CSIRO Publishing

Wildlife Research

Volume 28, 2001

All enquiries and manuscripts should be directed to:

Wildlife Research
CSIRO Publishing
PO Box 1139 (150 Oxford St)
Collingwood, Vic. 3066, Australia

Telephone: +61 3 9662 7622
Fax: +61 3 9662 7611
Email: wr@publish.csiro.au

Published by CSIRO Publishing
for CSIRO and the Australian Academy of Science

www.publish.csiro.au/journals/wr

Wildlife Research, 2001, **28**, 111–119

Kullback–Leibler information as a basis for strong inference in ecological studies

Kenneth P. Burnham[A] *and David R. Anderson*

Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, Fort Collins, CO 80523, USA.
[Employed by USGS, Division of Biological Resources.]
[A]Email: kenb@cnr.colostate.edu

Abstract. We describe an information-theoretic paradigm for analysis of ecological data, based on Kullback–Leibler information, that is an extension of likelihood theory and avoids the pitfalls of null hypothesis testing. Information-theoretic approaches emphasise a deliberate focus on the *a priori* science in developing a set of multiple working hypotheses or models. Simple methods then allow these hypotheses (models) to be ranked from best to worst and scaled to reflect a strength of evidence using the likelihood of each model (g_i), given the data and the models in the set (i.e. $L(g_i \mid data)$). In addition, a variance component due to model-selection uncertainty is included in estimates of precision. There are many cases where formal inference can be based on all the models in the *a priori* set and this multi-model inference represents a powerful, new approach to valid inference. Finally, we strongly recommend inferences based on *a priori* considerations be carefully separated from those resulting from some form of data dredging. An example is given for questions related to age- and sex-dependent rates of tag loss in elephant seals (*Mirounga leonina*).

Introduction

Theoretical and applied ecologists are becoming increasingly dissatisfied with the traditional testing-based aspects of frequentist statistics. Over the past 50 years a large body of statistical literature has shown the testing of null hypotheses to have relatively little utility, in spite of their very widespread use (Nester 1996). Inman (1994) provides a historical perspective on this issue by highlighting the points of a heated exchange in the published literature between R. A Fisher and Karl Pearson in 1935. In the applied ecology literature Yoccoz (1991), Cherry (1998), Johnson (1999) and Anderson *et al.* (2000) have written on this specific issue. The statistical null hypothesis testing approach is not wrong, but it is relatively uninformative and, thus, slows scientific progress and understanding.

Bayesian approaches are relatively unknown to ecologists and will likely remain so because this material is not commonly offered in statistics departments, except perhaps in advanced courses. Still, an increasing number of people think that Bayesian statistics offer an acceptable alternative (Gelman *et al.* 1995; Ellison 1996), while others are leery (Forster 1995; Dennis 1996). In addition, there are fundamental issues with the subjectivity inherent in many Bayesian methods and this has unfortunately divided the field of statistics for many decades. Also, much of Bayesian statistics has been developed from the viewpoint of decision theory. We find that science is most involved with estimation, prediction and understanding and, less so, with decision-making (see Berger 1985 for a discussion of decision-making).

The purpose of this paper is to introduce readers to the use of Kullback–Leibler information as a basis for making valid inference from the analysis of empirical data. We provide this introduction because information-theoretic approaches are simple, easy to learn and understand, compelling, and quite general. This class of methods allows one to select the best model from an *a priori* set, rank and scale the models, and include model selection uncertainty into estimates of precision. Information-theoretic approaches provide an effective strategy for objective data analysis (Burnham and Anderson 1998; Anderson and Burnham 1999). Finally, we provide a simple approach to making formal inference from more than a single model (multi-model inference, or MMI). We believe the information-theoretic approaches are excellent for the analysis of ecological data, whether experimental or observational, and provide a rational alternative to the testing-based frequentist methods and the computer-intensive Bayesian methods.

The central inferential issues in science are two-fold. First, scientists are fundamentally interested in estimates of the magnitude of the parameters or differences between parameters and their precision; are the differences trivial, small, medium, or large? Are the differences biologically meaningful? This is an estimation problem. Second, one often wants to know whether the differences are large enough to justify inclusion in a model to be used for further inference (e.g. prediction) and this is a model-selection problem. These central issues are not properly associated with statistical null hypothesis-testing. In particular, hypothesis-testing is a poor

 10.1071/WR99107 1035-3712/01/020111

approach to model selection or variable selection (e.g. forward or backward selection in regression analysis).

The application of information-theoretic approaches is relatively new; however, a number of papers using these methods have already appeared in the fisheries, wildlife and conservation biology literature. Research into the analysis of marked birds has made heavy use of these new methods (see the special supplement of *Bird Study*, 1999, Vol. 46). Program MARK (White *et al.* 2001) allows a full analysis of data under the information-theoretic paradigm, including model averaging and estimates of precision that include model-selection uncertainty. Distance sampling and analysis theory (Buckland *et al.* 1993) should often be based on this theory with an emphasis on making formal inference from several models. The large data sets on the threatened northern spotted owl in the United States have been the subject of large-scale analyses using these new methods (see Burnham *et al.* 1996). Burnham and Anderson (1998) provide a number of other examples, including formal experiments to examine the effect of a treatment, studies of spatial overlap in *Anolis* lizards in Jamaica, line-transect sampling of kangaroos at Wallaby Creek in Australia, predicting the frequency of storms in South Africa, and the time distribution of an insecticide (Dursban®) in a simulated ecosystem. Burnham and Anderson (1998: 96–99) provide an example of a simulated experiment on starlings (*Sturnus vulgaris*) to illustrate that substantial differences can arise between the results of hypothesis-testing and model-selection criteria. Another example relates to time-dependent survival of sage grouse (*Centrocercus urophasianus*) where Akaike's Information Criterion selected a model with 4 parameters whereas hypothesis tests suggested a model with 58 parameters (Burnham and Anderson 1998: 106–109). Information-theoretic methods have found heavy use in other fields of science (e.g. time series analysis).

Science Philosophy

First we must agree on the fact that there are no true models; instead, models, by definition, are only approximations to unknown reality or truth. George Box made the famous statement 'All models are wrong but some are useful'. In the analysis of empirical data, one must face the question 'What model should be used to best approximate reality given the data at hand?' (the best model depends on sample size). The information-theoretic paradigm rests on the assumption that good data, relevant to the issue, are available and these have been collected in an appropriate manner. Three general principles guide us in model-based inference in the sciences.

Simplicity and Parsimony. Many scientific concepts and theories are simple, once understood. In fact, Occam's Razor implores us to 'shave away all but what is necessary'. Albert Einstein is supposed to have said 'Everything should be made as simple as possible, but no simpler'. Parsimony enjoys a featured place in scientific thinking in general and in modelling specifically (see Forster and Sober 1994; Forster 2001) for a strictly science philosophy perspective).

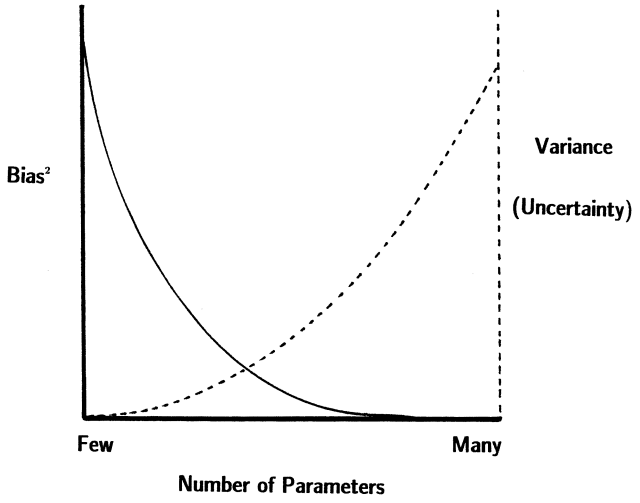

Fig. 1. The principle of parsimony: the conceptual trade-off between squared bias (solid line) and variance (i.e. uncertainty) versus the number of estimable parameters in the model. The best model has dimension (K_0) near the intersection of the two lines, while full reality lies far to the right of trade-off region.

Model selection (variable selection in regression is a special case) is a bias *v.* variance trade-off and this is the principle of parsimony (Fig. 1). Models with too few parameters (variables) have bias, whereas models with too many parameters (variables) may have poor precision or tend to identify effects that are, in fact, spurious (slightly different issues arise for count data *v.* continuous data). These considerations call for a balance between under- and over-fitted models – the so-called 'model selection problem' (see Forster 2000).

Multiple Working Hypotheses. Over 100 years ago, Chamberlin (1890, reprinted 1965) advocated the concept of 'multiple working hypotheses'. Here, there is no null hypothesis; instead, there are several well-supported hypotheses (equivalently, 'models') that are being entertained. The *a priori* 'science' of the issue enters at this important point. Relevant empirical data are then gathered, analysed, and the results tend to support one or more hypotheses, while providing less support for other hypotheses. Repetition of this general approach leads to advances in the sciences. New or more elaborate hypotheses are added, while hypotheses with little empirical support are gradually dropped from consideration. At any one point in time, there are multiple hypotheses (models) still under consideration. An important feature of this multiplicity is that the number of alternative models should be kept small (Zucchini 2000); the analysis of, say, hundreds of models is not justified except when prediction is the only objective, or in the most exploratory phases of an investigation.

Strength of Evidence. Providing information to judge the 'strength of evidence' is central to science. Null-hypothesis-testing provides only arbitrary dichotomies (e.g. significant *v.* non-significant) and in the all-too-often-seen case where the null hypothesis is obviously false on *a priori* grounds, the test result is superfluous. Royall (1997) provides an interesting discussion of the likelihood-based strength-of-evidence approach in simple statistical situations.

The information-theoretic paradigm is partially grounded in the three principles above. Impetus for the general approach can be traced to several major advances made over the past half century and this history will serve as an introduction to the subject.

Advance 1 – Kullback–Leibler information

In 1951 S. Kullback and R. A. Leibler published a now-famous paper that examined the scientific meaning of 'information' related to R. A. Fisher's concept of a 'sufficient statistic'. Their celebrated result, now called *Kullback–Leibler information*, is a fundamental quantity in the sciences and has earlier roots back to Boltzmann's (1877) concept of *entropy.* Boltzmann's entropy and the associated Second Law of Thermodynamics represents one of the most outstanding achievements of 19th century science.

Kullback–Leibler (K–L) information is a measure (a 'distance' in an heuristic sense) between conceptual reality, *f,* and approximating model, *g*, and is defined for continuous functions as the integral

$$I(f,\ g) = \int f(x) \log_e \left(\frac{f(x)}{g(x \mid \theta)} \right) dx$$

where *f* and *g* are *n*-dimensional probability distributions. K–L information, denoted $I(f, g)$, is the 'information' lost when model *g* is used to approximate reality, *f*. The analyst seeks an approximating model that loses as little information as possible; this is equivalent to minimising $I(f, g)$, over the set of models of interest (we assume there are *R a priori* models in the candidate set).

Boltzmann's entropy *H* is $-I(f, g)$, although these quantities were derived along very different lines. Boltzmann derived the fundamental relationship between entropy (*H*) and probability (*P*) as

$$H = \log_e(P)$$

and because $H = -I(f, g)$, one can see that entropy, information and probability are linked, allowing probabilities to be multiplicative whereas information and entropy are additive.

K–L information can be viewed as an extension of the famous Shannon (1948) entropy and is often referred to as 'cross entropy'. In addition, there is a close relationship between Jaynes' (1957) 'maximum entropy principle' or MaxEnt (see Akaike 1977, 1983*a*, 1985). Cover and Thomas (1989) provide a nice introduction to information theory in general. K–L information, by itself, will not aid in data analysis as both reality (*f*) and the parameters (θ) in the approximating model are unknown to us. H. Akaike made the next breakthrough in the early 1970s.

Advance 2 – Estimation of Kullback–Leibler information (AIC)

Akaike (1973, 1974) found a formal relationship between K–L information (a dominant paradigm in information and coding theory) and maximum likelihood (the dominant paradigm in statistics) (see deLeeuw 1992). This finding makes it possible to combine estimation (e.g. maximum likelihood or least squares) and model selection under a single theoretical framework – optimisation. Akaike's breakthrough was the finding of an estimator of the expected, relative K–L information, based on the maximised log-likelihood function. Akaike's derivation (which is for large samples) relied on K–L information as averaged entropy and this lead to 'Akaike's information criterion' (AIC),

$$\mathrm{AIC} = -2\log_e(L(\hat{\theta} \mid data)) + 2K,$$

where $\log_e(L(\hat{\theta} \mid data))$ is the value of the maximised log-likelihood over the unknown parameters (θ), given the data and the model, and *K* is the number of estimable parameters in that approximating model. In the special case of least-squares (LS) estimation with normally distributed errors for all *R* models in the set, and apart from an arbitrary additive constant, AIC can be expressed as

$$\mathrm{AIC} = n\log(\hat{\sigma}) + 2K,$$

where

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n}$$

and $\hat{\epsilon}_i$ are the estimated residuals from the fitted model. In this case the number of estimable parameters, *K*, must be the total number of parameters in the model, including the intercept and σ^2. Thus, AIC is easy to compute from the results of LS estimation in the case of linear models or from the results of a likelihood-based analysis in general (Edwards 1992; Azzalini 1996). Akaike's procedures are now called information-theoretic because they are based on the K–L information (see Akaike 1983*b*, 1992, 1994).

Assuming that a set of *a priori* candidate models has been defined and is well supported by the underlying science, then AIC is computed for each of the approximating models in the set (i.e. g_i, $i = 1, 2, \ldots, R$). The model for which AIC is minimal is selected as best for the empirical data at hand. This is a simple, compelling concept, based on deep theoretical foundations (i.e. entropy, K–L information, and likelihood theory). AIC is not a test in any sense: no single hypothesis

(model) is made to be the 'null', there is no arbitrary α level, and there is no arbitrary notion of 'significance'. Instead, there are concepts of evidence and a 'best' inference, given the data and the set of *a priori* models representing the scientific hypotheses of interest.

When K is large relative to sample size n (which includes when n is small, for any K) there is a small-sample (second-order) version called AIC_c,

$$AIC_c = -2\log_e(L(\hat{\theta})) + 2K + \frac{2K(K+1)}{(n-K-1)}$$

(see, for example, Hurvich and Tsai 1989), and this should be used unless $n/K > \sim 40$. Both AIC and AIC_c are estimates of expected, relative Kullback–Leibler information and are useful in the analysis of real data in the 'noisy' sciences. Assuming independence, AIC-based model selection is equivalent to certain cross-validation methods (Stone 1974, 1977) and this is an important property.

Akaike's general approach allows the best model in the set to be identified, but also allows the rest of the models to be easily ranked. Here, it is very useful (essentially imperative) to rescale AIC (or AIC_c) values such that the model with the minimum information criterion has a value of 0, i.e.

$$\Delta_i = AIC_i - \min AIC.$$

The Δ_i values are easy to interpret, and allow a quick 'strength of evidence' comparison and ranking of candidate hypotheses or models. The larger the Δ_i, the less plausible is fitted model i as being the best approximating model in the candidate set. It is generally important to know which model (hypothesis) is second best (the ranking) as well as some measure of its standing with respect to the best model. Some simple rules of thumb are often useful in assessing the relative merits of models in the set: models having $\Delta_i \leq 2$ have substantial support (evidence), those where $4 \leq \Delta_i \leq 7$ have considerably less support, while models having $\Delta_i > 10$ have essentially no support. An improved method for scaling models appears in the next section.

The Δ_i values allow an easy ranking of hypotheses (models) in the candidate set. One must turn to goodness-of-fit tests or other measures to determine whether any of the models is good in some absolute sense. For count data, we suggest a standard goodness-of-fit test; whereas standard measures such as R^2 and $\hat{\sigma}^2$ in regression and analysis of variance are often useful. Justification of the models in the candidate set is a very important issue. This is where the science of the problem enters the scene. Ideally, there ought to be a justification of models in the set and a defense as to why some models should remain out of the set. This is an area where ecologists need to spend much more time just thinking, well prior to data analysis and, perhaps, prior to data collection.

The principle of parsimony, or Occam's razor, provides a philosophical basis for model selection; Kullback–Leibler information provides an objective target based on deep, fundamental theory; and the information criteria (AIC and AIC_c), along with likelihood- or least-squares-based inference, provide a practical, general methodology for use in the analysis of empirical data. Objective data analysis can be rigorously based on these principles without having to assume that the 'true model' is contained in the set of candidate models – surely there are no true models in the biological sciences!

Advance 3 – Likelihood of a model, given the data

The simple transformation $\exp(-\Delta_i/2)$, for $i = 1, 2, \ldots, R$, provides the likelihood of the model (Akaike 1981) given the data: $L(g_i \mid data)$. This is a likelihood function over the model set in the same sense that $L(\theta \mid data, g_i)$ is the likelihood over the parameter space (for model g_i) of the parameters θ, given the data (x) and the model (g_i). The relative likelihood of model i versus model j is $L(g_i \mid data)/L(g_j \mid data)$; this ratio does not depend on any of the other models under consideration. Without loss of generality we may assume model g_i is more likely than g_j. Then if this ratio is large (e.g. >10 is large), model g_j is a poor model to fit the data *relative* to model g_i. The expression $L(g_i \mid data)/L(g_j \mid data)$ can be regarded as an *evidence ratio* – the evidence for model i versus model j.

It is often convenient to normalise these likelihoods such that they sum to 1, hence we use

$$w_i = \frac{\exp(-\Delta_i/2)}{\sum_{r=1}^{R}\exp(-\Delta_r/2)}$$

The w_i, called *Akaike weights*, are useful as the 'weight of evidence' in favor of model i as being the actual K–L best model in the set. The ratios w_i/w_j are identical to the original likelihood ratios, $L(g_i \mid data)/L(g_j \mid data)$; however, w_i, $i = 1, \ldots, R$ are useful in additional ways. For example, the w_i are interpreted approximately as the probability that model i is, in fact, the K–L best model for the data. This latter inference about model-selection uncertainty is conditional on both the data and the full set of *a priori* models considered. There are simple methods to provide a confidence set on the models, in the same sense as a confidence set for estimates of parameters, and to allow prior (Bayesian type) information to affect these weights (see Burnham and Anderson 1998: 126–128).

Advance 4 – Unconditional sampling variance

Typically, estimates of sampling variance are conditional on a 'given' model as if there were no uncertainty about which model to use (Breiman 1992 calls this a 'quiet scandal'). When model selection has been done, there is a variance component due to model-selection uncertainty that should be

incorporated into estimates of precision. That is, one needs estimates that are 'unconditional' on the selected model. Here the estimates are unconditional on any particular model, but conditional on the *R* models in the *a priori* set. A simple estimator of the unconditional variance for the parameter maximum likelihood estimator, $\hat{\theta}$, from the selected (best) model is,

$$\widehat{\text{var}}(\hat{\theta}) = \left[\sum_{i=1}^{R} w_i \sqrt{\widehat{\text{var}}(\hat{\theta}_i \mid g_i) + (\hat{\theta}_i - \bar{\hat{\theta}})^2} \right]^2$$

$$\text{where } \bar{\hat{\theta}} = \sum_{i=1}^{R} w_i \hat{\theta}_i$$

and this represents a form of frequentist 'model averaging'. The notation $\hat{\theta}_i$ here means that the parameter θ is estimated on the basis of model g_i, but θ is a parameter in common to all *R* models (such as occurs with prediction).

This estimator, from Buckland *et al.* (1997), includes a term for the conditional sampling variance, given model g_i (denoted as $\widehat{\text{var}}(\hat{\theta}_i|g_i)$ here) and a variance component for model-selection uncertainty, $(\hat{\theta}_i - \bar{\hat{\theta}})^2$. These variance components are multiplied by the Akaike weights, which reflect the degree of support or evidence for model *i*. The unconditional variance and its square root are appropriate measures of precision after model selection. The usual 95% confidence interval, $\hat{\theta} \pm 2\widehat{\text{se}}(\hat{\theta})$ should be based on the unconditional variance. Alternatively, intervals can be based on log- or logit-transformations (Burnham *et al.* 1987), profile likelihoods (Royall 1997) or bootstrap methods (Efron and Tibshirani 1993). Burnham and Anderson (1998, chapter 5) provide a number of Monte Carlo results on achieved confidence interval coverage when information-theoretic approaches are used in some moderately challenging data sets. Model averaging (see below) arises naturally when the unconditional variance is derived (Burnham and Anderson 1998: section 4.2.6).

Advance 5 – Multi-model inference (MMI)

Rather than base inferences on a single, selected best model from an *a priori* set of models, inference can be based on the entire set of models (multi-model inference, or MMI). Such inferences can be made if a parameter, say θ, is in common over all models (as θ_i in model g_i), or if the goal is prediction. Then, by using the weighted average for that parameter across models (i.e. $\bar{\hat{\theta}} = \Sigma w_i \hat{\theta}_i$) we are basing point inference on the entire set of models. This approach has both practical and philosophical advantages (see Hoeting *et al.* 1999 for a discussion of model averaging in a Bayesian context). Where a model-averaged estimator can be used it often has better precision and reduced bias compared with the estimator of that parameter from just the selected best model (Burnham and Anderson 1998: chapters 4 and 5).

Assessment of the relative importance of variables has often been based only on the best model (e.g. often selected using a stepwise testing procedure). Variables in that best model are considered 'important', excluded variables are considered not important. This is far too simplistic. Variable importance can be refined by making inference from all the models in the candidate set (see Burnham and Anderson 1998: 140–151). Akaike weights are summed for all models containing a given predictor variable x_j; we denote the resultant sum as $w_+(j)$. For each variable considered we can compute its predictor weight. The predictor variable with the largest predictor weight, $w_+(j)$, is estimated to be the most important; the variable with the smallest sum is estimated to be the least important predictor (as with all inferences there is uncertainty about the inferred order of variable importance). This procedure is superior to making inferences concerning the relative importance of variables based only on the best model. This is particularly important when the second or third best model is nearly as well supported as the best model, or when all models have nearly equal support. (There are 'design' considerations about the set of models to consider when a goal is assessing the importance of predictor variables, we do not discuss these considerations here – the key issue is one of balance of models with and without each variable.)

Advance 6 – Incorporation of overdispersion for count data

Much of the statistical analysis in wildlife and ecology deals with count data (e.g. capture–recapture) and overdispersion is a fact of life with such count data. When there is more variation than predicted by Poisson or multinomial probability distributions, the data are termed overdispered (Agresti 1990: 42). A partial dependence in the count data most often underlies the overdispersion; however, parameter heterogeneity is another contributor to overdispersion. Kullback–Leibler-based model-selection and inference methods have been adapted to deal with overdispersion based on ideas from quasi-likelihood methods and variance inflation (Wedderburn 1974). The usual models of count data implicitly assume a theoretical sampling variance. However, common violations of stochastic assumptions will lead to data more variable than assumed and can do so without affecting structural aspects of the model. In this case, there is an overdispersion coefficient, *c*, such that $c > 1$ and actual variances are obtainable as $c \times$ theoretical variances. Typically with overdispersion *c* is only a little larger than 1, say $1 < c < 4$; *c* is estimated on the basis of the data.

We denote the quasi-likelihood modifications to AIC and AIC_c as (Lebreton *et al.* 1992; see also, Hurvich and Tsai 1995; Burnham and Anderson 1998)

$$QAIC = \frac{-2\log(L(\hat{\theta}))}{\hat{c}} + 2K,$$

$$\text{and QAIC}_c = \frac{-2\log(L(\hat{\theta}))}{\hat{c}} + 2K + \frac{2K(K+1)}{n-K-1},$$

$$= \text{QAIC} + \frac{2K(K+1)}{n-K-1}$$

When no overdispersion exists $c = 1$, so the formulae for QAIC and QAIC_c then reduce to AIC and AIC_c, respectively. We note that $\hat{c} < 1$ should not be used (use 1) and when c is estimated it counts as a parameter and should be in K, the number of estimable parameters in the model (this last point was not mentioned or done in Burnham and Anderson 1998 – an oversight). Only one estimate of c should be used along with a set of models (varying $\hat{c}$ over the models produces invalid results). Often there will be a global model wherein all other models are nested within the global model. Then we obtain $\hat{c}$ from the goodness-of-fit Chi-square statistic (χ^2) for the global model and its degrees of freedom (d.f.):

$$\hat{c} = \chi^2/\text{d.f.}$$

More discussion and guidance on QAIC, $\hat{c}$ and variance inflation using $\hat{c}$ are given in Burnham and Anderson (1998).

An Example

Pistorius *et al.* (2000) evaluated age- and sex-dependent rates of tag loss in southern elephant seals and used information-theoretic methods as the basis for data analysis and inference. Specifically, they considered 4 models representing rates of tag loss being either constant or age- or sex-dependent. We will use their results and make a number of extensions for illustrative purposes. We are not attempting to present a reanalysis and reinterpretation of their data; instead, we wish only to show that additional steps might be considered. Details of this study are contained in Pistorius *et al.* (2000) and we assume the reader is familiar with this paper.

We performed a goodness-of-fit test (essentially Test 2, Burnham *et al.* 1987) on these data, partitioned by gender and found evidence of overdispersion (χ^2 for males = 157.20, d.f. = 77; χ^2 for females = 97.92, d.f. = 84; pooled χ^2 = 255.12, d.f. = 161). An estimate of the variance inflation factor was $\hat{c}$ = 255.12/161 = 1.58. This estimate of c may reflect primarily heterogeneity rather than a lack of independence. Interestingly, much of the lack of fit was attributed to a single cell in the data for both males and females (the same cell, by gender). Had these two cells been in line with what was expected from the general model, the estimate of the variance inflation factor would have been only 1.30.

QAIC, rather than AIC, was used for model selection and model-based estimates of sampling variance should be multiplied by 1.58. Pistorious *et al.* (2000) used the bootstrap to get robust estimates of sampling variance, thus their estimates of precision should appropriately reflect the overdispersion. The results are summarised in Table 1 and provide substantial support for the model that allows tag loss to be both age- and sex-dependent. Support for the model with age-dependence, but not sex-dependence, is more limited; the evidence ratio for the best model versus the second best model is 0.82/0.18 = 4.6.

Inference concerning tag loss could be made from the best model, whereby tag loss is a function of both age and sex. Alternatively, model averaging could be used to allow a robust inference of the derived parameters shown in Table 3 of Pistorius *et al.* (2000). In this case, the model parameters are the β_i and the age- and sex-dependent estimates of tag loss are derived from the β_i. Model averaging in this example would slightly minimise the difference in estimates of tag loss by gender, relative to those shown in the original paper. Clearly, there is essentially no support for the model whereby tag loss is independent of age and sex or the model where tag loss is only sex-dependent.

To measure the relative importance of variables the w_i values can be summed for all models (only 2 here) with age-dependence and all models with sex-dependence. In the example, $w_+(\text{age}) = 1$, whereas $w_+(\text{sex}) = 0.82$, confirming that age is the more important variable in explaining tag loss in these seals. In this example, model-selection uncertainty was minor as the data point to the model allowing both age- and sex-specific tag loss. Other examples where substantial model-selection uncertainty exists are given in Burnham and Anderson (1998: chapter 5).

Recommendations and Summary

There needs to be increased attention to separate those inferences that rest on *a priori* considerations from those resulting from some form of data dredging (see Mayo 1996).

Table 1. Model selection statistics for the southern elephant seal data to estimate tag loss
See Pistorius *et al.* (2000)

Model	$-\log(L)/\hat{c}$	K^A	QAIC	Δ_i	w_i
Age-constant, sex-constant	1,845	3	2,341	39	0.00
Age-dependent, sex-constant	1,815	4	2,305	3	0.18
Age-constant, sex-dependent	1,845	4	2,343	41	0.00
Age-dependent, sex-dependent	1,811	5	2,302	0	0.82

[A]This total includes the estimation of the overdispersion parameter c.

Essentially, no justifiable theory exists to estimate precision (or test hypotheses, for those still so inclined) when data dredging has taken place. The theory (mis)used is for *a priori* analyses, assuming the model was the only one fit to the data. This glaring fact is either not understood by practitioners and journal editors or is simply ignored. Two types of data dredging include (1) an iterative approach where patterns and differences observed after initial analysis are 'chased' by repeatedly building new models with these effects included and (2) analysis of 'all possible models'. Data dredging is a poor approach to making reliable inferences about the sampled population and both types of data dredging are best reserved for more exploratory investigations that probably should remain unpublished. The incorporation of *a priori* considerations is of paramount importance and, as such, editors, referees and authors should pay much closer attention to these issues and be wary of inferences obtained from *post hoc* data dredging.

At a conceptual level, reasonable data and a good model allow a separation of 'information' and 'noise.' Here, information relates to the structure of relationships, estimates of model parameters and components of variance. Noise then refers to the residuals: variation left unexplained. We can use the information extracted from the data to make proper inferences and achieve what Romesburg (1981) termed 'reliable information'. We want an approximating model that minimises information loss, $I(f, g)$, and properly separates noise (non-information or entropy) from structural information. In a very important sense, we are not trying to model the data; instead, we are trying to model the information in the data.

Information-theoretic methods are based on deep theory and are quite effective in making strong inferences from the analysis of empirical data. These methods are relatively simple to understand and practical to employ across a very large class of empirical situations and scientific disciplines. The methods are easy to compute by hand if necessary (assuming one has the parameter estimates $\hat{\theta}_i$, the conditional variances, $\hat{\text{var}}(\hat{\theta}_i|g_i)$, and the maximised log-likelihood values for each of the R candidate models from standard statistical software). Researchers can easily understand the heuristics and application of the information-theoretic methods presented here; we believe it is very important that people understand the methods they employ. Information-theoretic approaches should not be used unthinkingly; a good set of *a priori* models is essential and this involves professional judgment and integration of the science of the issue into the model set.

Publication of results under the information-theoretic paradigm would typically have substantial material in the *Methods* section to discuss and fully justify the candidate models in the set, whereas the *Results* section would typically present a table showing AIC_c or QAIC_c, K, the maximised $\log_e(L)$, Δ, and w for each of the R models, followed by an effective discussion of the scientific interpretation of the table entries. Further material, including many examples, on information-theoretic methods can be found in recent books by Burnham and Anderson (1998) and McQuarrie and Tsai (1998). Akaike's collected works have been recently published (Parzen *et al.* 1997) and this book will be of interest to the more quantitatively fit.

An interesting application of the information-theoretic approach is in conflict resolution in applied aspects of ecology and environmental science (see Anderson *et al.* 1999 for a general protocol). Here, there are opposing parties in a technical controversy and data are available that bear on the resolution of the disagreement. In such cases, models would be built to represent the position of each of the parties. For example, consider the case where there are 3 parties and each party might have 2 models that represent their general position; thus there are $R = 6$ models in the set. Computation of AIC_c and Δ for each model would allow a ranking of the various positions (models), while the Akaike weights would allow a scaling and weight of evidence for the opposing parties and their positions. This approach has not yet been tried in a real controversy to our knowledge (but see Anderson *et al.* 2001).

Acknowledgments

Dr Peter Boveng (NMFS) provided a summary of the seal tag-loss data for our use in computing goodness-of-fit tests and for arranging for our use of the data from senior author Pistorious. Dr Richard Barker and an anonymous referee offered valuable suggestions that allowed the manuscript to be improved.

References

Agresti, A. (1990). 'Categorical Data Analysis.' (John Wiley & Sons: New York.)

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

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* **19**, 716–723.

Akaike, H. (1977). On entropy maximization principle. In 'Applications of Statistics'. (Ed. P. R. Krishnaiah.) pp. 27–41. (North Holland: Amsterdam.)

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics* **16**, 3–14.

Akaike, H. (1983*a*). Statistical inference and measurement of entropy. In 'Scientific Inference, Data Analysis, and Robustness'. (Eds G. E. P. Box, T. Leonard and C.-F. Wu.) pp. 165–189. (Academic Press: London.)

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

Akaike, H. (1985). Prediction and entropy. In 'A Celebration of Statistics'. (Eds A. C. Atkinson and S. E. Fienberg.) pp. 1–24. (Springer: New York.)

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In 'Breakthroughs in Statistics. Vol. 1'. (Eds S. Kotz and N. L. Johnson.) pp. 610–624. (Springer-Verlag: London.)

Akaike, H. (1994). Implications of the informational point of view on the development of statistical science. In 'Engineering and Scientific Applications. Vol. 3. Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach'. (Ed. H. Bozdogan.) pp. 27–38. (Kluwer Academic Publishers: Dordrecht, The Netherlands.)

Anderson, D. R., and Burnham, K. P. (1999). General strategies for the analysis of ringing data. *Bird Study* **46**(suppl.), S261–270.

Anderson, D. R., Burnham, K. P., Franklin, A. B., Gutíerrez, R. J., Forsman, E. D., Anthony, R. G., White, G. C., and Shenk, T. M. (1999). A protocol for conflict resolution in analyzing empirical data related to natural resource controversies. *Wildlife Society Bulletin* **27**, 1050–1058.

Anderson, D. R., Burnham, K. P., and Thompson, W. L. (2000). Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management* **64**, 912–923.

Anderson, D. R., Burnham, K. P., and White, G. C. (2001). Kullback–Leibler information in resolving natural resource conflicts when definitive data exist. *Wildlife Society Bulletin*.

Azzalini, A. (1996). 'Statistical Inference Based on the Likelihood.' (Chapman and Hall: London.)

Berger, J. O. (1985). 'Statistical Decision Theory and Bayesian Analysis.' 2nd Edn. (Springer-Verlag: New York.)

Boltzmann, L. (1877). Uber die Beziehung zwischen dem Hauptsatze der mechanischen Warmetheorie und der Wahrscheinlicjkeitsrechnung respective den Satzen uber das Warmegleichgewicht. *Wiener Berichte* **76**, 373–435.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

Buckland, S. T., Anderson, D. R., Burnham K. P., and Laake, J. L. (1993). 'Distance Sampling: Estimating Abundance of Biological Populations.' (Chapman and Hall: London.)

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

Burnham, K. P., and Anderson, D. R. (1998). 'Model Selection and Inference: a Practical Information-Theoretic Approach. (Springer-Verlag: New York.)

Burnham, K. P., Anderson, D. R., White, G. C., Brownie, C., and Pollock, K. H. (1987). Design and analysis methods for fish survival experiments based on release–recapture. American Fisheries Society, Monograph No. 5. 437 pp.

Burnham, K. P., Anderson, D. R., and White, G. C. (1996). Meta-analysis of vital rates of the northern spotted owl. *Studies in Avian Biology* **17**, 92–101.

Chamberlin, T. (1965). The method of multiple working hypotheses. *Science* **148**, 754–759. [Reprint of 1890 paper in *Science*.]

Cherry, S. (1998). Statistical tests in publications of The Wildlife Society. *Wildlife Society Bulletin* **26**, 947–953.

Cover, T. M., and Thomas, J. A. (1991). 'Elements of Information Theory.' (John Wiley and Sons: New York.)

deLeeuw, J. (1992). Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In 'Breakthroughs in Statistics. Vol. 1'. (Eds S. Kotz and N. L. Johnson.) pp. 599–609. (Springer-Verlag: London.)

Dennis, B. (1996). Should ecologists become Bayesians? *Ecological Applications* **6**, 1095–1103.

Edwards, A. W. F. (1992). 'Likelihood.' Expanded Edn. (The Johns Hopkins University Press: Baltimore, Maryland.)

Efron, B., and Tibshirani, R. J. (1993). 'An Introduction to the Bootstrap.' (Chapman and Hall: New York.)

Ellison, A. M. (1996). An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* **6**, 1036–1046.

Forster, M. R. (1995). Bayes or bust: the problem of simplicity for a probabilistic approach to confirmation. *British Journal for the Philosophy of Science* **46**, 399–424.

Forster, M. R. (2000). Key concepts in model selection: performance and generalizability. *Journal of Mathematical Psychology* **44**, 205–231.

Forster, M. R. (2001). The new science of simplicity. In 'Simplicity, Inference and Econometric Modelling'. (Eds H. Keuzenkamp, M. McAleer and A. Zellner.) (Cambridge University Press.)

Forster, M. R., and Sober, E. (1994). How to tell simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal of the Philosophy of Science* **45**, 1–35.

Gelman, A, Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). 'Bayesian Data Analysis.' (Chapman and Hall: London.)

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science* **14**, 382–417.

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

Hurvich, C. M., and Tsai, C.-L. (1995). Model selection for extended quasi-likelihood models in small samples. *Biometrics* **51**, 1077–1084.

Inman, H. F. (1994). Karl Pearson and R. A. Fisher on statistical tests: a 1935 exchange from *Nature*. *The American Statistician* **48**, 2–11.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Review* **106**, 620–630.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management* **63**, 763–772.

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

Mayo, D. G. (1996). 'Error and Growth of Experimental Knowledge.' (University of Chicago Press: London.)

McQuarrie, A. D. R., and Tsai, C.-L. (1998). 'Regression and Time Series Model Selection.' (World Scientific Press: Singapore.)

Nester, M. (1996). An applied statistician's creed. *Applied Statistics* **45**, 401–410.

Parzen, E., Tanabe, K., and Kitagawa, G. (Eds) (1998). 'Selected Papers of Hirotugu Akaike.' (Springer-Verlag: New York.)

Pistorius, P. A., Bester, M. N., Kirkman, S. P., and Boveng, P. L. (2000). Evaluation of age- and sex-dependent rates of tag loss in southern elephant seals. *Journal of Wildlife Management* **64**, 373–380.

Romesburg, H. C. (1981). Wildlife science: gaining reliable knowledge. *Journal of Wildlife Management* **45**, 293–313.

Royall, R. M. (1997). 'Statistical Evidence: a Likelihood Paradigm.' (Chapman and Hall: London.)

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423 & 623–656.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* **39**, 44–47.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* **61**, 439–447.

White, G. C., Burnham, K. P., and Anderson, D. R. (2001). Advanced features of program MARK. In 'Integrating People and Wildlife for a Sustainable Future. Proceedings of the Second International

Wildlife Management Congress'. (Ed. R. Fields.) (The Wildlife Society: Bethesda, Maryland.)
Yoccoz, N. G. (1991). Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* **72**, 106–111.
Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology* **44**, 41–61.

Manuscript received 24 November 1999; accepted 4 September 2000