Fisheries Research 183 (2016) 119–127

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Evaluation of the prediction skill of stock assessment using hindcasting

Laurence T. Kell[a,*], Ai Kimoto[b], Toshihide Kitakado[c]

[a] ICCAT Secretariat, C/Corazón de María, 8, 28002 Madrid, Spain
[b] Bluefin Tuna Resources Division, National Research Institute of Far Seas Fisheries, Fisheries Research Agency, 5-7-1 Orido, Shimizu, Shizuoka 424-8633, Japan
[c] Faculty of Marine Science, Tokyo University of Marine Science and Technology, Department of Marine Biosciences, 5-7, Konan 4, Minato-ku, Tokyo 108-8477, Japan

ARTICLE INFO

Article history:
Received 9 February 2016
Received in revised form 18 May 2016
Accepted 19 May 2016
Handled by A.E. Punt

Keywords:
Abundance indices
Cross-validation
Projection
Retrospective analysis
Stock assessment
Taylor diagrams

ABSTRACT

A major uncertainty in stock assessment is the difference between models and reality. The validation of model prediction is difficult, however, as fish stocks can rarely be observed and counted. We therefore show how hindcasting and model-free validation can be used to evaluate multiple measures of prediction skill. In a hindcast a model is fitted to the first part of a time series and then projected over the period omitted in the original fit. Prediction skill can then be evaluated by comparing the predictions from the projection with the observations. We show that uncertainty increased when different datasets and hypotheses were considered, especially as time-series of model-derived parameters were sensitive to model assumptions. Using hindcasting and model-free validation to evaluate prediction skill is an objective way to evaluate risk, i.e., to identify the uncertainties that matter. A hindcast is also a pragmatic alternative to hindsight, without the associated risks. While the use of multiple measures helps in evaluating prediction skill and to focus research onto the data and the processes that generated them.

1. Introduction

The provision of fisheries management advice requires the assessment of stock status relative to reference points, the prediction of the response of a stock to management, and checking that predictions are consistent with reality. In most fishery management frameworks a stock is defined on operational rather than an ecological or evolutionary basis (Waples and Gaggiotti, 2006). In this paper a stock is defined as a population or subpopulation of a species for which parameters such as growth, recruitment, mortality, and fishing mortality are regarded as being homogeneous, and which have the main effect on determining the dynamics; extrinsic factors such as immigration and emigration are traditionally ignored.

Stock assessments sometimes proven to be wrong in retrospect, due to poor model assumptions or to data that do not reflect the key processes (Schnute and Hilborn, 1993). To evaluate uncertainty often a number of scenarios are considered corresponding to alternative model structures and dataset choices (Hilborn, 2016). It is difficult, however, to empirically validate stock assessment models as it is seldom possible to observe fish populations directly. Therefore techniques such as retrospective analysis, where a model is fitted to increasing periods of data to identify systematic inconsistencies (Mohn, 1999), or simulation are used. Deroba et al. (2015) summarised an extensive state-of-the-art simulation exercise to compare stock assessment models. This was limited to the evaluation of historical and current estimates of stock status based on self- and cross-tests. Both approaches evaluate consistency rather reliability, where a reliable model provides accurate results despite uncertainty.

One approach to address uncertainty in historical estimates of stock status is to integrate multiple diverse datasets to try and extract as much information as possible about modelled processes (Fournier et al., 1998). An implicit assumption is that integrated models can compensate for lack of good data. Models are by definition, however, simplifications of reality and model misspecification can lead to degradation of results when there are multiple potentially conflicting data sets. For example Payne et al. (2009) showed that including all available data in stock assessments may lead to high noise levels and poor-quality assessments, and recommended that the choice of data should be based on rational and justifiable selection criteria. It is therefore critical to determine what drives an assessment (Francis and Hilborn, 2011).

* Corresponding author.
E-mail address: Laurie.Kell@iccat.int (L.T. Kell).

http://dx.doi.org/10.1016/j.fishres.2016.05.017
0165-7836/

To check that predictions are consistent with reality it is necessary to evaluate prediction skill (e.g., Walters and Punt, 1994; Patterson et al., 2001; Ralston et al., 2011); a statistical evaluation of the accuracy of a prediction relative to a reference model or dataset. Prediction skill can be used to compare alternative models or observations used for prediction to a reference set of estimates or data (e.g., Jin et al., 2008; Weigel et al., 2008; Balmaseda et al., 1995). If data are regarded as being representative of the dynamics of the stock then they can be used as a model-free validation measure (Hjorth, 1993), and the best performing scenarios (e.g., choice of models and data) can be identified by comparing predictions with observations. Stock biomass cannot actually be observed so if estimates of population abundance were compared in the hindcast this would be model-based validation.

Errors and uncertainty in historical parameter estimates, particularly in the most recent years, propagate into predictions. Different stock assessment packages often use different methods for estimating and propagating those errors, and the choices made will affect the robustness of management advice (Patterson et al., 2001; Magnusson et al., 2012). Validation of predictions is therefore as least as important as examining diagnostics for fits to historical data. More effort, however, appears to be going into the latter than the former, unlike in other fields such as meteorology and oceanography (e.g., Murphy and Winkler, 1987; Doswell III et al., 1990; Schaefer, 1990; Roebber, 2009), where the ability to predict is more important than the description of past states.

Hindcasting is widely used, in oceanography and meteorology where the state of a system is observable, to evaluate prediction skill (Huijnen et al., 2012). Hindcasting is a conceptually simple form of cross-validation, which has no parametric or theoretic assumptions allowing it to be used for comparisons across different models and datasets. In a hindcast, a model is first fitted using a truncated time series, dynamics are projected forward using the model and predictions compared to recent observations not used in fitting (e.g., Christoffersen and Pelletier, 2004; Pastoors et al., 2007; Heath et al., 2004). Although hindcasting is not commonly used in stock assessment it combines two individual procedures routinely used, namely retrospective analysis and projection.

An objective of the paper is to show how hindcasting and using multiple measures for prediction skill can help in the development of robust stock assessment advice frameworks. We use hindcasting to evaluate the prediction skill of series of catch per unit effort (CPUE) used as indices of stock abundance, across a range of stock assessment scenarios. Since time series of CPUE are often the most influential inputs to stock assessment models (Francis and Hilborn, 2011) it is important to be aware of the limitations of these data when fitting models with them.

2. Materials and methods

We chose Atlantic and Mediterranean bluefin (*Thunnus thynnus*) as a case study to show how hindcasting can provide insight into stock assessment uncertainty, to illustrate the benefits of the approach and help identify ways forward. A reason for the choice is because it is a stock of high value with well documented uncertainty about current stock status and response of the stock to management (Fromentin et al., 2014; Leach et al., 2014).

2.1. Stock assessment

Atlantic bluefin is assessed by the International Commission for the Conservation of Atlantic Tuna (ICCAT) using Virtual Population Analysis (VPA). VPA sums catch numbers-at-age backwards down a cohort, adjusted by losses due to natural mortality (Pope, 1972). Indices of relative stock abundance allow the numbers in the oldest age class of a cohort (terminal numbers-at-age) to be derived using maximum likelihood.

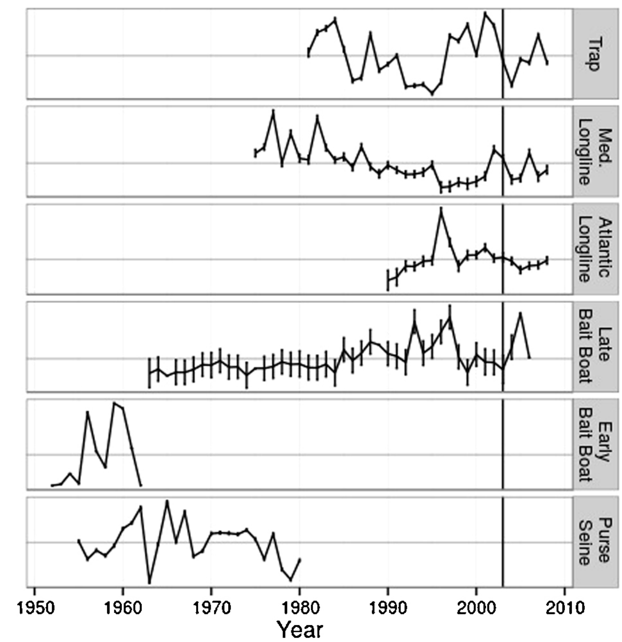

Fig. 1. Time series of CPUE, the error bars are the CVs derived from the standard errors of the GLM predictions; the vertical line corresponds to 2003 the start of the hindcast (standardized to have a mean of 0 and variance of 1).

Two main datasets are used in the Atlantic bluefin assessment (ICCAT, 2015), i.e., catch-at-age and CPUE. Catch-at-age data are derived from total reported catches and samples of size data using age slicing. The raw catch and effort data are standardised using generalised linear models (GLMs) to remove the effect of factors that bias CPUE when used as an index of abundance (Maunder et al., 2006). Although the majority of the catch comes from the Mediterranean purse seine fishery, catches taken with this gear do not provide a reliable estimate of stock abundance and the CPUE from this gear are not used in the assessment.

Time series of catch numbers-at-age start in 1950 and the assessment included six CPUE series from fleets using four fishing gears (Table 1; Fig. 1). Data after 2008 are not used in this study since the implementation of the bluefin recovery plan affected catch rates and the selection of age classes by the fisheries. Fig. 1 shows the time series of CPUE, the error bars are the CVs derived from the standard errors of the GLM predictions. Only four series covered the recent period, namely trap, Mediterranean and Atlantic long line and late period bait boat. All the gears target large bluefin tuna except the bait boats that target juveniles. As the purse seine and early period bait boats series do not cover the recent period they were not included in the hindcast analysis.

The time series of CPUE are in biomass and represent all ages in the catch. Therefore when fitting the VPA these have to be transformed into numbers-at-age based on the vulnerability of age-classes to the fishery and the mass-at-age, i.e.

$$\hat{I}_{iy} = q_i \delta_i \Sigma_a v_{ia} w_{iay} \tilde{N}_{ay} \quad (1)$$

where q_i is the catchability coefficient, δ_i an adjustment for time of fishing, v_{ia} the relative vulnerability-at-age, w_{iay} mass-at-age, and $\tilde{N}_{ay}$ the estimated of numbers-at-age. The subscripts are i for the CPUE series, a for age and y for year.

v is given by

$$v_{ia} = \frac{\Sigma_y C_{iay} F_{ay} / C_{ay}}{\max_a \{C_{iay} F_{ay} / C_{ay}\}} \quad (2)$$