complex models if the variation in the data was incorrectly modelled. When overdispersion was ignored and only Poisson models were fitted to the data, the model PD(s) was most often the AIC best model (56%); thus, the correct inference that only seed encounter rate was affected by soil moisture was reduced by 5% (Table 2). More importantly, however, ignoring overdispersion resulted in the most complex model proposed, PD(s + h), being the best model 20% of the time, and its chance of selection was not reduced even when model nesting was taken into account.

We also investigated the utility of model weighting. As the underlying model is non-linear, it is not recommended that model parameters be weighted (Burnham and Anderson 2002); instead, we investigated whether model weights improved our predictions of the expected number of seeds consumed. Specifically, we simulated 100 data sets for a range of effect sizes due to watering ($s_W - s_D$), varying s_D from 0.1 to 1.0 whilst retaining s_W=1.0. For each data set, we calculated the expected number of seeds consumed for each of the ten factor pairings using the AIC best model and model averaging and compared them with the true expected counts, $Tf(x)$. We quantified the accuracy of both approaches for each data set by calculating the absolute, relative error, averaged across the ten factor pairings. Overall, we observed a decrease in accuracy as effect size increased; however, the relation was certainly non-linear (Fig. 4). More interestingly, in this example, the two approaches produced very similar levels of accuracy and predicted very similar consumption rates, even when a single model did not have a dominant weight (Fig. 3b).

Discussion

We have presented reasons why we feel the IT-AIC approach is well suited for analysing the types of data that are often collected during behavioural ecology studies, including simple controlled experiments and observational studies. Behavioural studies are often motivated by reconciling multiple proposed hypotheses, and IT-AIC provides a natural and straightforward framework for comparing such hypotheses, more so than NHST.

Table 2 Summary of 100 AIC model selection analyses for the foraging example discussed in the main text

Model	*K*	Overdispersion considered					Overdispersion ignored		
		AIC best	Simple rule (2)	Simple rule (6)	Nesting rule (2)	Nesting rule (6)	AIC best	Simple rule (6)	Nesting rule (6)
PD(−)	2	0	0	0	0	0	0	5	5
PD(s)	3	0	3	5	3	5	56	86	86
PD(h)	3	0	2	5	2	5	24	58	58
PD(s + h)	4	0	2	9	2	9	20	100	20
NBD(-)	3	2	6	26	6	26	NA	NA	NA
NBD(s)	4	61	85	97	83	95	NA	NA	NA
NBD(h)	4	30	52	88	50	86	NA	NA	NA
NBD(s + h)	5	7	99	100	7	7	NA	NA	NA
Mean selected		1	2.49	3.30	1.53	2.33	1	2.49	1.69

Results are presented for analyses that only considered Poisson distributions (PD) when describing variation in the data (overdispersion ignored) and analyses when negative binomial distributions (NBD) were also considered (overdisperison considered). See main text for descriptions of each model. The data were generated according to model NBD(s), parameterized by $T=20$, $h_1=h_2=0.5$, $s_1=1$, $s_2=0.5$ and $\phi=0.075$. The number of times each model was the one with the lowest AIC value from the 100 random data sets is presented (AIC best). Also presented is the number of times each model was chosen for various selection criteria. Simple rule refers to selecting all models having a Δ value less than the threshold indicated in brackets. Nesting rule refers to the method where, in addition to the simple rule, models are not selected if they are more complicated versions of a model having a lower AIC value. Also presented is the average number of models selected by each rule (Mean selected). *K* denotes the number of model parameters

To illustrate issues relating to model selection in behavioural ecology, we presented two examples of IT-AIC analyses applied to behavioural data. The first, simple example illustrated three major points: first, in simple contexts, inferences drawn from IT-AIC and NHST analyses will often be similar; second, estimates of effect sizes generated using IT-AIC model averaging will tend to be more stable than those based solely on the best AIC model; third, in spite of this, model averaging does not reliably improve the accuracy of estimated effect sizes. With our second, more complex example, we illustrated the effective use of IT-AIC analyses for fitting process-based models directly to raw data (see further below). Our second example reinforced the findings of the simpler study that model averaging does not reliably improve prediction accuracy. In that context, we made suggestions regarding thresholds for deciding whether a candidate model is well supported and for dealing with nested models during model selection and model averaging.

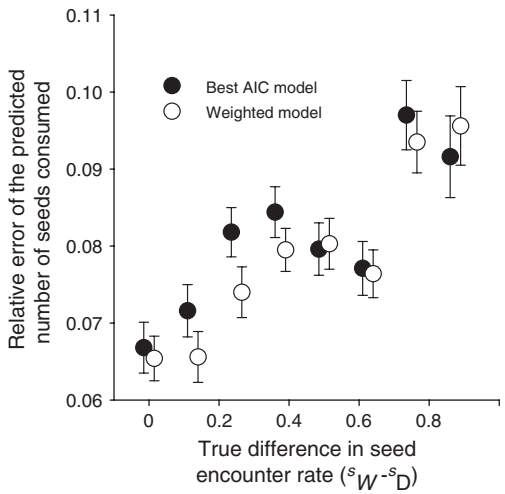

Fig. 4 A comparison of prediction accuracy for the seed predation experiment when prediction is estimated using the AIC best model and when model averaging is used. Soil moisture only affects seed encounter rate. Means and standard errors of the mean are presented based on 100 simulated studies for eight different effect sizes. Accuracy is quantified as the relative absolute error, $|\hat{y}-y|/y$, where $y = Tf$ is the true expected number of seeds consumed and $\hat{y}$ is the estimated expectation using either prediction approach

Two particular advantages of IT-AIC that arise from our second example include the ability of this technique to aid the identification of appropriate error structures (see also Hobbs and Hilborn 2006) and the flexibility of the approach to deal with data that do not conform to some of the more restrictive assumptions of many NHST analyses. Our example specifically illustrates how a biological understanding of a system can be enhanced when models are developed based explicitly on proposed mechanisms and fitted directly to data. Often, such models predict non-linear relations, non-normal variation and/or heteroscedasticity which complicate or even prevent the use of commonly adopted NHST approaches. Even those NHST-based approaches developed to address non-normally distributed errors (e.g. generalised linear models) and non-linear mean relations (e.g. generalised additive models) make some assumptions of the data. How best to

incorporate these more general models into a model selection framework continues to be an active area of research (e.g. Hu and Shao 2008). Here, we have shown that the IT-AIC approach provides an alternative and often more straightforward approach for analysing naturally complex data.

In this paper, we only considered relatively simple studies and models; however, the IT-AIC approach can be applied in a straightforward manner to more complex studies involving multiple fixed factors and covariates. A more challenging situation arises when data are not independent; for example, the same forager could have been used for multiple seed density or soil treatments. In this case, treating all data as independent may result in incorrect inference regarding treatment effects (Schielzeth and Forstmeier 2009). Mixed effects models can be used to control for non-independence under the NHST framework. Schielzeth and Forstmeier (2009) and van de Pol and Wright (2009) provide some good examples illustrating that non-independence is common when performing behavioural studies; they also show how random effects can be incorporated into linear models to tease apart within-subject effects from between-subject effects. Non-independence can be tackled using the IT-AIC approach by noting mechanistically how non-independence will likely affect the data observed and constructing appropriate model likelihoods. For example, models could be constructed by proposing that individual seed encounter rate (s_i) and seed handling time (h_i) were drawn from probability distributions specific to the group the individual is in. In this case, the model would have parameters describing these probability distributions. This latter approach is similar to the approaches suggested by Schielzeth and Forstmeier (2009) and van de Pol and Wright (2009) whereby slopes and intercepts are treated as random effects. For the foraging example presented here, however, these NHST approaches would not be directly applicable; in particular, the expected relationship between seed density and consumption is non-linear, and the expected variation is non-normal. Unfortunately, although the IT-AIC approach can account for non-linear relations and non-normally distributed variation, it is seldom straightforward to implement numerical integration schemes in order to calculate model likelihoods. When likelihoods can be calculated, however, examining which models were selected using IT-AIC would provide insight into the degree and the nature of variation among individuals, both within and between groups; this could prompt the researcher to investigate further mechanisms or factors causing the differences. Other patterns of variation often found in behavioural data (e.g. inflated zero counts) can also be modelled in a mechanistic fashion (Martin et al. 2005) and support for the proposed mechanisms quantified using IT-AIC.

Our simulation studies demonstrate how Δ values can be used to generate a candidate set that contains models that are most consistent with the data. In order to have approximately a 95% chance of including the truly most parsimonious model in the candidate set, we suggest selecting models with Δ value less than 6. In addition, to avoid retention of overly complex models (i.e., models having additional parameters that result in a minimal increase in fit), we also suggest that models be removed from the candidate set if they are more complex versions of models having a lower AIC value (see Table 2). Rejection of models based on their nesting is not a new idea (e.g. Burnham and Anderson 2002; Richards 2008); unfortunately, however, nesting is seldom considered in published studies employing AIC analyses.

Burnham and Anderson (2002) and Lukacs et al. (2007) suggest the use of Akaike weights to quantify the probability that a model is the best of those considered, given the data. We have not identified the true best model for the examples presented here, as it requires estimation of the Relative Expected Kullback–Leibler Distance (REKLD); consequently, we cannot use these examples to assess the utility of these weights as probability estimates. It is important to note that the true best model may not be the model that generated the data in these studies, but may be a simpler model if model fitting can result in poor parameter estimates (Richards 2005). However, when REKLD have been estimated, the correspondence between model weights and the true best model has been equivocal (Richards 2005). Our simulated examples resulted in large variation in model weights from one data set to the next, suggesting that model weights are likely to be relatively uninformative indicators of the true best model. Model weights ignore model nesting and so are also prone to overestimating support for overly complex models.

Burnham and Anderson (2002) and Lukacs et al. (2007) also suggest that model weights be used to incorporate model uncertainty, thereby providing more reliable estimates of model parameters (if models are linear) and model predictions. We have examined this suggestion using our simulated examples and found mixed results. Whilst we agree that the precision of model parameters or model predictions will often be improved by model averaging (Fig. 2c, d), we found little evidence that model averaging improved accuracy with any consistency (Figs. 2b and 4). Our findings support calls for further research on the utility of models weights and their application to model averaging (Buckland et al. 1997; Burnham and Anderson 2002; Richards 2005). Thus, at this stage, we do not see any great advantages arising from using model weights in an AIC analysis. Instead, we suggest the more cautionary approach of generating a candidate model set using model Δ values (see above) and basing inference on the simpler models selected.

In conclusion, we feel behavioural ecologists would often benefit from adopting the IT-AIC approach over NHST. For simple analyses, the benefits are largely philosophical, and operational differences are likely to be slight (see also Mundry 2010). By contrast, when more complex analyses are undertaken, IT approaches provide a more straightforward framework for testing multiple biological hypotheses and are often more amenable to the analysis of commonly collected behavioural data. In particular, the IT-AIC approach often allows biological hypotheses to be made more explicit during an analysis, which we believe can help to improve the chance of correctly inferring mechanism. Ultimately, as the IT-AIC approach simply assesses the relative consistency of models, given the data at hand, the success of an IT-AIC analysis depends on the models proposed and the quality of the data (e.g. Nakagawa and Freckleton 2008). This dependency highlights the need for careful thought when proposing alternate hypothesis before data analysis which, in itself, may also contribute to a greater understanding of a system.

Acknowledgements We would like to thank Arthur Goldsmith and Rob Freckleton for helpful advice and discussions, three reviewers for their helpful comments, and Laszlo Garamszegi and Shinichi Nakagawa for inviting us to contribute to this issue. MJW was supported by a David Phillips Fellowship.

Appendix

A likelihood function for the functional response experiment

Recall that data consist of 50 triplets (x, i, y), where x is the initial density of seed, i = D or W (indicating dry or wet soil, respectively), and y is the number of seeds consumed. For each triplet, (x, i, y), each of the eight models provides the mean handling time and search rate, h_i and s_i, respectively. These two parameters can then be substituted into Eq. 2 to give the mean seed encounter rate, f. If the model assumes overdispersed data (i.e. the data are described by the NBD), then the likelihood of the model parameters, given the datum, is

$$L(s_i, h_i, \phi | x, y) = \frac{\Gamma(y+a)}{\Gamma(y+1)\Gamma(a)} \left(\frac{b/T}{1+b/T}\right)^a \left(\frac{1}{1+b/T}\right)^y, \qquad \text{(A.1)}$$

where T is the duration of the experiment, ϕ is a parameter describing the amount of variation among individuals, Γ is the complete gamma function, $a = f/\phi$ and $b = 1/\phi$. Using this formula to describe the NBD, the variance inflation factor is simply $(1+\phi)$ (see Richards 2008 for details). In the limit as ϕ tends to zero, Eq. A.1 approaches the likelihood assuming a Poisson distribution. The likelihood of the model, given all the data, is the product of the likelihoods for each datum. Note that Eq. A.1 can be calculated in Microsoft Office Excel using the GAMMALN function, and maximum likelihood estimates can be found using the SOLVER add-in; hence, a complete AIC analysis can be implemented using a spreadsheet.

References

Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. J Wildl Manage 64:912–923

Bolker BM (2008) Ecological models and data in R. Princeton University Press, Princeton

Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. Biometrics 53:603–618

Burnham KP, Anderson DR (2001) Kullback–Leibler information as a basis for strong inference in ecological studies. Wildlife Res 28:111–119

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer, New York

Burnham KP, Anderson DR, Huyvaert K (2010) AIC_c model selection in ecological and behavioral science: some background, observations, and comparisons. Behavioral Ecology & Sociobiology. doi:10.1007/s00265-010-1029-6

Cohen J (1994) The earth is round (P<.05). Am Psychol 49:997–1003

Freckleton RP (2010) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behavioral Ecology & Sociobiology. doi:10.1007/s00265-010-1045-6

Garamszegi LZ (2010) Information-theoretic approaches to statistical analysis in behavioural ecology: an introduction. Behavioral Ecology & Sociobiology. doi:10.1007/s00265-010-1028-7

Gurney WSC, Nisbet RM (1998) Ecological dynamics. Oxford University Press, Oxford

Hegyi G, Garamszegi LZ (2010) Using information theory as a substitute for stepwise regression in ecology and behavior. Behavioral Ecology & Sociobiology (in press)

Heyes CM, Dawson GR (1990) A demonstration of observational-learning in rats using a bidirectional control. Q J Exp Psychol B 42:59–71

Hilborn R, Mangel M (1997) The ecological detective: confronting models with data, vol 28. Princeton University Press, Princeton

Hobbs NT, Hilborn R (2006) Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecol Appl 16:5–19

Hu B, Shao J (2008) Generalized linear model selection using R^2. J Stat Plan Inference 138:3705–3712

Johnson DH (1999) The insignificance of statistical significance testing. J Wildl Manage 63:763–772

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. Trends Ecol Evol 19:101–108

Krebs JR, Davies NB (1978) Behavioural ecology: an evolutionary approach. Blackwell, Oxford

Lind J, Cresswell W (2005) Determining the fitness consequences of antipredation behavior. Behav Ecol 16:945–956

Lukacs PM, Thompson WL, Kendall WL, Gould WR, Doherty PF, Burnham KP, Anderson DR (2007) Concerns regarding a call for pluralism of information theory and hypothesis testing. J Appl Ecol 44:456–460

Martin TG, Wintle BA, Rhodes JR, Kuhnert PM, Field SA, Low-Choy SJ, Tyre AJ, Possingham HP (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol Lett 8:1235–1246

Martinez-Abrain A (2007) Are there any diffferences? A non-sensical question in ecology. Acta Oecol 32:203–206

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

Mitchell CJ, Heyes CM, Gardner MR, Dawson GR (1999) Limitations of a bidirectional control procedure for the investigation of imitation in rats: odour cues on the manipulandum. Q J Exp Psychol B 52:193–202

Mundry R (2010) Issues in information theory based statistical inference—a commentary from a frequentist's perspective. Behavioral Ecology & Sociobiology. doi:10.1007/s00265-010-1040-y

Nakagawa S, Cuthill IC (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev 82:591–605

Nakagawa S, Freckleton RP (2008) Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol 23:592–596

Nakagawa S, Freckleton RP (2010) Model averaging, missing data and multiple imputation: a case study for behavioural ecology. Behavioral Ecology & Sociobiology. doi:10.1007/s00265-010-1044-7

R Development Core Team (2005) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Richards SA (2005) Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology 86:2805–2814

Richards SA (2008) Dealing with overdispersed count data in applied ecology. J Appl Ecol 45:218–227

Rogers D (1972) Random search and insect population models. J Anim Ecol 41:369–383

Ruxton GD, Colegrave N (2006) Experimental design for the life sciences, 2nd edn. Oxford University Press, Oxford

Schielzeth H, Forstmeier W (2009) Conclusions beyond support: overconfident estimates in mixed models. Behavioral Ecology 20:416–420

Stephens PA, Buskirk SW, Hayward GD, Martinez del Rio C (2005) Information theory and hypothesis testing: a call for pluralism. J Appl Ecol 42:4–12

Stephens PA, Buskirk SW, Hayward GD, Del Rio CM (2007a) A call for statistical pluralism answered. J Appl Ecol 44:461–463

Stephens PA, Buskirk SW, Martinez del Rio C (2007b) Inference in ecology and evolution. Trends Ecol Evol 22:192–197

van de Pol MV, Wright J (2009) A simple method for distinguishing within- versus between-subject effects using mixed models. Anim Behav 77:753–758

Vander Wall SB (2000) The influence of environmental conditions on cache recovery and cache pilferage by yellow pine chipmunks (*Tamias amoenus*) and deer mice (*Peromyscus maniculatus*). Behavioral Ecology 11:544–549

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 75:1182–1189

Ydenberg RC, Brown JS, Stephens DW (2007) Foraging: an overview. In: Stephens DW, Brown JS, Ydenberg RC (eds) Foraging: Behavior and Ecology. University of Chicago Press, Chicago, pp 1–28