ELSEVIER

Modelling fish growth: Model selection, multi-model inference and model selection uncertainty

Stelios Katsanevakis*

*Department of Zoology-Marine Biology, School of Biology, University of Athens, Panepistimioupolis,
15784 Athens, Greece*

Abstract

Model selection based on information theory is a relatively new paradigm in biological sciences with several advantages over the classical approaches. The aim of the present study was to apply information theory in the area of modelling fish growth and to show how model selection uncertainty may be taken into account when estimating growth parameters. The methodology was applied for length–age data of four species of fish, taken from the literature. Five-candidate models were fitted to each dataset: von Bertalanffy growth model (VBGM), generalized VBGM, Gompertz growth model, Schnute–Richards growth model, and logistic. In each case, the 'best' model was selected by minimizing the small-sample, bias-corrected form of the Akaike information criterion (AIC). To quantify the plausibility of each model, given the data and the set of five models, the 'Akaike weight' w_i of each model was calculated. The average model was estimated for each case based on w_i. Following a multi-model inference (MMI) approach, the model-averaged asymptotic length $\bar{L}_\infty$ for each species was estimated, using all five models, by model-averaging estimations of L_∞ and weighting the prediction of each model by w_i. In the examples of this study, model selection uncertainty caused a magnification of the standard error of the asymptotic length of the best model (up to 3.9 times) and thus in all four cases estimating L_∞ from just the best model would have caused overestimation of precision of the asymptotic length. The VBGM, when used for inference, without being the best model, could cause biased point estimation and false evaluation of precision. Model selection uncertainty should not be ignored even if VBGM is the best model. Multi-model inference by model-averaging, based on Akaike weights, is recommended for making robust parameter estimations and for dealing with uncertainty in model selection.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Bertalanffy; Gompertz; Growth models; Logistic; Schnute–Richards

1. Introduction

The process of individual growth in fish is rooted in physiological processes and is the net result of two opposing processes, catabolism and anabolism (von Bertalanffy, 1938). For population analysis, expression of the mean individual body growth of fish in a mathematical expression is needed, relating the size of the species to its age (in the present study, length will be used as a measure of size). Several models have been proposed to estimate the mean growth of individual fish in a population, some of these are based on purely empirical relationship, while others have a theoretical basis and are arrived at by differential equations that link the anabolic and catabolic processes.

The most studied and commonly applied model among all the length–age models is the von Bertalanffy growth model (VBGM) (von Bertalanffy, 1938). Other commonly used alternatives are the generalized VBGM (Pauly, 1979), the Gompertz growth model (Gompertz, 1825), the logistic model (Ricker, 1975), and the Schnute–Richards model (Schnute and Richards, 1990).

The most common approach in length–age data analyses is to 'pick' one model, usually VBGM, based on the shape of the desired curve and on biological assumptions, and fit it to the data. Inference and estimation of parameters and their precision are based solely on that fitted model. Another approach is to fit more than one model to the data and then use a criterion like minimizing residual sum of squares or maximizing adjusted R^2 to select the 'best' model. When the models are nested, a statistical hypothesis testing approach can be used with an F test or with a likelihood ratio test when the probability distribution of the error structure is specified (Quinn and Deriso, 1999).

* Tel.: +30 2107274634.
  E-mail addresses: skatsan@biol.uoa.gr, stelios@katsanevakis.com.

During the past 20 years, modern statistical science has been moving away from traditional methodologies based on statistical hypotheses testing. In particular, traditional approaches of hypotheses testing, when applied to model selection, have been often found to be mediocre and of limited value (Akaike, 1981) and it is suggested that its application will diminish in future (Burnham and Anderson, 2002). Hypotheses testing schemes are based on arbitrary *a* levels (commonly 0.05 or 0.01), multiple testing is challenging, and tests between models that are not nested are problematic. The adjusted coefficient of multiple determination (adj.-R^2) that is often used in model selection was found to be a very poor approach (McQuarrie and Tsai, 1998).

Model selection based on information theory is a relatively new paradigm in biological sciences and is quite different from the usual methods based on null hypothesis testing. The basis of the information theory approach to model selection and inference is Kullback–Leibler (K–L) information (or K–L distance), $I(f, g) = \int f(x) \log(f(x)/g(x|\theta)) \, dx$ (Kullback and Leibler, 1951). $I(f, g)$ is the 'information' lost when model *g* (with parameters θ) is used to approximate full reality or truth *f*; equivalently $I(f, g)$ is interpreted as the distance from the approximating model to full reality (Burnham and Anderson, 2002). Information theory proposes the minimization of K–L distance as a fundamental basis for model selection. However, K–L distance cannot be computed without knowledge of both full reality *f* and the parameters of every candidate model *g*. Akaike (1973) devised a method to approximate K–L distance, based on the empirical log-likelihood function. His method, Akaike's information criterion or AIC, is summarized in the formula $\text{AIC} = -2 \log(\mathcal{L}(\hat{\theta}|\text{data})) + 2K$, where $\log(\mathcal{L}(\hat{\theta}|\text{data}))$ is the numerical value of the log-likelihood at its maximum point, $\hat{\theta}$ the vector of the estimated parameters of the model, and *K* is the number of estimated parameters. It has to be emphasized that AIC is not a 'test' in any sense and that, contrary to the use of other criteria like adj.-R^2, there is a sound theoretical basis for AIC. The derivation of AIC and related issues are given in detail in Burnham and Anderson (2002).

According to the information theory approach, data analysis is taken to mean the integrated process of *a priori* specification of a set of candidate models (based on the science of the problem), model selection based on the principle of parsimony according to AIC, and the estimation of parameters and their precision. The principle of parsimony implies the selection of a model with the smallest possible number of parameters for adequate representation of the data, a bias versus variance tradeoff.

The information theory method frees the researcher from the limiting concept that the proper approximating model is somehow 'given'. When a model is 'picked' in some way, independent of the data, and used to approximate the data as a basis for inference, both the uncertainty associated with model selection and the benefits of selecting a parsimonious model are ignored. This strategy incurs substantial costs in terms of reliable inferences, because uncertainty in model selection is assumed to be zero. If model selection uncertainty is ignored, precision is likely overestimated, estimated confidence intervals of the parameters are often below the nominal level, and predictions may be less accurate than expected. When the data support evidence of more than one model, model-averaging the predicted response variable across models is advantageous in reaching a robust inference that is not conditional on a single model. Rather than estimating parameters from only the 'best' model, parameter estimation can be made from several or even all the models considered. This procedure is termed multi-model inference (MMI) and has several theoretical and practical advantages (Burnham and Anderson, 2002). Although there are few instances of the usage of AIC for selection of fish growth models (e.g. Tsangridis and Filippousis, 1994; Imai et al., 2002), there is no instance, to my knowledge, of MMI based on information theory in this field.

In the present article, model selection and MMI based on information theory approach is applied for length–age data. Some examples are given using data from the literature and the results are compared with those where a single model (VBGM) is arbitrarily picked or where a single 'best' model is used for inference.

2. Methods

2.1. Data sets

Four sets of length–age data were taken from the literature: for the yellowfin tuna (*Thunnus albacares*) (both sexes) in the western equatorial Atlantic (Lessa and Duarte-Neto, 2004), for the striped seabream (*Lithognathus mormyrus*) (both sexes) in the Canarian archipelago (Pajuelo et al., 2002), for the male sandbar shark (*Carcharhinus plumbeus*) in the waters of northeastern Taiwan (Joung et al., 2004), and for the rougheye rockfish (*Sebastes aleutianus*) (both sexes) in southeastern Alaska (Quinn and Deriso, 1999). These examples were picked randomly from the literature and were not selected for any desired properties. In the literature, researchers adopt several approaches for data analysis of length–age data. They may use raw observed, raw back-calculated, mean observed (per age class), or mean back-calculated (per age class or per annulus) data. In the examples of the present study, a variety of approaches were applied. To model the data of the rougheye rockfish, average observed lengths of each age class were used, and weighted non-linear least squares (nl-LS) with the sample size of each age class as weights. Mean back-calculated lengths were used for the yellowfin tuna (fork lengths; with weighted nl-LS), while all (for every annulus) mean back-calculated total lengths were used for the sandbar shark (weighted nl-LS) and for the striped seabream (non-weighted nl-LS).

2.2. Set of candidate models

A set of five-candidate models was used for each dataset to model fish growth: VBGM, generalized VBGM, Gompertz, Schnute–Richards, and logistic; g_1–g_5, respectively. The underlying principle of the VBGM is that the growth rate of fish tends to decrease linearly with size, as indicated in the equation $(\mathrm{d}L/\mathrm{d}t) = k_1(L_\infty - L)$, where k_1 is a relative growth rate parameter (with units yr^{-1}) and L_∞ is the asymptotic length (L_∞ has the same biological meaning in all the models described