SC/66a/EM/1

# Why BIC is not always (and may never be) an appropriate criterion for selecting terms in complex ecological models

J.P. McKinlay and W.K. De La Mare

INTERNATIONAL
WHALING COMMISSION

# Why BIC is not always (and may never be) an appropriate criterion for selecting terms in complex ecological models

J. P. MCKINLAY and W. K. de la MARE
Australian Antarctic Division, 203 Channel Highway, Kingston, Australia, 7050.
*Contact email: bill.delamare@aad.gov.au*

**Abstract**

At recent meetings of the Scientific Committee the question of whether to prefer AIC or BIC in model selection has been a matter of contention. Drawing on both the statistical literature and results from a simulation experiment conditioned on analyses of minke whale nutritive condition, this paper examines the relative merits of model selection based on AIC or BIC and provides some recommendations on appropriate practice in the development and presentation of statistical analyses that use model selection. The conclusion is that the choice of which information criterion to use depends on the purpose of the analysis, the sample size and specifics of the realised experimental design. In the specific case of analyses of minke whale biological parameters, results indicate that, based on current sample sizes, it is likely that BIC is under-estimating the complexity necessary to adequately describe system behaviour. Model misspecification of this kind is critical given that models are then used to calculate prospective lethal-take sample sizes under research permits.

**Introduction**

In recent years there has been protracted discussion by the Scientific Committee (SC) and its working groups concerning statistical model selection, particularly in relation to complex models of biological parameters. Much of this discussion has centred on whether Akaike's Information Criterion (AIC; Akaike 1973) or Schwarz' Bayesian Information Criterion (BIC; Schwarz 1978) is more appropriate for model selection. While there exists a large number of information criteria (IC) tailored to particular models or data types (e.g. Ward 2008, Hooten and Hobbs 2015), AIC and BIC remain the best known and most widely used. They are superficially quite similar, philosophically quite different, and can provide divergent results in some circumstances. In this paper we examine recent advice provided to the SC in relation to the use of AIC and BIC for model selection, provide some advice of our own, and consider what effect ignoring that advice would have on a simulated model selection exercise conditioned on the analyses we undertook in de la Mare *et al*. (2014). For the purposes of this work we define model selection in the restricted sense of selecting a covariate set to adequately explain a response, excluding matters relating to the distribution of response or covariates, which are assumed fixed between competing models. 'Adequately', in this sense, is central to the aim of this paper, and can broadly be thought of as balancing goodness of fit with simplicity.

It is worth reviewing recent advice to the SC on the topic of model selection. Most recently, the "Report of the expert workshop to review the Japanese JARPA II Special Permit research program" (SC/65b/Rep02) provided relatively detailed advice in relation to analyses of nutritive condition of minke whales, which we feel is important enough to reproduce in full below:

> "Given the discussion in a number of papers presented by both proponents and observers, the Panel offers the following summary of the merits or otherwise of AIC and BIC, both commonly used metrics for model selection. The two metrics, while mathematically similar (twice the negative log-likelihood plus a penalty term), arise from different underlying arguments (Kass and Raftery, 1995). BIC will tend to select simpler models than AIC if the number of data points exceeds 7, all things being equal (because the penalty term for additional parameters is larger than for AIC). Simulations have shown that AIC often selects a more complex (wrong) model over the simpler (and correct) model (Kass and Raftery, 1995). Whether this necessarily means that BIC is always better than AIC is not clear because AIC attempts to find the best approximating model rather than true model (which would be rarely in the set of candidate models). BIC is, however, generally preferred to AIC for 'large' data sets. However, what constitutes 'large' in any particular case depends inter alia on the set of models under consideration. Thus, for any one problem the selection between AIC and BIC is seldom definitive. Many practicing statisticians consequently often apply both AIC and BIC, examine the sensitivity to the different models selected and apply expert judgement." (SC/65b/Rep02, p37)

We commend the Expert Panel for providing guidance in relation to this difficult topic, and find that with one or two exceptions we largely agree with the explanation given above, at least in the context that Kass and Raftery (1995) presented their work. We particularly appreciate the Expert Panel's comments that a choice between AIC or BIC is seldom definitive, is influenced by sample size, and that sample size considerations are linked to the nature and complexity of the system being studied. The Panel seems to indicate that concordance between IC indicates some degree of reassurance that things haven't gone too awry, while lack of concordance puts us into the arena of "expert judgement". Of course, one person's expert judgement might be seen as lack of judgement by

another, so where does this really leave us? Luckily, there is a growing body of evidence in the ecological literature about the relative performance of different IC, a point we explore in the next section.

We wish to bring to the attention of the SC that the advice provided by the JARPA II review panel was specifically in relation to analyses of blubber thickness and related metrics of body condition (Konishi and Walloe 2014; SC/F14/J13) and stomach contents (Konishi *et al*. 2014; SC/F14/J14). Nutritive data were reanalysed, ostensibly following the advice of the Expert Panel, by the Working Group on Ecosystem Modelling in 2014 (IWC/65/Rep01, Annex K1, Appendix 2). We found it unfortunate that these new analyses only seemed to follow some of the Expert Panel's advice. In particular, the new work only reported BIC values, gave no assessment of concordance or otherwise between AIC and BIC, and no sensitivity analyses in the event that the criteria selected appreciably different models. We hope the present paper will serve to highlight the merit of the Panel's advice.

In the remainder of this paper we set out our rationale for why we think that BIC should not be used as the sole criterion for variable selection in developing models of complex ecological systems. We first contrast AIC and BIC by providing their formulaic (but not theoretical) derivation, in each case paying particular attention to the size of the penalty term that serves to balance model fit with model complexity. In this section we also consider the special case of nested models, showing the link between the degree of improvement in IC required to retain a model term and the likelihood ratio test. For the nested case, this demonstrates the notional *p*-value that would be required for each of AIC and BIC to reject a simple model in favour of more complex one. In the next section we draw upon the modern statistical literature, as well as two decades of applied ecological experience since the publication of Kass and Raftery (1995), to show that opinions about use of BIC are divided. On balance, this review indicates that BIC is generally not supported for developing models in complex ecological systems. We follow this with our own simulation study into the behaviour of AIC and BIC using data conditioned on models similar to those used for nutritive analyses of minke whales. Our results indicate that BIC is unsuitable as the sole determinant of model complexity for the models we considered. We comment briefly about the probity of significance testing model terms after model selection, before offering some concluding recommendations. While noting that it is difficult to be entirely prescriptive in matters of model selection, we suggest a minimum set of 'best practice' approaches for using and reporting model selection based on IC.

**AIC and BIC Compared**

Consider a set of candidate models that we wish to compare in order to select the best model (for some defined meaning of 'best'). Assume $y$ are our observed data, described by a density $p(y|\Theta)$ as a function of parameter set $\Theta$. Define the deviance as $D(\Theta) = -2\log\{ p(y|\Theta)\}$. Classical model comparisons for nested models are achieved by a likelihood ratio test comparing the difference in log-likelihoods to a chi-squared distribution with degrees of freedom equal to the difference in free parameters between the competing models. For non-nested models, one alternative is the Akaike Information Criterion (AIC) given by

$$\mathrm{AIC} = -2\log\{p(y|\hat{\theta})\} + 2k$$

where $\hat{\theta}$ is the maximum likelihood estimate and $k$ is the number of parameters in the model. Based on Kullback–Leibler divergence (a concept from information theory that quantifies information loss), AIC provides a relative measure of fit of a model to a given set of data. Assuming all else being equal (i.e. near-equivalent model checking diagnostics), the quality of different models can be assessed by comparing AIC values. In the case of nested models, Murtaugh (2014) shows the close relationship that exists between differences in AIC values and traditional p-values derived from likelihood ratio tests. AIC is known to perform poorly if $k$ is large in relation to sample size ($n$), leading to the finite-sample adjustment criterion $\mathrm{AIC}_c$ (Hurvich and Tsai 1989, Sugiura 1978). For the purposes of illustration we present the form for a univariate linear model, noting that the complexity of calculation changes with model type.

$$\mathrm{AIC}_c = -2\log\{p(y|\hat{\theta})\} + 2k + \frac{2k(k+1)}{n-k-1}$$

Burnham and Anderson (2002) recommend that $\mathrm{AIC}_c$ be routinely used over AIC since the differences are important at small sample sizes and the two criteria correspond asymptotically.

An alternate metric for comparing between models is Swartz's Bayesian Information Criterion (BIC), given by

$$\mathrm{BIC} = -2\log\{p(y|\hat{\theta})\} + k\log(n)$$

The BIC is derived as an easy approximation to the log of a Bayes factor, a concept closely linked to Bayesian hypothesis testing. Asymptotically, it favours models that correspond to the most probable given the data based on the Bayesian posterior. Recent work by Flynn *et al*. (2011) suggests finite-sample performance of BIC may, like AIC, be poor in some circumstances, however this work in not widely available in software and is not considered further here.

For a given criterion, one typically computes the IC values for a set of candidate models then selects the model with the smallest value (or potentially more than one model if several have similar IC values). An important thing to note in contrasting the AIC and BIC is that both criteria are composed of minus twice the log-likelihood (-2LL) plus a penalty term designed to balance model fit with model complexity. In the case of AIC, this penalty is twice the number of parameters in the model, so more parameters incurs higher penalty. Stone (1977) demonstrated that this is asymptotically equivalent to conducting model choice by leave-one-out cross-validation, which leads naturally to the conclusion that AIC is tailored to fit the observed data well (asymptotic efficiency), and predict new data well.

In contrast, the BIC is said to be asymptotically consistent; that is, as sample size increases to infinity it will select the correct model from a candidate set, providing the correct model is in the set (Shao, 1997). The penalty term for BIC links the number of parameters to the log of sample size; penalties are in general greater in comparison with AIC and smaller models (fewer parameters) are favoured. AIC and BIC correspond only at very small sample sizes (around 7). These asymptotic characteristics lead to the conclusion that AIC and BIC will impact components of the bias-variance trade-off differently. Optimal models determined by AIC will asymptotically have lower variance but higher bias than optimal models determined by BIC, and vice versa. The optimality aspects of AIC and BIC, efficiency and consistency respectively, are not thought able to be shared by an individual criterion, as shown by Yang (2005) in a regression framework. While in some sense asymptotic properties are academic, many authors use the concept to justify one IC or another (e.g. an appeal to finding the true model as $n$ approaches infinity, in the case of BIC). In applied situations it is challenging to identify what sample size will approach asymptotic behaviour, or the degree to which our measured variables adequately represent the system under study.

It is of interest to consider the comparative size of penalty between AIC and BIC for the same model structure fitted to the same data. We do this by discounting the common -2LL between criteria and evaluate the penalty over a grid of parameter and sample size values (Figure 1).

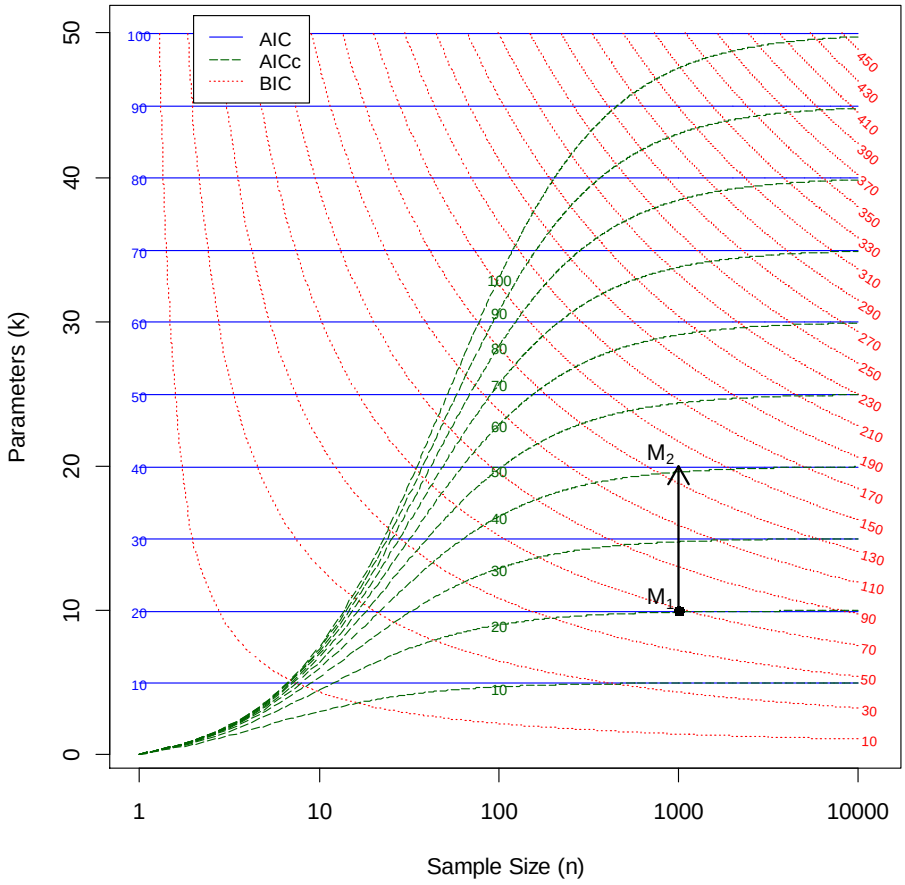


Figure 1. Magnitude of IC penalty term (contour lines) evaluated over a grid of sample size and number of model parameters. The arrow indicates the respective complexity penalties that would be applied to models comprised of 10 parameters ($M_1$) and 20 parameters ($M_2$), each fitted to 1000 data points.

Consider a comparison between a model with 10 parameters ($M_1$) and a model with 20 parameters ($M_2$), both fitted to the same data ($n$=1000). Under AIC the penalty for $M_1$ is 20 and the penalty for $M_2$ is 40, a difference in penalty of 20 units. In contrast, under BIC the penalty for $M_1$ is about 70, for $M_2$ it's about 138, giving a difference in penalty of around 68 units. In this example, this means that under AIC model selection the improvement in fit (-2LL) between the smaller and larger model must be at least 20, the difference in penalties, before the larger model can be considered an equivalent or superior fit. Similarly, under BIC the improvement in fit would need to exceed 68 before the larger model was considered an equivalent or better fit.

How do we decide between what constitutes an equivalent model (in which case we would prefer the simpler model) and what constitutes an improved model? The degree to which an improvement in fit exceeds the difference in penalties is measured by the difference in AIC or BIC values themselves. Authors differ in their views about how large changes in IC ($\Delta$AIC or $\Delta$BIC) must be to prefer one model over another. Murtaugh (2014) provides a summary of published advice for interpreting $\Delta$AIC, which ranges from weak support (1-2) to very strong support (10-14) for a difference between models. Values of $\Delta$BIC are interpreted on a similar scale (Raftery and Kass 1995). We note that this type of advice is subjective, in a similar way that adopting $\alpha$=0.05 is subjective in conventional testing.

It is possible to contrast $\Delta$AIC and $\Delta$BIC values with conventional $p$-values from likelihood ratio tests, at least in the case of nested models. While we recognise significance testing and model selection using IC derive from different underlying paradigms, it is nonetheless useful to cast IC in terms of $p$-values since the latter are widely used and understood. For the nested model situation, where $k_d$ is the number of parameters difference between the larger and smaller model, it is straightforward to calculate the notional $p$-values for which AIC and BIC will retain an extra term by producing an equivalently small IC value (i.e. the larger model will produce an equivalently small AIC value as the simpler model). As discussed previously, most practitioners generally require an improvement (reduction) in AIC or BIC values before a more complex model is considered superior to a smaller model. For our purposes we assess differences of 2 and 8, generally thought to offer relatively weak and relatively strong evidence for a difference between models.

We assess AIC for the notional $p$-value (under a likelihood ratio test) that would allow a single parameter ($k_d$=1), a moderate interaction (e.g. between a three- and six-level categorical variable, $k_d$=10) and a complex interaction (e.g. 3-level x 6-level x 3-level, $k_d$=20) to enter a model (Table 1). The R code is straightforward:

```
aicsig <- function(kd, diff.aic) 1-pchisq(kd*2+diff.aic, df=kd)
outer(c(1, 10, 20), c(2, 8), "aicsig")
```

Table 1. Equivalent likelihood ratio test significance levels that would be required to retain terms with $k_d$ parameters, as determined from the AIC penalty term and an assumed $\Delta$AIC of 2 or 8.

| Type | # Parameters | $\Delta$AIC = 2 | $\Delta$AIC = 8 |
|---|---|---|---|
| Single parameter | $k_d$ = 1 | 0.045500 | 0.001565 |
| Moderate interaction | $k_d$ = 10 | 0.015105 | 0.001805 |
| Complex interaction | $k_d$ = 20 | 0.002766 | 0.000425 |

These results show that for a weak improvement in AIC (i.e. 2) a model term costing one degree of freedom will be retained if the term has a notional significance level of $p < 0.0455$, a value quite close to the conventional level of 0.05. Similarly, moderate sized interaction (10 df) will be retained with significance $p < 0.0152$, and a complex interaction (20 df) with $p < 0.0028$. This ignores a finite-sample correction that is negligible for $n > 50$ (Murtaugh 2014). It is clear that these significance levels are comparable to conventional standards for low numbers of parameters, but more stringent for complex terms. Larger improvements in AIC lead to higher notional significance levels under the likelihood ratio test paradigm. For example, if we were to require an improvement in AIC of 8 (strong evidence) before we retained a complex interaction, this would equate to requiring a significance level of $p < 0.0005$ by conventional standards.

In contrast, BIC is dependent not only on $k$ but also on sample size $n$, here evaluated for small (100), medium (1000) and large (10 000) sample sizes (Table 2). Again, the R code is straightforward so we reproduce it in full:

```
bicsig <- function(n, kd, diff.bic) {1-pchisq(kd*log(n)+diff.bic, kd)}
outer(c(100, 1000, 10000), c(1, 10, 20), "bicsig", diff.bic=2)
outer(c(100, 1000, 10000), c(1, 10, 20), "bicsig", diff.bic=8)
```

Table 2. Equivalent likelihood ratio test significance levels that would be required to retain terms with $k_d$ parameters, at small, moderate and large sample sizes, as determined from the BIC penalty term and an assumed ΔBIC of 2 or 8.

| | Type | # Parameters | Small *n* (100) | Medium *n* (1000) | Large *n* (10 000) |
|---|---|---|---|---|---|
| ΔBIC = 2 | Single parameter | $k_d = 1$ | 0.010168 | 6.0732e-07 | 1.408607e-11 |
| | Moderate interaction | $k_d = 10$ | 0.002840 | 2.744927e-11 | <0. 1e-16 |
| | Complex interaction | $k_d = 20$ | 0.000813 | 7.771561e-16 | <0. 1e-16 |
| ΔBIC = 8 | Single parameter | $k_d = 1$ | 0.000385 | 4.74585e-08 | 1.20703e-12 |
| | Moderate interaction | $k_d = 10$ | 0.000113 | 1.87250e-12 | <0. 1e-16 |
| | Complex interaction | $k_d = 20$ | 0.000033 | <0. 1e-16 | <0. 1e-16 |

Results show that BIC is appreciably more conservative than AIC for retaining an additional single parameter at small sample sizes ($p < 0.0102$ and $p < 0.0004$ for ΔBIC 2 and 8, respectively), with the required *p*-value increases with the number of parameters $k_d$, sample size, and ΔBIC. What is perhaps not well appreciated is that conservatism increases sharply with moderate increases in sample size, particularly so for interactions. The number of formal coefficients for a given interaction of categorical variables is given by:

$$k_{a\times b} = (k_a - 1)(k_b - 1)$$

where $k_a$ and $k_b$ are the number of levels for categorical variables *a* and *b*. For interactions involving continuous variates (separate slopes and intercepts models) the number of formal coefficients is given by:

$$k_{a\times b} = 2(k_b - 1)$$

where *b* is a categorical variable for which separate slopes and intercepts are to be estimated.

Consequently, the number of coefficients increase quite quickly for even modest numbers of levels for categorical variables and the equivalent likelihood ratio test significance levels are virtually zero for retaining even moderate interactions. At medium sample sizes ($n$=1000), and with a requirement for weak evidence (ΔBIC = 2), even single degree of freedom parameters will only be retained in models if they exceed a conventional significance threshold of $p < 0.0000006$! Criticisms are often levelled at the use and interpretation of *p*-values (e.g. Barker and Ogle 2014, Lavine 2014), and we are sympathetic to some of those concerns. In particular, we are not in favour of interpreting *p*-values as definitive in isolation of effect size and sample size considerations. Nonetheless, we believe the notional *p*-value thresholds for accepting additional terms under BIC, even at modest sample sizes, are much more conservative than typical statistical practice would indicate.

**What does the literature say?**

So which is better, AIC or BIC? That depends on who you ask. From an applied perspective in the ecological literature, AIC is a clear winner in terms of popularity. In a Web of Science search covering 1993-2013, Aho *et al*. (2014) surveyed ecological publications that used IC and found 139 (84%) used AIC, 23 (14%) used BIC and only 4 (2%) used other IC. In a comparative sense, there is a large literature contrasting AIC, BIC and other IC metrics with real and simulated data (Gelman *et al*. 2013, Ward 2008, Symonds and Moussalli 2011). Results from these studies sometimes reflect author's experiences, philosophical preferences, the types of data analysed, or the limitations of simulation frameworks. Perhaps a more relevant question is which IC is better suited for a particular task. Shao (1997), studying the asymptotic behaviour of AIC and BIC in linear models, found that AIC-like measures are most useful when there is no fixed-dimensional correct model, and BIC-like measures are to be preferred when a fixed-dimension true model exists. Aho *et al*. (2014) reviewed this question in the context of ecological applications. They point out that it is now generally accepted that there are two classes of model selection procedures: a class that lead in predictive accuracy (AIC and similar), and a class of confirmation/falsification methods that are consistent (BIC and similar) (Table 3).

Table 3. The worlds of AIC and BIC contrasted (reproduced from Aho *et al*. 2014).

| ***Factor*** | ***AIC*** | ***BIC*** |
|---|---|---|
| Mathematical characteristics | | |
| Derivation | Estimated information loss. | Approximate Bayes factor. |
| Optimality criterion | Asymptotic efficiency. | Asymptotic consistency. |
| Close cousins | Data splitting, Mallows' Cp , PRESS. | Hannan-Quinn, Geweke and Meese, Bayes factors, and Bayesian hypothesis testing. |
| World View | | |
| Problem statement | Multiple incompletely specified or infinite parameter models. | A small number of completely specified models/hypotheses. |
| Perspective | ''All models are wrong, but some are useful.'' | ''Which model is correct?'' |
| Simulation structure | d >> n | d << n |
| With increased n . . . | Best model grows more complex. | Procedure focuses in on one best model. |
| Applications | | |
| Context | Exploratory analysis; model selection to address which model will best predict the next sample; imprecise modelling; tapering effects. | Confirmatory analysis; hypothesis testing; model selection to address which model generated the data; Low dimension, precisely specified models. |
| Ecological examples | Complex model selection applications, e.g., predictive models for community, landscape, and ecosystem ecology; time series applications including forecasting. | Controlled experiments, for instance in physiology/enzymatics/genetics with a limited number of important, well-understood, biological predictors; models including expected or default (null) frameworks, e.g., enzyme kinetics models, Hardy-Weinberg equilibrium, or RAD curves, one of which is expected to be correct. |

Notes: The number of parameters in the true model is *d*; sample size is *n*. Abbreviations are: PRESS, predicted residual sum of squares; and RAD, ranked abundance distribution.

According to Aho *et al*. (2014), AIC is typified by a world view that sees reality as complex (high dimensional), not easily characterised (we typically don't measure everything required to capture the complexity), and not sufficiently resolved (our sample sizes are too small). In contrast, BIC lives in a world of relatively low dimension, where sample sizes are generous in comparison to the number of parameters required to adequately capture system behaviour, and where underlying theory posits a small number of precisely specified candidate models. Into which of these 'world views' might the estimation of trends in minke whale nutritive condition fall? We contend that a very complex reality generates the data (the time- and space-varying Southern Ocean ecosystem), that we have relatively few data to attempt to describe these processes (d > n), and that we are unlikely to captured all the variables that would need to be measured to represent the system. We believe that current models, built from relatively few data, contain slight to moderate effects (i.e. model terms providing slight to moderate improvement in model fit) that are systemically important but that are not sufficiently retained under the penalisation imposed by BIC model selection. Burnham and Anderson (2002) introduce the concept of tapering effects to describe this gradation in effect size. A feature of complex systems with tapering effects is that increased model complexity can be expected with increased sample size as small to moderate effects become increasingly resolved.

Returning to the Expert Panel advice regarding IC, a less-than-careful reading of the Panel's advice might seem to imply that BIC should be preferred over AIC. However, qualifications about the availability of the true model in the candidate set (a point we are not so concerned about, see below) and what constitutes a large sample size (which cannot be easily assessed in practice), any such preference is not clear-cut. The reference provided by the Panel, namely Kass and Raftery (1995), does offer a strong critique against the use of AIC in favour of BIC, and uses several applied analysis examples to support their case. However, the description of those supporting analyses indicates they are in fact well suited to the type of problems BIC is designed to assess; low dimensional systems with a few strong effects. Had they assessed a different class of problem, we wonder if their conclusions would have been similar.

One mainstay defence of the BIC approach is an appeal to Okham's razor, which postulates we should develop models according to the 'law of parsimony'; as simple as necessary, but no more. On this point it is pertinent to note the view of prominent Bayesian Andrew Gelman (2009) who, in a discussion of Robert *et al*. (2009), says that:

> *In the social science problems I've seen, Ockham's razor is at best an irrelevance and at worst can lead to acceptance of models that are missing key features that the data could actually provide information on. As such, I am no fan of methods such as BIC that attempt to justify the use of simple models that do not fit observed data*.

We find this last point particularly compelling; models should provide an adequate fit to observed data and predict future data well, and if they do not they should be considered unsatisfactory. Gelman and Rubin (1995) develop a convincing argument, based on the implicit prior on $\Theta$ assumed by BIC, against routinely using BIC for model selection. An even more polarised view is given by Burnham and Anderson (2011) who, dealing predominantly with ecological data, say that "*There are a host of reasons why BIC is a poor criterion; we believe it should not be used with real data*". Much of their argument centres on the idea that in ecological contexts the true generating model is almost never contained within the candidate set of models considered. We find this point disputable (see for example Casella *et al*. 2009), and the position "never use BIC with real data" too much a generalisation. Instead, we prefer the view that AIC and BIC are simply suited to different situations. Finally on this point, it is worth noting that the jury is out on the consistency of Bayes factors, and by extension the BIC, when the number of important regressors increases with sample size (Moreno *et al*. 2010, Wang and Maruyama 2015). This remains an active area of statistical research.

We also recognise that the use of AIC is not without its perils. Arnold (2010) and Freckleton (2011) warn against inclusion of uninformative parameters in models selected by AIC and provide guidance to solutions for this problem, including careful consideration of models and regressors, avoiding automated step-wise selection procedures and considering model averaging. We feel this is generally sound advice. The Expert Panel also provided advice in this regard (though in relation to model selection in general, not AIC or BIC in particular), namely that the correlation structure of covariates should be examined with a view to determining an uncorrelated subset either by excluding highly correlated variables or by developing new uncorrelated covariates as substitutes (e.g. by using principal components analysis, PCA) (SC/65b/Rep02, p36). While we agree that the correlation between covariates should be examined carefully during model building, we would like to offer a cautionary note that developing uncorrelated predictors that simultaneously preserves data structure is not always possible in any practical sense. For example, how does one determine uncorrelated predictor sets when patterns of correlation may be dependent on several other, potentially interacting, predictors? In the presence of interactions we are therefore cautious about the use of PCA or other dimension reduction techniques unless there are well developed subject-matter driven arguments to do so. The utility of exploratory analysis in examining predictor and covariate structure prior to developing formal models cannot be emphasised enough.

There are many different approaches to model selection and, of the information theoretic class of procedures, AIC and BIC are simply two of a wide and diverse family (e.g. see Konishi and Kitagawa 2008, Rao and Wu 2001, Ward 2008). It was never our intention to provide a review of all IC model selection procedures, but rather to point out that the two most widely used IC have appreciably different properties and are best suited to different research problems.

**A case in point: AIC versus BIC in condition analyses of minke whales**

We conducted a simulation study based on JARPA / JARPA II patterns of data collection to examine the frequency that model selections based on either AIC or BIC choose the "correct" model. We used the R function stepAIC from the recommended MASS package in R v3.1.3 (Venables and Ripley 2002, R Core Team 2015). While we would not normally advocate automated application of stepwise model selection procedures on real data, we find the approach acceptable for assessment using simulations. We note that stepAIC respects the marginality of terms during addition/deletion. Stepwise selection based on BIC is achieved by providing the appropriate penalty term.

Simulated data were generated using the standard linear modelling framework

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where **y** is a vector of simulated data, **X** is the model matrix, **α** is a vector of model coefficients and **ε** is a vector of i.i.d random normal errors. Only linear models can be tested because there is no stepwise selection procedure available for mixed effects models in R. In any case, in mixed effects modelling one needs to decide whether selection is focussed on the fixed effects or the random effects.

In the trials reported here the model matrix is derived either from model fit d3 (the de la Mare *et al*. 2014 low AIC model with several interactions) or J17 (the Konishi and Walloe 2014 model with low BIC, designated therein as BT11jarpa17). The models are modified to delete any covariates that are not publically available via the standard reporting for catch data. Consequently the terms related to diatoms have been omitted from the model matrix. The simulated total sample size is varied by randomly resampling entire rows from the data matrix.

Model d3 is used to generate the "complex" data. Model J17 is used to generate the "simple" data. In order to generalise the simulations, the coefficients are multiplicatively "dithered" in each trial by a random number from a uniform distribution with the range 0.6 to 1.4. The central values for the coefficients are similar to those estimated from fitting the models.

The complex model (d3) is specified by:

***BT11 ~ DateNum*Sex + DateNum*Year:LonSect:LongNum + DateNum:LonSect:Ice + YearNum:LonSect + BLm***

The terms of this model are:

**DateNum*Sex** allows the growth in blubber to be different for each sex for both the slope and the intercept.

***DateNum*Year:LonSect:LongNum*** allows the growth blubber to be different in each year and longitudinal sector, and within that sector to vary linearly with longitude

***DateNum:LonSect:Ice*** allows the growth in blubber in each sector to be different in the strata *near* the ice from that *far* from the ice (see de la Mare *et al* 2014).

***YearNum:LonSect*** allows the year trend to be different for each longitudinal sector

***BLm*** provides a linear correction for blubber thickness due to body size (or related size term in some analyses).

This model allows whales in each sector and year to accumulate blubber thickness at different rates with blubber growth rates also able to differ by sex and between *near* and *far* ice strata. As discussed in Wotherspoon *et al.*, 2014 this is the type of model that is biologically reasonable.

The simple model (J17) is given by:

***BT11 ~ YearNum + BLm^3 + DateNum + LonSect + Sex + DateNum:LongNum***

The only interaction term in this model is between the continuous variables **DateNum** and **LongNum**, which thus adds only one coefficient to the model and assumes that there is a linear relationship between the growth rate in blubber thickness and longitude.

Three scenarios are explored:

- A complex true model is fitted to complex data (true model is reachable [nested])
- simple true model is fitted to simple data (true model is reachable [nested])
- complex model is fitted to simple data (true model not nested in fitted model because the "true interaction" term is not represented in the complex model)

The degree to which models are under or over-fitted in each of 100 simulations is based on the number of parameters in the selected model compared with the number in the "true" model (Table 4). In trying the recover complex models from complex data (Scenario 1), results show that AIC does remarkably well, achieving close to the true number of parameters in all simulations. In contrast, BIC on average only recovered about 64% of the true number of parameters, in only 20/100 simulations recovered 90% or more of the true number of parameters, and in no instance recovered the full parameter set. Differences between AIC and BIC were similar in Scenario 2; AIC was better at recovering true model structure in a low-dimensional setting, but BIC did perform substantially better than in scenario 1. Scenario 3 was used to test fitting a high dimensional model to data generated from a simple model, and BIC was better in this instance. AIC overfitted the data by selecting more than twice as many

predictors as was required. Surprisingly, BIC also overfitted by about 50%. As expected, model selection becomes more reliable with increasing sample sizes and AIC is less prone to underfitting at the lower sample sizes.

The important issue revealed by these trials is, in a given realised experimental "design", what sample size is necessary before it is reasonable to assume that the asymptotic consistency property of BIC model selection is likely to be effective. In the context of the JARPA nutritive condition the available sample size appears to be too small. The simulations suggest that in this case a total sample size for a fixed effects linear model needs to be of the order of 10 000 or more to be reasonable confident that BIC is not underfitting the data. With some further development, the types of simulations in this paper could provide an appropriate practical approach for establishing a reasonable lower bound for sample size when undertaking BIC based model selection for a given data set. For the smaller sample sizes, the use of AIC model selection is less likely to lead to underfitting. A simulation approach may be even more important in mixed effects models where the effective sample sizes are usually much less than the nominal sample sizes.

**Significance testing after model selection**

One issue that the panel did not touch on in its general advice was the validity of significance tests carried out after selecting a model using IC criteria. Statistical inference after model selection generally results in confidence intervals that are too narrow because it usually neglects the probability associated with whether a given variable is selected through the model-selection procedure (Miller 2002, Burnham and Anderson 1998, Berk *et al*. 2009). Moreover, the distributions of test statistics can depart very substantially from the form assumed used in calculating significance levels (Leeb and Pötscher, 2006, Berk *et al*. 2009). Berk *et al*. summarise this neatly:

> *Conventional statistical inference requires that a model of how the data were generated be known before the data are analyzed. Yet in criminology, and in the social sciences more broadly, a variety of model selection procedures are routinely undertaken followed by statistical tests and confidence intervals computed for a ''final'' model.* [Beck *et al*.] *... examine such practices and show how they are typically misguided. The parameters being estimated are no longer well defined, and post-model-selection sampling distributions are mixtures with properties that are very different from what is conventionally assumed. Confidence intervals and statistical tests do not perform as they should.*

Of course, these observations are not restricted to the social sciences but apply to any approach based on model selection in complex circumstances whether based on IC, hypothesis testing or even ad-hoc procedures of model selection.

**Conclusions and recommendations**

It should be clear from this and past work that we do not accept that BIC should be used as the sole means of model selection when fitting statistical models in complex ecological systems. We demonstrate that our position is well supported in the ecological and statistical literature. We have shown that, in the specific analysis case of interest to the SC (minke whale biological parameters), BIC does not perform well compared with AIC in terms of recovering complex model structure from complex data (Scenario 1). Although BIC performed somewhat better than AIC when trying to fit a complex model to simple data (Scenario 3), we believe that Scenario 3 is highly unlikely in the context of large-scale ecological systems.

We recognise that the science of statistical model selection sometimes allows, even requires, a degree of artistic license, but we nonetheless feel there are several points that might usefully be observed when analysing data and reporting results.

1. *Standard information*
   Presentation of model selection results based on IC should include the log-likelihood, the number of free parameters, and (minimally) both the AIC and BIC values. Other IC variants should be reported if warranted by the analysis problem. The common sample size for all comparisons should also be reported.
2. *Similar models and model averaging*
   Sometimes a single model will not be identified as a clear 'winner' based on IC scores (irrespective of the IC used). There may be several models with similar IC scores, potentially formed from quite different covariate sets, and all may imperfectly approximate some underlying process in different ways (Berk *et al*. 2010). Approximately equivalent models (in terms of difference in IC from the best model amongst the candidate set) should be explored to determine if they give appreciably different results (i.e. in terms of predictive ability or significance of parameter estimates, depending on the goal). Advice on "approximately equivalent" varies, with overviews of provided by Murtaugh (2014) and Neath and Cavanaugh (2012). Hoeting *et al*. (1999) and Burnham *et al*. (2011) give practical advice about model averaging.
3. *Concordance between different types of IC*
   In the event that model selection by AIC and BIC (or other IC approaches) select appreciably different sets of predictor variables, and assuming obvious model deficiencies are not evident by standard diagnostic

procedures, then this should be interpreted as indicating uncertainty in deciding useful candidate models. Results from competing models should be sensitivity tested to determine how choice of IC affects the analysis goal (e.g. predictive accuracy, significance of model terms).

4. *Evaluate the sample size*
   Use simulation or other methods to explore whether sample sizes are large enough to be confident that model selection will have appropriate properties in the context of the specific analysis.
5. *To be conservative*
   For the purposes of conducting power analyses to determine prospective sample sizes (e.g. lethal-take of minke whales for research), we believe there is a strong argument for adopting more complex, rather than less complex, models. Our reasoning is that we have a duty to err on the side of caution; we must assume that the system under study is at least as complex as our largest models indicate, and sample sizes should be large enough to ensure adequate data to address the questions we pose in that case. If the system turns out to be less complex than anticipated, then we are no worse off in terms of our ability to address our research goals. The reverse would not hold: if we start by assuming a less complex system than in fact exists, then the calculated sample sizes and resulting data may be insufficient for the research purposes.
6. *Design variables*
   Correct statistical inference sometimes requires models that incorporate fixed and random terms to accommodate incomplete control in the sampling design. These terms may themselves be crossed or nested in respect to other covariates. Deficiencies in data collection are taken to include lack of balance in spatial and temporal coverage of sampling (possibly leading to confounding) and non-random selection of observational units. We propose that it makes little sense to remove design variables via model selection procedures when their express purpose is to adjust data so that valid inferences can be drawn. Such variables should be quarantined from removal. Reasonable inference in complex mixed models can be challenging; Bolker *et al.* (2009) and Claeskens and Hjort (2008) offers some general guidance.
7. *Respect the possibility of interactions*
   We recommend the use of IC that allow the possibility — rather than impossibility — of interactions. It make little sense to us to use an IC, like BIC, that a priori imposes such stringent penalties against interactions between variables. Our calculations show that BIC, even at moderate sample sizes, requires interactions to be extremely significant by conventional levels before they would be considered for inclusion in models. For large sample sizes, BIC becomes even more conservative in selecting interactions. This restriction seems incongruent with decades of ecological research demonstrating the prevalence of complex interactions in biological systems.
8. *Do not over-interpret post model selection statistical inference*
   Statistical inference after model selection generally results in confidence intervals that are too narrow because it usually neglects the probability associated with whether a given variable is selected through the model-selection procedure. In general, full Bayesian methods that take into account model uncertainty are likely to be more reliable.

There are many topics important to the subject of model selection that this paper has not touched on. In particular, we single out cross-validation and methods employing shrinkage estimators as strong rivals to the IC approach to model selection (Hastie *et al.* 2009). Others aspects concerning model selection, such as model averaging and alternate IC, have been touched on only superficially. While important, and potentially highly relevant, these additional approaches are secondary to what we hope is the main message of this paper: that no model selection procedure is perfect or even superior. Each has strengths and weaknesses that will differentially affect estimation and inference dependent on the data and system under study, and that unthinking adherence to a single method is likely to provide poor results in some cases. We believe use of BIC to select models for estimating trends in minke whale body condition provides just such a case.

**References**


Aho, K., Derryberry, D. and Peterson, T. 2014 Model selection for ecologists: the worldview of AIC and BIC. Ecology. 95(3): 631-636.

Akaike, H. 1973 Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Caski (Eds.), Proceeding of the second international symposium on information theory. Budapest: Akademiai Kiado.

Arnold, T.W. 2010 Uninformative Parameters and Model Selection Using Akaike's Information Criterion. Journal of Wildlife Management 74(6): 1175-1178.

Barker, J.J. and Ogle, K. 2014 To *P* or not to *P*? Ecology 95(3):621-626.

Berk, R., Brown, L. and Zhao, L. 2010 Statistical inference after model selection. Journal of Quantitative Criminology 26(2):217–236.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H. and White, J-S.S. 2009 Generalized Linear Mixed Models: A Practical Guide for Ecology and Evolution. Trends in Ecology & Evolution 24(3): 127–135.

Burnham, K.P. and Anderson, D.R. 1998 Model selection and inference: a practical information-theoretic approach. Springer Verlag, New York. 353pp

Burnham, K.P. and Anderson, D.R. 2002 Model selection and multimodel inference: a practical information-theoretic approach. Springer, New York. 488p.

Burnham, K.P., Anderson, D.R. and Huyvaert, K.P. 2011 AIC model selection and multimodel inference in behavioural ecology: some background, observations, and comparisons. Behavioural Ecology and Sociobiology 65(1):23-35

Casella, G., Girón, F.J., Martínez, M.L. and Moreno, E. 2009 Consistency of Bayesian procedures for variable selection. The Annals of Statistics 37(3):1207–1228.
Claeskens, G. and Hjort, N.L. 2008 Model Selection and Model Averaging. Leiden: Cambridge University Press.
de La Mare, W., Candy, S., McKinlay, J., Wotherspoon, S. and Double, M. 2014 What can be concluded from the statistical analyses of JARPA/JARPA II body condition data? SC/F14/O06.
Flynn, C., Hurvich, C.M. and Simonoff, J.S. 2011 Efficiency and Consistency for Regularization Parameter Selection in Penalized Regression: Asymptotics and Finite-Sample Corrections. NYU Working Paper, 2451/31317.
Freckleton, R.P. 2011 Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. Behavioral Ecology and Sociobiology, 65(1):91–101.
Gelman, A. 2009 Bayes, Jeffreys, Prior Distributions and the Philosophy of Statistics. Statistical Science 24:176–178
Gelman, A.E. and Rubin, D.B. 1995 Avoiding model selection in Bayesian social research. Sociological Methodology 25:165–173
Gelman, A., Hwang, J. and Vehtari, A. 2013 Understanding Predictive Information Criteria for Bayesian Models. Statistics and Computing 24(6): 997–1016.
Hastie, T., Tibshirani, R. and Friedman, J. 2009 The Elements of Statistical Learning. 2nd edition. Springer Series in Statistics, New York.
Hooten, M.B. and Hobbs, N.T. 2015 A Guide to Bayesian Model Selection for Ecologists. Ecological Monographs 85:3-28.
Hoeting, J. A., Madigan, D., Raftery, A.E. and Volinsky, C.T. 1999 Bayesian Model Averaging: A Tutorial. Statistical Science: 382–401.
Hurvich, C.M., and Tsai, C-L. 1989 Regression and time series model selection in small samples. Biometrika 76:297–307.
Kass, R.E. and Raftery, A.E. 1995 Bayes Factors. Journal of the American Statistical Association 90(430): 773–795.
Konishi, S.and Kitagawa, G. 2008 Information Criteria and Statistical Modeling. Springer, New York. 273p.
Konishi, K. and Walloe, L. 2014 Time trends in the energy storage in the Antarctic minke whales during the JARPA and JARPA II research periods. SC/F14/J13.
Konishi, K., Hakamada, T., Kiwada, H., Kitakado, T. and Walloe, L. 2014 Decrease in stomach contents in the Antarctic minke whale (Balaenoptera bonaerensis) in the Southern Ocean. SC/F14/J14.
Lavine, M. 2014 Comment on "Murtaugh.2014 In defence of P-values". Ecology 95(3):642-645.
Leeb, H. and Pötscher, B.M. 2006 Can one estimate the conditional distribution of post-model-selection estimators? Annals of Statistics 34(5):2554–2591
Miller A.J. 2002 Subset selection in regression (2nd edition). Monographs in Statistics and Applied Probability 95:1:237. Chapman and Hall/CRC, Boca Raton, Florida
Moreno, E., Girón, F.J. and Casella, G. 2010 Consistency of objective Bayes factors as the model dimension grows. Annals of Statistics 38(4):1937–1952.
Murtaugh, P.A. 2014 In defence of P values. Ecology 95(3): 611–617.
Neath, A.A. and Cavanaugh, J. E. 2012 The Bayesian information criterion: background, derivation, and applications. Computational Statistics 4(2):199–203.
R Core Team 2015 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Robert, C. P., Chopin, N. and Rousseau, J. 2009 Harold Jeffreys's Theory of Probability Revisited. Statistical Science 24:141–172
Schwarz, G. 1978 Estimating the dimension of a model. Annals of Statistics, 6, 461–464.
Shao, J. 1997An Asymptotic Theory for Linear Model Selection. Statistica Sinica 7(2): 221–242.
Stone, M. 1977 An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. Roy. Statist. Soc. Ser. B 39(1): 44-47.
Sugiura, N. 1978 Further analysis of the data by Akaike's information criterion and the finite corrections. Communications in Statistics, Theory and Methods A7:13–26.
Symonds, M.R. E. and Moussalli, A. 2011 A Brief Guide to Model Selection, Multimodel Inference and Model Averaging in Behavioural Ecology Using Akaike's Information Criterion. Behavioral Ecology and Sociobiology 65(1): 13–21.
Venables, W.N. and Ripley, B.D. 2002 Modern applied statistics with S, 4th Edition. Springer. 495p.
Wang, M. and Maruyama, Y. 2015 Consistency of Bayes factor for nonnested model selection when the model dimension grows. Bernoulli, *submitted*.
Ward, E.J. 2008 A Review and Comparison of Four Commonly Used Bayesian and Maximum Likelihood Model Selection Tools. Ecological Modelling 211(1-2): 1–10.
Wotherspoon, S., Double, M.C., McKinlay, J., Candy, S., Andrews-Goff, V. and de la Mare, W.K. 2014. JARPA and JARPA II cannot monitor trends in the Antarctic ecosystem due to flawed sampling strategies. Paper presented to IWC review of JARPAII SC/F14/O05 : 3pp.
Yang, Y. 2005 Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. Biometrika 92:937–950

Table 4. Recovery of the true number of generating model parameters for each of three simulation scenarios using both AIC and BIC model selection on the same simulated data. The sample size n = 4727 matches the available sample size from JARPA.

| Measure | True DF | n = 1000 | | n = 4727 | | n = 10000 | |
|---|---|---|---|---|---|---|---|
| | | AIC | BIC | AIC | BIC | AIC | BIC |
| Fitting complex model to complex data - d3 fitted to d3 | | | | | | | |
| Mean number parameters selected | **109** | 91.1 | 54.3 | 105.6 | 69.5 | 107.0 | 98.1 |
| Number of models correctly selected | | 6 | 0 | 42 | 0 | 66 | 0 |
| Number of models with at least 90% of true parameters | | 71 | 0 | 100 | 20 | 100 | 87 |
| Fitting simple model to simple data - j17 fitted to j17 | | | | | | | |
| Mean number parameters selected | **11** | 8.1 | 4.2 | 10.0 | 8.3 | 10 | 9.5 |
| Number of models correctly selected | | 44 | 1 | 98 | 49 | 100 | 88 |
| Number of models with at least 90% of true parameters | | 67 | 1 | 100 | 68 | 100 | 90 |
| Fitting complex model to simple data - d3 fitted to j17 | | | | | | | |
| Mean number parameters selected | **11** | 16.0 | 6.5 | 22.0 | 12.2 | 25.6 | 17.2 |
| Number of models correctly selected | | n/a | n/a | n/a | n/a | n/a | n/a |
| Number of models with at least 90% of true parameters | | 89 | 15 | 100 | 86 | 100 | 99 |