JOURNAL OF Evolutionary Biology

doi: 10.1111/j.1420-9101.2010.02210.x

REVIEW

Multimodel inference in ecology and evolution: challenges and solutions

C. E. GRUEBER, S. NAKAGAWA, R. J. LAWS & I. G. JAMIESON

Department of Zoology, University of Otago, Dunedin, New Zealand

Keywords:

Akaike Information Criterion;
generalized linear mixed models;
inbreeding;
information theory;
lethal equivalents;
model averaging;
random factors;
standardized predictors.

Abstract

Information theoretic approaches and model averaging are increasing in popularity, but this approach can be difficult to apply to the realistic, complex models that typify many ecological and evolutionary analyses. This is especially true for those researchers without a formal background in information theory. Here, we highlight a number of practical obstacles to model averaging complex models. Although not meant to be an exhaustive review, we identify several important issues with tentative solutions where they exist (e.g. dealing with collinearity amongst predictors; how to compute model-averaged parameters) and highlight areas for future research where solutions are not clear (e.g. when to use random intercepts or slopes; which information criteria to use when random factors are involved). We also provide a worked example of a mixed model analysis of inbreeding depression in a wild population. By providing an overview of these issues, we hope that this approach will become more accessible to those investigating any process where multiple variables impact an evolutionary or ecological response.

Introduction

There has been a recent and significant change in the way that ecologists and evolutionary biologists analyse and draw biological inferences from their data. As an alternative to traditional null hypothesis testing (sometimes referred to as the 'frequentist' approach), an information theoretic or 'IT' approach examines several competing hypotheses simultaneously to identify the best set of models (i.e. hypotheses) via information criteria such as Akaike's information criterion (Burnham & Anderson, 1998, 2002; Anderson *et al.*, 2000). In addition, the IT approach makes inferences based on weighted support from several models, i.e. model averaging (detailed below).

The IT approach, and specifically model averaging, has numerous advantages over traditional hypothesis testing of a single null model where support is measured by an arbitrary probability threshold. Instead, similar to Bayesian approaches, several models can be ranked and weighted to provide a quantitative measure of relative support for each competing hypothesis. In cases where two or more models achieve similarly high levels of support, model averaging of this 'top model set' can provide a robust means of obtaining parameter estimates (both point and uncertainty estimates) and making predictions (Burnham & Anderson, 2002). By comparison, more traditional approaches such as stepwise methods, although also resulting in a final model, completely ignore model uncertainty (e.g. Whittingham *et al.*, 2006). Starting with a strong base in the field of wildlife management and mark-recapture studies to estimate population abundance and survival probabilities (Lebreton *et al.*, 1992; Schwarz & Seber, 1999), the IT approach is now being used in many areas of ecology and evolution including landscape ecology, behavioural ecology, life history evolution, phylogenetics and population genetics (Johnson & Omland, 2004; Carstens *et al.*, 2009). Although many biologists agree with the principles behind using this approach, the ways and means of applying a multimodel procedure and model averaging to various types of biological problems are still in their infancy (see also Richards, 2005).

Meanwhile, linear mixed-effects modelling and its extension to generalized linear mixed-effects models (GLMMs) are now used widely in ecology and evolutionary biology (Paterson & Lello, 2003; Bolker *et al.*, 2009). GLMMs are extremely useful as they permit the inclusion of random effects as well as fixed effects to complex and realistic hierarchical biological systems,

Correspondence: Catherine E. Grueber, Department of Zoology, University of Otago, PO Box 56, Dunedin 9054, New Zealand.
Tel.: +64 3 479 7986; fax: +64 3 479 7584;
e-mail: c_grueber@yahoo.co.nz

simultaneously dealing with non-normal response variables (such as binary and count data). The recent popularity of GLMMs is not surprising, as they are an overarching statistical tool that encompasses older tools such as *t*-tests, ANOVA, ANCOVA and generalized linear models (GLMs), and indeed many of the issues we discuss herein can be applied to other modelling approaches. Unfortunately, the handling of the random effects in the IT environment, especially when model averaging is employed, is not straightforward, as the best method of estimating Akaike Information Criterion (AIC) (see Box 1), when random effects are included is unclear

Box 1: a summary of the alternatives to AIC

Forms of the AIC, such as AIC_C (small sample size correction, Table 1) and Quasi-AIC (QAIC: controls for overdispersion), remain the most widely used information criteria for ranking models in the IT approach. However, there is debate surrounding the utility of AIC (e.g. Spiegelhalter *et al.*, 2002; Stephens *et al.*, 2007), and various alternatives have been proposed. The different criteria in use today may be appropriate in different circumstances (Murtaugh, 2009), but all information criteria are in fact approximations of Bayes Factors (BFs) (Congdon, 2006a) with certain assumptions such as large sample sizes. The BF is a ratio between two models, reflecting 'true' model probabilities given data support, i.e. posterior model probabilities (other information criteria approximate these posterior model probabilities) (Jefferys, 1961 *in* Congdon, 2006b):

$$BF_{1,2} = \frac{p(y|M_1)}{p(y|M_2)} \quad (1)$$

where $p(y|M_i)$ is the marginal likelihood of model i. Therefore, BFs seem to be the ideal index for model selection and averaging. However, calculations of BFs directly become quickly complicated when comparing more than two models. Although several methods for using BFs for model averaging have been suggested, it seems that currently available methods are highly technical and difficult to implement (Congdon, 2006a). Practical implementations of BFs for multimodel comparisons are an active frontier of statistical research (R. Barker, personal communication) and thus advances in the area are anticipated in the near future.

In the interim, a particular alternative to AIC, the weighted Bayesian Information Criterion (BIC) has been proposed as superior to AIC in IT model averaging approaches (Link & Barker, 2006), as it tends to favour more parsimonious models [c.f. AIC which tends to favour complex models (Burnham & Anderson, 2002; Link & Barker, 2006)] and does not require approximation of likelihood. However, BIC still does not accurately quantify k for random effects (Table 1), and AIC and BIC can in fact give similar results for particular data sets (Murtaugh, 2009). Another criterion, also in the Bayesian context, is the Deviance Information Criterion [DIC (Spiegelhalter *et al.*, 2002)], which improves on BIC by the incorporation of the term k_D: effective number of parameters. DIC is a promising metric for use with mixed models; however, its application to model averaging is not yet implemented in widely used statistical packages nor has it been widely tested with either simulations or empirical data. DIC is both philosophically and mathematically more similar to AIC than BIC (Spiegelhalter *et al.*, 2002) in that DIC suffers similar problems to AIC (R. Barker, personal communication, Table 1). Conditional AIC [cAIC (Vaida & Blanchard, 2005; Liang *et al.*, 2008)] is another interesting prospect in that it too can control for the number of effective parameters. However, Vaida & Blanchard (2005) state that specification of the number of parameters (i.e. whether to count each random effect as 1, as per AIC_C, or to use the effective number of parameters, as per cAIC) depends on the question being investigated. Notably, cAIC is yet to be widely implemented in statistical packages allowing its use for model averaging.

Table 1 presents the formulae for the aforementioned information criteria, although this is by no means an exhaustive list of information criteria. Other information criteria found in the statistical literature include: the Focused Information Criterion (FIC) (Claeskens & Hjort, 2003; Claeskens *et al.*, 2007), Akaike's Bayesian Information Criterion, the Generalized Information Criterion (GIC), the Extended (Bootstrap) Information Criterion (EIC), the Predictive Information Criterion and Takeuchi's Information criterion [TIC; reviewed in Konishi & Kitagawa (2008)]. Alternatives to AIC that still rely on maximum likelihood estimation and k are subject to the same issues as AIC for model averaging under IT in generalized linear mixed modelling. Overall, information criteria can be assigned to either of two broad categories: those suited for model selection (such as BIC) and those suited for minimizing predictive error (such as AIC and others outlined above) (Yang, 2005). The type of criteria chosen depends on the question being answered (Yang, 2005), which in turn influences how the number of degrees of freedom should be calculated (Vaida & Blanchard, 2005; Bolker *et al.*, 2009).

Table 1 Information criteria for model selection.

Information criterion	Formula*	References
Akaike Information Criterion	$AIC = -2 \cdot \ln L + 2k$	Akaike (1973)
AIC – small sample size correction	$AIC_C = -2 \cdot \ln L + \frac{2k(k+1)}{n-k-1}$	Hurvich & Tsai (1989)
Quasi-AIC	$QAIC = \frac{-2 \ln L}{\hat{c}} + 2k$	Lebreton *et al.* (1992)
Conditional AIC	$cAIC = -2 \cdot \ln L + 2k_C$	Vaida & Blanchard (2005); Liang *et al.* (2008)
Bayesian Information Criterion	$BIC = -2 \cdot \ln L + k \ln (n)$	Schwarz (1978)
Deviance Information Criterion	$DIC = -2 \cdot \ln L + 2k_D$	Spiegelhalter *et al.* (2002)

*L = likelihood function = $p(y|\theta)$, or, if random factors are explicitly separated as parameters (as in cAIC) = $p(y|\theta,u)$. NB, $-2 \cdot \ln L$ is also known as the 'deviance'. k = number of parameters in the model; n = sample size; $\hat{c}$ = overdispersion parameter; k_C = effective number of degrees of freedom (cAIC); k_D = effective number of parameters (DIC). See listed references for additional details of formula components.

Table 2 Overview of practical issues associated with IT approaches and model averaging in evolution and ecology covered in this manuscript, with their tentative solutions.

Practical problem	Tentative solution
General challenges in the IT approach	
Translating biological hypotheses into statistical models	This is likely to remain the most difficult aspect of using an IT approach with model averaging in ecology and evolution, because of the complexity of biological processes
Which information criterion to use when comparing models	AIC_C is most widely used; where random effects are present, this problem is at present unresolved. See also Box 1
Whether to model average	If the weight of the 'best' model < 0.9, model averaging is recommended
Practical challenges for model averaging an ecological data set	
Narrowing a list of predictors from the measured input variables	Use 'biologically reasonable' variables; only transform if there is an *a priori* justification. Consider whether *a priori* examination and/or removal of individual variables is appropriate
Presence of strongly correlated variables	Depends on the nature of the correlation (see text); aim to select the variables that are most biologically important
Generating a model set	One method is to generate a global model of all biologically relevant parameters, and then generate all possible submodels from this. However, if the global models fails to converge, it may be necessary to reduce its complexity/size
Incompatibility of global model parameters	Tailor the model set to include only plausible models
How to compute the model average (natural average or zero method)	Depends on the aim of the study (see text)
How to define a top model set (what cut-off to use)	Consider how many models (*S*) will be captured by a given cut-off. 'Too many' (based on *N*) is discouraged because of the risk of spurious results, but specific recommendations for *S* are lacking
How to evaluate model goodness-of-fit	In nonmixed models one can calculate R^2; however; calculation of model fit is much more technical in mixed models, thus presenting a practical difficulty
How to use the model for prediction	The model can give 'conditional estimates', e.g. predictions for a factor of interest at the mean of all other parameters
Special issues for complex models	
Defining random intercepts or slopes	Always fit slope if possible, otherwise use just the intercept
Nested models in the top model set	It is recommended to remove models from the set that are complex versions of simpler ones, but clear guidelines are currently lacking
Whether to force inclusion of a parameter of interest in the model set/final model	Perform with caution if using the zero method of model averaging. Also, force inclusion of a parameter fixes its relative importance at 1, making this metric no longer useful
How to interpret the effect sizes of interactions and their main effects	Centring variables permits interpretation of main effects when interactions are present
How to interpret effect sizes when predictors are on different scales	Standardization on 0.5 SD results in effect sizes that are on comparable scales

AIC, Akaike Information Criterion; IT, information theoretic.

(Bolker, 2009). Additional difficulties of the IT approach become quickly evident when compiling a biologically meaningful model set [i.e. the difficulties of translating biological hypotheses to statistical models (Dochtermann & Jenkins, 2010)]. Even if one succeeds in compiling a model set, the model averaging procedure is complicated when interaction and polynomial terms are included (Dochtermann & Jenkins, 2010). Furthermore, it is not entirely clear how to proceed when a top model set for averaging does not include a particular factor of interest.

Despite having a relatively good understanding of the basic theory behind the IT approach, we encountered a number of problems when applying this approach to what initially appeared as a relatively straightforward but fundamental analysis: modelling the effects of inbreeding in wild populations (Grueber *et al.*, 2010; Laws & Jamieson, 2010; Laws *et al.*, 2010). It is these difficulties, and the general lack of specific guidelines for overcoming these in the literature at present, that led to this paper.

The aim of this paper is to highlight some of the common practical obstacles and challenges faced when performing mixed modelling under IT and recommend potential solutions where they exist. Our manuscript is intended to accompany recent papers that review particular statistical issues with the IT approach (for example Johnson & Omland, 2004; Richards, 2005; Link & Barker, 2006; Bolker *et al.*, 2009; Carstens *et al.*, 2009; and a recent 'special issue' of *Behavioral Ecology and Sociobiology* [2011, Vol. 65, No. 1]). The current manuscript provides methodological guidelines for practitioners in ecology and evolution who have already decided that the IT approach is appropriate for their data and the reader is directed to relevant reviews for additional detail. The issues addressed here, with their tentative solutions, are summarized in Table 2. We further illustrate the practical difficulties posed when using IT and model averaging approaches, through reference to a worked example (see Appendix), which provides clear, step-by-step

instructions for effective analysis and standardization of reporting using the IT method. The worked example focuses on modelling the fitness effects of inbreeding on a life history trait that is also affected by several demographic variables, and in which the analysis requires model averaging to predict survival estimates for different levels of inbreeding (Grueber *et al.*, 2010). By providing a systematic overview of tentative solutions to practical challenges faced, we hope that the IT approach will become more accessible to those interested in the analysis of any process where multiple variables impact an evolutionary or ecological response.

Beyond simple model selection

One of the key philosophies that distinguishes the IT approach from traditional null hypothesis testing is the evaluation of relative support for a given hypothesis over others (Burnham & Anderson, 2002), similar to the concepts of a Bayesian framework. As such, each model to be compared constitutes a biological hypothesis, yet one of the first problems encountered when using an IT approach to modelling ecological processes is in translating biological hypotheses into statistical models.

Defining appropriate input and predictor variables

The primary step is to determine which input variables to include, and whether or how to transform these into predictor variables (explanatory or independent variables) (see Appendix: Step 1). Note that we make a distinction here between input variables (raw parameters that are measured) and predictor variables (the variables used in the model, which can also include interactions and polynomial terms) (Gelman, 2008).

Burnham & Anderson (2002) suggest that only predictors with strong biological reasoning (based on *a priori* investigation) should be included from the outset, to prevent overparameterization. In complex ecological systems, it is plausible that any number of factors could have an important effect on the response variable; therefore, one should consider the sample size rule-of-thumb of 10 : 1 subjects to predictors in multiple regression (Harrell, 2001). In addition, there are a large number of possible second- and higher-order interactions and transformations (e.g. log-transformation) that may be applied to input variables. Unless there is an *a priori* biological reason for expecting such conversions to improve the fit to the data (for example, to improve the normality of residuals), there is little justification for including these in the predictor set. Incidentally, regression analysis by GLMM does not require predictors (input variables) to be normally distributed, although in some cases, normalization transformations can reduce residual variance and therefore affect inference regarding parameter estimates (Faraway, 2005).

Where there are large numbers of possible predictors, it might seem natural to explore each variable independently prior to generating models to identify factors impacting strongly on the response. Doing so informally, ideally graphically, is exactly what exploratory data analysis is about (Tukey, 1977; Zuur *et al.*, 2010). However, advocates of the IT approach such as Burnham & Anderson (2002) are in principle against exploratory data analysis, because it results in *post hoc* creation of statistical models and thus biological hypotheses. They recommend that predictors should be selected on the basis of genuine prior knowledge, such as from pilot studies or the scientific literature (Burnham & Anderson, 2002).

An additional point to consider is collinearity amongst predictors, which has received little attention despite being a characteristic of many ecological studies (Freckleton, 2010). Collinearity amongst predictors can be a problem in model selection, as a number of models each containing different (but correlated) predictors may provide similar fits to the data and thus present difficulties when choosing the 'best' model and determining true relationships (Freckleton, 2010). Using simulations, Freckleton (2010) demonstrated that when predictors are correlated, IT approaches and model averaging performed just as well or even better than ordinary least squares methods at parameter estimation. However, Freckleton cautioned that measurement errors in correlated predictors can cause problems in any analysis. Whether to combine collinear variables (for example into principal components) depends on the nature of the variables themselves and the relationships that are expected (for examples see Freckleton, 2010). Incidentally, the high prevalence of correlated predictors in ecological data sets suggests to us the importance of exploratory data analysis of predictors.

Random factors

The benefit of using GLMMs is that the inclusion of random factors provides a means of dealing with nonindependence of data (e.g. individuals that breed from one year to the next, or breeding sites repeatedly used by different pairs), or for hierarchical study designs (e.g. individuals from the same social group, site, or taxon). Schielzeth & Forstmeier (2009) suggest that both random intercepts (to account for variation between group means, or 'inter-individual' variation where individuals are sampled repeatedly) and random slopes (to account for variation in group responses, or 'within-individual' variation) should be fitted where possible [see also Fig. 1 in van de Pol & Wright (2009)]. Using both random intercepts and slopes reduces the incidence of Type I and Type II errors and reduces the chance of overconfident estimates (unrealistically low standard error, SE) (Schielzeth & Forstmeier, 2009). However, fitting random slopes requires relatively large sample sizes for model convergence, especially if the data set

contains many groups with only a few observations (obviously, a slope cannot be fitted to only one data point, although it is very common in ecological data to have many individuals with only single observations). Therefore, we recommend attempting to fit both random intercepts and slopes unless the model does not converge, in which case fitting a random intercept only is preferable to not including the random variable at all (see Appendix: Step 1).

Generating a set of models to compare

Once it has been established which predictors are to be included, the next step is to generate a 'model set' of hypotheses (see Appendix: Step 2). The easiest way to generate a model set is to derive all possible submodels from a set of predictors of interest (but not necessarily all *possible* predictors, see previous section), including an intercept-only model (which should also contain any random factors), and then compare these (e.g. Symonds & Johnson, 2008). This method of generating a model set is acceptable insofar as each model is ecologically justifiable (Dochtermann & Jenkins, 2010). From a practical point-of-view, the easiest way to accomplish this in a statistical package such as R (R Core Development Team, 2009) is to generate a global model containing all the predictors of interest and then derive submodels from this [see Appendix: Step 2; see also Symonds & Moussalli (2010) for a summary of other software that perform AIC-based analyses].

There are, however, a number of potential obstacles to generating a model set in this way, such as what to do if the global model does not converge (possibly because of overparameterization in cases where sample size is small). There are two types of nonconvergence that can occur: the first is the failure to estimate parameters; the second is the overestimation of SE or confidence intervals, which can occur in the absence of any error messages from software (Bolker *et al.*, 2009). One solution to either of these forms of nonconvergence is to follow the recommendation of Bolker *et al.* (2009) and reduce the size and complexity of the global model. Interactions can be removed first (particularly those where the main effects are weak), and then undertaking *a priori* investigation of individual factors and removing one-by-one those main effects that either appear to have least impact on the response, or are of least biological interest, until the model converges. An alternative is to generate a submodel set manually; for example, if 10 parameters are to be investigated but the global model cannot converge, it may be desirable to generate a model set of all submodels with a maximum of five parameters each. However, automation would be required, as this example would result in 638 possible models (not including interactions or polynomials), far too many to generate by hand. Even so, by taking this approach, one is likely to fall victim to the 'problem of too many models' (Burnham & Anderson, 2002; Dochtermann & Jenkins, 2010), leading to potentially spurious results. In addition, care should be taken to avoid generating submodels that may be biologically implausible. For example, in cases where predictors are mutually exclusive or otherwise incompatible, models containing combinations of these should not be included in the model set. Again, we support the recommendations of Zuur *et al.* (2010) and reinforce the importance of exploratory data analysis and careful consideration of predictors.

Specific treatments or factors of interest

When there is a particular factor of interest (such as a particular experimental treatment, or population parameter such as inbreeding in the worked example in Appendix), it may seem reasonable to restrict the model set such that it only includes models that contain this focal parameter. However, this method should be used with caution as models excluding the focal parameter could possibly provide a superior fit to the data. For example, it may turn out that a particular covariate, such as age, explains the majority of the variation in the response variable and that the inclusion of the focal parameter, inbreeding, explains no additional variation; inbreeding may in fact introduce additional uncertainty. In the worked example, we chose not to restrict our model set (see Appendix: Steps 2 and 3, Table S3). Ultimately, the decision of whether to restrict a model set to contain only models with a factor of interest depends in part on the subsequent method used to model average, which we describe below.

Model selection and model averaging

If the model set is large, there may be no single best model: a number of models in the set may differ in their data fit by only small amounts, as defined by an information criterion. Under these circumstances, it is best to employ an IT model averaging approach, a procedure that accounts for model selection uncertainty to obtain robust parameter estimates or predictions. This procedure entails calculating a weighted average of parameter estimates, such that parameter estimates from models that contribute little information about the variance in the response variable are given little weight. Various information criteria have been presented to determine the amount of information contained in a given model (Table 1). At present, the most commonly used is the Akaike Information Criterion, AIC (Akaike, 1973), and its correction for small sample size [AIC_C (Hurvich & Tsai, 1989)] although AIC may be more suitable than AIC_C when modelling certain nonlinear ecological responses (Richards, 2005). Simulation studies have shown that in certain circumstances, choosing the 'best' model (based on AIC_C for example) may provide similar parameter estimates when compared to model averaging. However, model-averaged results can be more

stable than those based on choosing the best model, as the former is less likely to erroneously conclude that weak parameter estimates are zero (Richards, 2005; Richards *et al.*, 2010). It should be borne in mind, however, that assigning the incorrect sign to a weak parameter estimate is a possibility in any regression (Gelman & Tuerlinckx, 2000), and further research as to the effects of model averaging on this type of error would be useful.

An important issue with the broad application of AIC_C to GLMMs is in the calculation of the number of parameters (k) when random factors are included (Spiegelhalter *et al.*, 2002; see also Box 1). Tentative solutions are provided in the development of alternative information criteria for use in IT model averaging, especially under a Bayesian framework (Box 1). Additionally, in GLMM analysis, the residual variance of non-Gaussian data may be modelled as either multiplicative overdispersion (the overdispersion parameter which appears in QAIC, see Table 1) or additive overdispersion (a residual variance as in linear mixed models; see Browne *et al.*, 2005). These different implementations can obviously influence information criterion calculations (Nakagawa & Schielzeth, 2010). Although both methods of modelling overdispersion are suited for fitting GLMMs, different software packages may use either approach, affecting how the variance components (i.e. random effects) should be treated and interpreted (Nakagawa & Schielzeth, 2010). Overall, when focussing on linear regression-type analysis, AIC_C remains the most widely used criterion; it is also the most easily applied because it is implemented in model averaging packages in R [such as *MuMIn* (Bartoń, 2009)] and most other major statistical packages (Symonds & Moussalli, 2010).

Once it has been identified that model averaging is necessary, the next step is to determine which models to average (see Appendix: Step 3). This can be influenced by the question being asked: for example, broad questions, such as whether inbreeding affects fitness, will require a larger model set than more specific questions, such as whether one island exhibits greater fledging success than another island. Under an IT framework, it is assumed that the 'true' model is in the model set (Burnham & Anderson, 2002), but averaging the full model set, or a large proportion of it, is not recommended not only because parameter estimates from models with very poor weights are likely to be spurious (Anderson & Burnham, 2002) but also because the full model set may include redundant models (such as biologically meaningless models or nested models). Indeed, where S (the number of models in the set) is very high relative to N (the sample size), excessive model uncertainty (and thus high error associated with parameter estimation) can be expected and even the best model will have a very small Akaike weight (Burnham & Anderson, 2002). On the other hand, limiting the model set too stringently may result in exclusion of the 'best' model. There are a number of recommendations for the cut-off criterion to use to delineate a 'top model set', such as using the top $2AIC_C$ of models (Burnham & Anderson, 2002), top $6AIC_C$ (Richards, 2008), top $10AIC_C$ (Bolker *et al.*, 2009) or 95% confidence (summed weight, Burnham & Anderson, 2002).

An added complication is how to decide what to do if a particular factor of interest (such as an experimental treatment) is not present in a model captured within the top model set (see Appendix: Step 4). Solutions in such cases are to either conclude that there is little evidence that the factor of interest explains variation in the response variable or extend the cut-off criteria to include at least one model that contains the factor of interest (for example, in cases where a parameter estimate is essential to further analysis). The latter solution may result in very large model sets, and/or inconsistent cut-off criteria for different response variables. High cut-offs are discouraged as they can lead not only to spurious results as described earlier but also to the inclusion of overly complex models (Richards, 2008). Such overly complex models may have similar weight as simpler versions in the set, and model averaging these can potentially result in overweighting the parameters they contain. Simulation studies have shown that removing complex models from the set does not necessarily impact the chance of selecting parsimonious models and also reduces the total number of models selected (Richards *et al.*, 2010). A tentative solution therefore is to exclude models from the set that are more complex versions of those with lower AIC_C (Burnham & Anderson, 2002; Richards, 2008). However, careful scrutiny of these complex models may reveal that they are characterized by the presence of unique predictors of potentially strong biological importance and therefore in such cases should not be removed. Determining how to resolve the issue of nested models is likely to depend on the context of the particular study, but there are currently few clear guidelines on this.

After a top model set is defined, the method used to compute the model-averaged parameters should also be chosen carefully. There are two methods by which the estimate and error for each parameter are weighted (detailed in Burnham & Anderson, 2002; Nakagawa & Freckleton, 2010). In the so-called natural average method (Burnham & Anderson, 2002; p. 152), the parameter estimate for each predictor is averaged only over models in which that predictor appears and is weighted by the summed weights of these models. Alternatively, in the so-called zero method (Burnham & Anderson, 2002), a parameter estimate (and error) of zero is substituted into those models where the given parameter is absent, and the parameter estimate is obtained by averaging over all models in the top model set. Thus, the zero method decreases the effect sizes (and errors) of predictors that only appear in models with small model weights (particularly when the predictors have weak effects), diluting the parameter estimates of

these predictors (shrinkage towards zero) (Lukacs *et al.*, 2010).

Although no clear distinction has been made as to the circumstances under which either of these two methods is more appropriate, Nakagawa & Freckleton (2010) recommend that the zero method should be used when the aim of the study is to determine which factors have the strongest effect on the response variable. Conversely, when there is a particular factor of interest and it is possible that this factor may have a weak effect compared to other covariates, the natural average method should be used to avoid shrinkage towards zero (see Appendix: Step 3). Under the natural average method, the choice of whether to include a parameter of interest is inconsequential, as this method only averages parameters over models in which they appear anyway. Thus, the presence of additional models in the set, that do not include the parameter of interest, will have no influence on the calculation of the effect size or SE of the focal parameter. However, restricting the top model set to only those models that contain a parameter of interest will fix the relative importance of this parameter at 1, making this metric no longer useful (see Appendix: Table S3).

Determining whether the final model provides a good fit to the data presents technical challenges when random factors are present. In the case on nonmixed models, R^2 can be calculated (Burnham & Anderson, 2002), but this is difficult in mixed models (Gelman & Hill, 2007). Further implementation of these methods is required in widely used statistical software such as R.

Interpretation of model estimates

When model-averaged estimates are derived, it is essential to interpret both the direction (positive or negative) of parameter estimates and their magnitudes (effect sizes) in relation to one another (see Appendix: Step 4). Such an assessment can be problematic when input variables are measured on different scales (Gelman, 2008), and interactions are present. Interactions prevent the interpretation of main effects (van de Pol & Wright, 2009), because resultant estimates are usually not comparable to each other. These problems are common with any multiple regression analysis and are not unique to the IT approach *per se*. The process of model averaging can complicate these problems further as it combines parameter estimates derived from models both with and without interaction and polynomial terms (note that the model-averaged intercepts are usually not interpretable). Fortunately, these problems are largely solved by centralizing predictors (see Appendix: Steps 2 and 4), and there is generally a strong justification for doing so, especially where interactions and polynomials are present (Gelman, 2008; Schielzeth, 2010). Centralizing predictors is essential when model averaging is employed, and standardization facilitates the interpretation of the relative strength of parameter estimates.

In linear regression, the interpretation of main effects is impaired when (significant) interactions are present, but this issue is largely resolved if input variables are centred, and inferences are made at points within the biologically meaningful range of the parameter, such as the mean (detailed in Schielzeth, 2010). In addition, it is recommended that input variables (not predictors) are standardized to a mean of 0 and a SD of 0.5 before model analysis (see Appendix: Step 2). The value 0.5 is used, rather than 1 SD, as this allows the standardization of binary predictors [and/or categorical variables, as 'dummy variables' are created (Schielzeth, 2010)] and continuous predictor variables to a common scale (Gelman, 2008; see also Hereford *et al.* (2004) for a discussion of standardization in the context of quantitative genetics). When interpreting the model, it is therefore important to remember that parameter estimates are on this scale. Such standardizations have sometimes been criticized (King, 1986; Bring, 1994; Hereford *et al.*, 2004; Schielzeth, 2010) because parameter estimates are on the transformed scales, which are difficult to interpret biologically. However, back-transformations (described below) of these estimates are straightforward and we recommend that where point estimates of the response variable are derived, authors present them in the original scale (see Appendix: Step 5).

Using the model for prediction

In many cases, the final model is ultimately used to generate a point estimate for the response variable under a given set of circumstances (i.e at fixed points for each predictor variable). In studies of inbreeding, for example, we are interested in comparing the predicted survival point estimates of highly inbred vs. outbred individuals (e.g. Keller & Waller, 2002). There are nearly unlimited combinations of predictor levels ('conditions') that could conceivably be substituted into the model statement to evaluate survival estimates, and the choice of levels made will depend on the question being investigated. For example, one may choose to use a 'worst-case-scenario' (by substituting in extreme values for the predictors) to compare the responses at one site to those of another, to compare conservation management strategies or any others. When predictors have been centred and standardized following the approach of Gelman (2008), one can substitute 0 as the mean and $(x_i - \bar{x})/(2 \cdot \sigma_x)$ for different levels (x_i) of a parameter of interest (with a mean $\bar{x}$ and standard deviation σ_x) (see Appendix: Step 5). It is essential to remember to back-transform the result. Effects of a parameter of interest should be computed at the mean of all other parameters as a matter of routine, to allow comparisons across studies.

Conclusion

The issues presented here are not intended as an exhaustive survey of the practical difficulties associated

with the application of model averaging under an IT framework. For example, this paper has not explored the problems presented by missing data. Model comparisons using IT approaches require data sets with no missing data, as deleting cases containing missing values can severely affect the results of model selection under IT approaches (Nakagawa & Freckleton, 2010). This has been recently covered in detail by other authors (Nakagawa & Freckleton, 2010). Nonetheless, in the current discussion, we have identified a number of areas for more research:

- Which IT criteria should be used when comparing models, given the difficulties presented by including random factors?
- In determining the cut-off for a top model set when examining a factor of interest – how many models is 'too many' for model averaging?
- How should we decide which nested models to remove from the model set?
- How do we quantify model fit in mixed-effects models?

In addition, we emphasize the importance of standardizing variables where model averaging is employed, as to fail to do so renders the results of model averaging uninterpretable in the presence of interactions (c.f. Schielzeth, 2010).

Whereas the debate continues amongst the statisticians in this general area – amongst Frequentists, Information Theoreticians and Bayesians (e.g. Stephens *et al.*, 2005, 2007; Lukacs *et al.*, 2007; McCarthy, 2007) – ecologists and evolutionary biologists continue to derive interesting and important hypotheses, collect data to test their hypotheses, and analyse and (hopefully) publish their results. Resolution of some of the pertinent issues noted above may still be a considerable time away and future work on these problems using simulated data, particularly exploring the use of AIC-based metrics (Box 1), will be a promising area of research. In the meantime, practitioners require pathways and signposts to tentatively guide them through what could be considered the analytical and statistical fog of the new era of information theory and model averaging. Until that fog lifts, it is hoped that the guidelines provided here can improve the consistency and standard of reporting of results in ecological and evolutionary studies using IT approaches.

Acknowledgments

We thank S Richards, J Slate, F Allendorf and H Spencer for their constructive comments on an earlier version of this manuscript. Our research in conservation genetics of threatened New Zealand species is funded by the Department of Conservation (Contract no. 3576), Landcare Research (OBI subcontract no. C09 × 0503), Takahe Recovery Programme and University of Otago. CEG acknowledges the support of a Tertiary Education Commission Top Achiever's Doctoral Scholarship. SN is supported by the Marsden Fund (UOO0812).

References

Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. In: *Second International Symposium on Information Theory* (B.N. Petrov & F. Csaki, eds), pp. 267–281. Akademiai Kiado, Budapest.

Anderson, D.R. & Burnham, K.P. 2002. Avoiding pitfalls when using information-theoretic methods. *J. Wildl. Manage.* **66**: 912–918.

Anderson, D.R., Burnham, K.P. & Thompson, W.L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. *J. Wildl. Manage.* **64**: 912–923.

Bartoń, K. 2009. MuMIn: multi-model inference. R package, version 0.12.2. Available at: http://r-forge.r-project.org/projects/mumin/.

Bates, D. & Maechler, M. 2009. lme4: Linear mixed-effects models using S4 classes. R package, version 0.999375-31. Available at: http://CRAN.R-project.org/package=lme4.

Bolker, B.M. 2009. Learning hierarchical models: advice for the rest of us. *Ecol. Appl.* **19**: 588–592.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. *et al.* 2009. Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol. Evol.* **24**: 127–135.

Bring, J. 1994. How to standardize regression coefficients. *Am. Stat.* **48**: 209–213.

Browne, W.J., Subramanian, S.V., Jones, K. & Goldstein, H. 2005. Variance partitioning in multilevel logistic models that exhibit overdispersion. *J. R. Stat. Soc. Ser. A-Stat. Soc.* **168**: 599–613.

Burnham, K.P. & Anderson, D.R. 1998. *Model Selection and Multimodel Inference*. Springer, Berlin.

Burnham, K.P. & Anderson, D.R. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer, Berlin.

Carstens, B.C., Stoute, H.N. & Reid, N.M. 2009. An information-theoretical approach to phylogeography. *Mol. Ecol.* **18**: 4270–4282.

Claeskens, G. & Hjort, N.L. 2003. The focused information criterion. *J. Am. Stat. Assoc.* **98**: 900–916.

Claeskens, G., Croux, C. & Van Kerckhoven, J. 2007. Prediction-focused model selection for autoregressive models. *Aus. N. Z. J. Stat.* **49**: 359–379.

Congdon, P. 2006a. Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Comput. Stat. Data Anal.* **50**: 346–357.

Congdon, P. 2006b. *Bayesian Statistical Modelling*, 2nd edn. Wiley, West Sussex.

Darwin, C. 1876. *The Effects of Crossing and Self-fertilization in the Vegetable Kingdom*. John Murray, London.

Dochtermann, N. & Jenkins, S. 2010. Developing and evaluating multiple hypotheses in behavioral ecology. *Behav. Ecol. Sociobiol.*, doi: 10.1007/s00265-010-1039-4.

Faraway, J.J. 2005. *Linear Models with R*. CRC Press, Boca Raton.

Fisher, R.A. 1948. *The Theory of Inbreeding*. Oliver and Boyd Ltd, Edinburgh.

Forslund, P. & Pärt, T. 1995. Age and reproduction in birds – hypotheses and tests. *Trends Ecol. Evol.* **10**: 374–378.

Freckleton, R.P. 2010. Dealing with collinearity in behavioral and ecological data: model averaging and the problems of measurement error. *Behav. Ecol. Sociobiol.* **65**: 91–101.

Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Stat. Med.* **27**: 2865–2873.

Gelman, A. & Hill, J. 2007. *Data Analysis Using Regression and Multileval/Hierarchical Models*. Cambridge University Press, Cambridge.

Gelman, A. & Tuerlinckx, F.A. 2000. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Comput. Stat.* **15**: 373–390.

Gelman, A., Su, Y.-S., Yajima, M., Hill, J., Pittau, M.G., Kerman, J. *et al.* 2009. arm: data analysis using regression and multilevel/hierarchical models. R package, version 9.01. Available at: http://CRAN.R-project.org/package=arm.

Grueber, C.E. & Jamieson, I.G. 2008. Quantifying and managing the loss of genetic variation through the use of pedigrees in a non-captive endangered species. *Conserv. Genet.* **9**: 645–651.

Grueber, C.E., Nakagawa, S., Laws, R.J. & Jamieson, I.G. 2010. Inbreeding depression accumulates across life-history stages of the endangered takahe. *Conserv. Biol.* **24**: 1617–1625.

Haldane, J.B.S. 1924. The mathematical theory of natural and artificial selection. Part II: the influence of partial self-fertilization, inbreeding, assortative mating, and selective fertilization on the composition of Mendelian populations, and on natural selection. *Proceedings of the Cambridge Philosophical Society* **1**, 158–163.

Harrell, F.E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logisitic Regression, and Survival Analysis*. Springer, New York.

Hereford, J., Hansen, T.F. & Houle, D. 2004. Comparing strengths of directional selection: how strong is strong? *Evolution* **58**: 2133–2143.

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

Jamieson, I.G., Roy, M.S. & Lettink, M. 2003. Sex specific consequences of recent inbreeding in an ancestrally inbred population of New Zealand takahe. *Conserv. Biol.* **17**: 708–716.

Jefferys, H. 1961. *Theory of Probability*, 3rd edn. University of Oxford Press, Oxford.

Johnson, J.B. & Omland, K.S. 2004. Model selection in ecology and evolution. *Trends Ecol. Evol.* **19**: 101–108.

Keller, L.F. & Waller, D.M. 2002. Inbreeding effects in wild populations. *Trends Ecol. Evol.* **17**: 230–241.

Keller, L.F., Reid, J.M. & Arcese, P. 2008. Testing evolutionary models of senescence in a natural population: age and inbreeding effects on fitness components in song sparrows. *Proc. R. Soc. B* **275**: 597–604.

King, G. 1986. How not to lie with statistics: avoiding common mistakes in quantitative political science. *Am. J. Pol. Sci.* **30**: 666–687.

Konishi, S. & Kitagawa, G. 2008. *Information Criteria and Statistical Modeling*. Springer, New York.

Laws, R.J. & Jamieson, I.G. 2010. Is lack of evidence of inbreeding depression in a threatened New Zealand robin indicative of reduced genetic load? doi: 10.1111/j.1469-1795.2010.00388.x.

Laws, R.J., Townsend, S.M., Nakagawa, S. & Jamieson, I.G. 2010. Limited inbreeding depression in a bottlenecked population is age but not environment dependent. doi: 10.1111/j.1600-048X.2010.05164.x.

Lebreton, J.D., Burnham, K.P., Clobert, J. & Anderson, D.R. 1992. Modeling survival and testing biological hypotheses using marked animals – a unified approach with case studies. *Ecol. Monogr.* **62**: 67–118.

Liang, H., Wu, H.L. & Zou, G.H. 2008. A note on conditional AIC for linear mixed-effects models. *Biometrika* **95**: 773–778.

Link, W.A. & Barker, R.J. 2006. Model weights and the foundations of multimodel inference. *Ecology* **87**: 2626–2635.

Lukacs, P.M., Thompson, W.L., Kendall, W.L., Gould, W.R., Doherty, P.F., Burnham, K.P. *et al.* 2007. Concerns regarding a call for pluralism of information theory and hypothesis testing. *J. Appl. Ecol.* **44**: 456–460.

Lukacs, P.M., Burnham, K.P. & Anderson, D.R. 2010. Model selection bias and Freedman's paradox. *Ann. Inst. Stat. Math.* **62**: 117–125.

McCarthy, M.A. 2007. *Bayesian Methods for Ecology*. Cambridge University Press, Cambridge.

Morton, N.E., Crow, J.F. & Muller, H.J. 1956. An estimate of the mutational damage in man from data on consanguineous marriages. *Proc. Natl. Acad. Sci. USA* **42**: 855–863.

Murtaugh, P.A. 2009. Performance of several variable-selection methods applied to real ecological data. *Ecol. Lett.* **12**: 1061–1068.

Nakagawa, S. & Freckleton, R.P. 2010. Model averaging, missing data and multiple imputation: a case study for behavioural ecology. *Behav. Ecol. Sociobiol.* **65**: 103–116.

Nakagawa, S. & Schielzeth, H. 2010. Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol. Rev.* **85**: 935–936.

Paterson, S. & Lello, J. 2003. Mixed models: getting the best use of parasitological data. *Trends. Parasitol.* **19**: 370–375.

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

R Core Development Team 2009. R: a language and environment for statistical computing. version 2.9.0. Available at: http://www.r-project.org.

Richards, S.A. 2005. Testing ecological theory using the information-theoretic approach: examples and cautionary results. *Ecology* **86**: 2805–2814.

Richards, S.A. 2008. Dealing with overdispersed count data in applied ecology. *J. Appl. Ecol.* **45**: 218–227.

Richards, S.A., Whittingham, M.J. & Stephens, P.A. 2010. Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. *Behav. Ecol. Sociobiol.* **65**: 77–89.

Schielzeth, H. 2010. Simple means to improve the interpretability of regression coefficients. *Meth. Ecol. Evol.* **1**: 103–113.

Schielzeth, H. & Forstmeier, W. 2009. Conclusions beyond support: overconfident estimates in mixed models. *Behav. Ecol.* **20**: 416–420.

Schwarz, G.E. 1978. Estimating the dimension of a model. *Ann. Stat.* **6**: 461–464.

Schwarz, C.J. & Seber, G.A.F. 1999. Estimating animal abundance: review III. *Stat. Sci.* **14**: 427–456.

Spiegelhalter, D.J., Best, N.G., Carlin, B.R. & van der Linde, A. 2002. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B* **64**: 583–616.

Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Del Rio, C.M. 2005. Information theory and hypothesis testing: a call for pluralism. *J. Appl. Ecol.* **42**: 4–12.

Stephens, P.A., Buskirk, S.W., Hayward, G.D. & Del Rio, C.M. 2007. A call for statistical pluralism answered. *J. Appl. Ecol.* **44**: 461–463.

Symonds, M.R.E. & Johnson, C.N. 2008. Species richness and evenness in Australian birds. *Am. Nat.* **171**: 480–490.

Symonds, M.R.E. & Moussalli, A. 2010. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behav. Ecol. Sociobiol.* **65**: 13–21.

Tukey, J.W. 1977. *Exploratory Data Analysis*. Addison Wesley, Reading, PA.

Vaida, F. & Blanchard, S. 2005. Conditional Akaike information for mixed-effects models. *Biometrika* **92**: 351–370.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B. & Freckleton, R.P. 2006. Why do we still use stepwise modelling in ecology and behaviour? *J. Anim. Ecol.* **75**: 1182–1189.

Wright, S. 1922. The effects of inbreeding and crossbreeding on guinea pigs III: crosses between highly inbred families. *US Dept. Agric. Bull.* **1121**: 1–60.

Yang, Y. 2005. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation *Biometrika* **92**: 937–950.

Zuur, A.F., Leno, E.N. & Elphick, C.S. 2010. A protocol for data exploration to avoid common statistical problems. *Meth. Ecol. Evol.* **1**: 3–14.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Table S1 Input data for the working example, comprising 217 observations of fledging success from 64 individuals, across multiple breeding seasons.

Table S2 Full model set of all submodels derived using the dredge function.

Table S3 Summary results of the working example after model averaging using different methodologies.

Figure S1 R model summary output for the global model after standardization.

Figure S2 R output from model averaging of standardized parameters.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Appendix: Worked example for performing model averaging under GLMM in R

This paper explores issues associated with model selection under an IT framework using GLMMs, and we provide here a worked example modelling the effect of inbreeding in an endangered species. Although the worked example focuses on inbreeding depression, the guidelines we present are sufficiently general that they could be applied to any area of study where model averaging is employed. With the advent of molecular markers, and the increasing interest in the conservation and management of small populations, the study of inbreeding has had renewed focus as one of the oldest topics in evolutionary biology (Darwin, 1876; Wright, 1922; Haldane, 1924; Fisher, 1948). The deleterious consequences of matings amongst relatives (inbreeding depression) are normally measured using lethal equivalents, where 1 lethal equivalent is defined as the number of deleterious genes per haploid genome whose cumulative effect is equivalent of 1 lethal gene (Keller & Waller, 2002). We demonstrate how to use the final model for prediction by calculating lethal equivalents (see below). Finally, we chose to perform our analysis using R (R Core Development Team, 2009), as this software is freely available and widely used. Symonds & Moussalli (2010) present a summary of other software packages that permit AIC-based analysis.

Background to the data set

The data were collected over several seasons and consisted of marked individuals, some of which were sampled multiple times. The analysis required model averaging to predict survival estimates for different levels of inbreeding. This example is a real-life conservation problem associated with small island populations of a flightless and highly endangered bird, the takahe (*Porphyrio hochstetteri*) (for further background to the study of inbreeding in this population see Jamieson *et al.*, 2003; Grueber & Jamieson, 2008; Grueber *et al.*, 2010). Here, the response variable is the probability that a hatched takahe egg will successfully fledge. The data set used for this analysis is provided in the Supporting Information (Table S1) and includes 217 observations of hatching (= the number of binomial trials) and fledging (= the number of binomial successes) from 64 individuals (see also Jamieson *et al.*, 2003; Grueber *et al.*, 2010).

Step 1: defining model parameters

The data set used here includes four input variables: (i) age (a continuous variable), (ii) inbreeding coefficient (*f*, coded as '*F*' in the analysis, a continuous variable), (iii) time period since population founding ('YearID': early, mid or late, an ordinal variable) and (iv) island site (a categorical variable with four levels). We controlled for breeding with multiple partners by also including a random factor for individual identity (IndID). Because this random factor has many levels (there are 64 individuals in the data set), but each level has only a few data points, we could not model random slopes. Random intercepts are denoted in the models below as (1|IndID).

In the manuscript (section 'Defining appropriate input and predictor variables'), we discuss the importance of including all interesting predictor variables including plausible (i.e. interpretable) interactions/polynomials in the analysis. Thus in our example the global model includes Age^2, as previous studies have revealed this relationship in bird populations (Forslund & Pärt, 1995), and we include the interaction $f \times$ Age based on other studies that observed this relationship (e.g. Keller *et al.*, 2008). As the response (fledge or not) was coded in the data as a two-column matrix of [Hatch, Fledge], it was recoded in R as the number of [successes, failures] using the function cbind:

$$\text{PrFledge} \leftarrow \text{cbind}(\text{Fledge}, \text{Hatch} - \text{Fledge}).$$

Step 2: generating a model set

To generate a model set in the working example, we first fit a global GLMM using the lmer function implemented in the *lme4* package (Bates & Maechler, 2009). In R, this is defined as:

$$\begin{aligned}\text{global.model} \leftarrow \text{lmer}(&\text{PrFledge} \sim \text{I}(\text{Age}^2) + \text{Age} \\ &+ \text{factor(Island)} + \text{YearID} + \text{F} + \text{F} : \text{Age} \\ &+ (1|\text{IndID}), \text{data} = \text{data}, \text{family} = ''\text{binomial}'')\end{aligned}$$

Once the global model is defined one can standardize the input variables using Gelman's (2008) approach, as this will be essential for interpreting the parameter estimates after model averaging (we detail this approach in the section 'Interpretation of model estimates' of the main manuscript). The standardize function is available within the *arm* package (Gelman *et al.*, 2009):

$$\begin{aligned}\text{stdz.model} \leftarrow \text{standardize}(&\text{global.model}, \\ &\text{standardize.y} = \text{FALSE})\end{aligned}$$

The function summary(stdz.model) can be used to generate a summary of the standardized global model (see Fig. S1), including information criteria (AIC, BIC, raw log likelihood and deviance), as well as details about the random factor. Parameter estimates are also provided, along with their SE and '*z*-scores' (actually these are modified 'half *z*-scores' because the standardization uses 2SD). We remind the reader that although a model may be fitted to the data (without producing an error message), extreme SE values are indicative of a poorly converging model (Bolker *et al.*, 2009).

The next step in generating a full submodel set (including the null model) from the global model is to use the dredge function implemented in the *MuMIn* package (Bartoń, 2009):

$$\text{model.set} \leftarrow \text{dredge}(\text{stdz.model})$$

In the example, this resulted in a total model set (S) of 40 models (Table S2). We chose not to restrict the model set to only those models containing inbreeding. However, results obtained when the model set was restricted are provided in Supplementary Material (Table S3, see also below).

Step 3: model averaging

In the working example, we obtained the top $2AIC_C$ of models using the function get.models implemented in the *MuMIn* package:

$$\text{top.models} \leftarrow \text{get.models}(\text{model.set}, \text{subset} = \text{delta}<2)$$

which results in a set of six models. Using a cut-off of $4AIC_C$ yields 21 models. Alternatively, one could obtain a 95% confidence model set:

$$\begin{aligned}\text{top.models} \leftarrow \text{get.models}(&\text{model.set}, \text{cumsum}(\text{weight}) \\ &\leq 0.95)\end{aligned}$$

which totalled 31 of the 40 possible models. Because of the high number of models in the latter two approaches, we proceed with the $2AIC_C$ cut-off, although for this particular data set, similar effect sizes are reached when using different AIC_C cut-offs (Table S3). This top model set is then averaged using the NA (natural average: nonshrinkage) method rather than the zero method as this example is focussed on the particular effect of inbreeding, and it is possible that this factor may have a weak effect compared to other covariates:

$$\text{model.avg}(\text{top.models}, \text{method} = ''\text{NA}'')$$

For this particular data set, it was observed that the alternative methods of averaging do result in different effect sizes for the parameter of interest (Table S3) and that this should therefore be made carefully (see section 'Model selection and model averaging' of the main manuscript).

Step 4: interpreting model-averaged results

The six models that were included in the 'top model' set are provided in the 'Model summary' of the R output for the model.avg function in the *MuMIn* package (Fig. S2). The model.avg function recalculates the model weights based on the new submodel set of top models. Age^2 is not present in the final model because it was not in the top model set. We interpreted this result as indicating that Age^2 is not a useful predictor of fledging success in takahe. The results of the model averaging are summarized in Table A1; remember that the parameter estimates are standardized effect sizes and are therefore on a comparable scale.

Table A1 Summary results of the working example after model averaging: effects of each parameter on fledging success in takahe (*Porphyrio hochstetteri*).

Parameter	Estimate*	Unconditional SE	Confidence interval	Relative importance
(Intercept)	0.146	0.265	(−0.374, 0.666)	
Island2†	−0.745	0.310	(−1.35, −0.138)	0.25
Island3	−0.572	0.371	(−1.30, 0.154)	''
Island4	−0.448	0.642	(−1.71, 0.811)	''
Age	0.500	0.287	(−0.063, 1.06)	1.00
f	−0.538	0.314	(−1.15, 0.079)	0.71
YearID	−0.117	0.290	(−0.686, 0.451)	0.10
Age × *f*	−1.190	0.732	(−2.63, 0.243)	0.38

*Effect sizes have been standardized on two SD following Gelman (2008).

†Island1 was the reference category.

It is most useful to report unconditional SE because it incorporates model selection uncertainty (Table S1), as opposed to standard SE which only considers sampling variance. If extreme SE or confidence intervals occur, this is indicative of at least one of the models in the set failing to converge (Bolker *et al.*, 2009). In the worked example, Age was the most important predictor with *f* (inbreeding coefficient) having 71% relative importance to Age. All confidence intervals for the parameter estimates include zero, so there is little evidence in this example that any of the predictor variables affect fledging success (Table S1). However, it could still be relevant to use the model to predict point estimates of survival for certain conditions (see below).

Step 5: using the model for prediction

Here we demonstrate using the model for prediction by calculating lethal equivalents. Given that the log of overall fitness is expected to decline linearly with increases in the inbreeding coefficient *f*, the slope of this relationship (−*B*) is used as a standardized measure of inbreeding depression (Keller & Waller, 2002). This estimate was first calculated by Morton *et al.* (1956) using linear regression and eqn A1:

$$B = -\ln(S_f/S_0)/f \qquad \text{(A1)}$$

where S_f is the probability of survival at inbreeding level *f* (by convention 0.25, first-order relatives) and S_0 is the probability of survival at $f = 0$, with $2B$ equal to the number of lethal equivalents per diploid organism. The final GLMM we have derived in the worked example allows us to calculate lethal equivalents including environmental and demographic factors, as well as the random factor.

Using the parameter estimates from the final model (Table S1), we calculate lethal equivalents by deriving point estimates to compare fledging probability when $f = 0$ (the breeder is not inbred) and $f = 0.25$ (the parents of the breeder were first-degree relatives). To make such point estimates from a complex model, one must specify fixed levels for each of the covariates in the final averaged model. In the current example, we make estimates at the population mean for all other parameters that were found to be important in the final model (i.e. Island, Age and Year ID), as this is likely to provide the most useful comparison to other, similar studies.

Bearing in mind that the predictors have been standardized to a mean of 0 and SD of 0.5 (Gelman, 2008), it is important to solve the model by substituting *standardised* predictors, i.e. 0 for the mean, or $(x_i - \bar{x})/(2 \cdot \sigma_x)$ for other values (i). In this data set, the mean of $f = 0.0316$ and the SD = 0.0600, calculated from the input file. Thus, we solve the model for both $f = 0$ and 0.25, at the mean of all other parameters (using a weighted mean for the categorical factor of island). For example, the predicted survival when $f = 0$ on island 1, is (using the figures in Table A1):

$$\begin{aligned} p(\text{Fledge}_{\text{Island1}}) = {} & 0.146 - 0.745 * 0 - 0.572 * 0 - 0.448 * \\ & 0 + 0.500 * 0 - 0.538 * ([0 - 0.0316]/ \\ & [2 * 0.0600]) - 0.117 * 0 - 1.190 * 0 * 0 \\ = {} & 0.288 \end{aligned}$$

The weighted average of survival estimates across all islands is:

$$\begin{aligned} p(\text{Fledge}) = {} & (0.288 * N_{Island1} - 0.457 * N_{Island2} \\ & + -0.284 * N_{Island3} - 0.160 * N_{Island4})/N_{total} \\ = {} & (0.288 * 11 - 0.458 * 22 - 0.284 * 21 \\ & - 0.160 * 10)/64 \\ = {} & -0.226 \end{aligned}$$

When $f = 0.25$ survival estimates are also calculated for each island in turn using the method above, and the weighted average across islands is −1.347.

As this example models a binomial response variable (fitted with a logit link), these point estimates are the probability of success on a logit scale. We back-transform using:

$$p = 1/(1 + 1/e^x) \qquad \text{(A2)}$$

where x is the probability of survival on the logit scale. The invlogit function (available in the package *arm* [Gelman *et al.*, 2009]) can perform this calculation in R. Thus, the probability that a fledged egg will hatch (the 'conditional survival') when the parental $f = 0$ and at the mean of all other parameters is 0.444. The probability when the parental $f = 0.25$ (equivalent to sib–sib mating) is 0.206, only 46% of the fledging success of outbred individuals. In studies of inbreeding, these values are normally substituted into eqn 1 to calculate lethal equivalents, and in this example, $2B = 6.1$. In addition to any other inferences made from a final model, we propose that point estimates should always be calculated using means for covariates and weighted means for factors, to permit comparisons across study populations.

Uncertainty of these point estimates can also be established, as the model.avg function in R outputs the lower and upper bounds of the confidence intervals for each parameter estimate (see Supplementary Material Fig. S2). These values can be substituted into a model formula as 'parameter estimates' to generate predicted survival estimates at both the lower and upper bounds of the 95% confidence interval. Following our worked example, where $f = 0$, we generate the lower bound of the 95% confidence interval for fledging probability on Island 1 thus (using the figures in Table A1):

$$\begin{aligned}\text{Lower } 95\%\ \text{CI}(\text{Fledge}_{\text{Island1}}) &= -0.374 - 1.350 * 0 \\ &\quad - 1.300 * 0 - 1.710 * 0 + 0.063 * 0 \\ &\quad - 1.150 * ([0 - 0.0316]/[2 * 0.0600]) \\ &\quad - 0.686 * 0 - 2.63 * 0 * ([0 - 0.0316]/[2 * 0.0600]) \\ &\quad = -0.071\end{aligned}$$

Each island must be computed separately, and then a weighted average obtained:

$$\begin{aligned}&(-0.071 * N_{\text{Island1}} - 1.421 * N_{\text{Island2}} - 1.371 * N_{\text{Island3}} \\ &\quad - 1.781 * N_{\text{Island4}})/N_{\text{total}} \\ &\quad = (-0.071 * 11 - 1.421 * 22 - 1.371 * 21 \\ &\quad - 1.781 * 10)/64 \\ &\quad = -1.229\end{aligned}$$

Again, this will need to be back-transformed (inverse logited), to give 0.226: the lower bound of the 95% confidence interval for the predicted probability of fledging when $f = 0$. The upper bound can be computed similarly, and the confidence intervals in this example are asymmetrical, because the response is binomial and as such bound between 0 and 1.

Finishing our calculation of lethal equivalents, the upper and lower 95% confidence bounds of survival probability for $f = 0$ and $f = 0.25$ can be substituted into eqn 1 to generate upper and lower bounds for the 95% confidence interval of lethal equivalents:

$$\begin{aligned}\text{Lower bound } 95\%\ \text{CI: } B &= -\ln(0.026/0.226)/0.25 \\ &= 8.66\end{aligned}$$

$$\begin{aligned}\text{Upper bound } 95\%\ \text{CI: } B &= -\ln(0.719/0.685)/0.25 \\ &= -0.196\end{aligned}$$

Note that the 'lower bound' produces a positive value, and the 'upper bound' produces a negative value. This is because of the sign change in eqn 1 so the lower bound should be interpreted as 'maximal inbreeding depression' and the upper bound as 'minimal inbreeding depression'. Here, the confidence interval for lethal equivalents includes zero, consistent with the observation that the confidence interval for the parameter estimate included zero (Table S1). It should be noted that these methods provide only approximated confidence intervals and that more work is needed to improve these approximation methods.

Received 1 July 2010; revised 22 November 2010; accepted 25 November 2010