Fisheries Research xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Fisheries Research

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

Some insights into data weighting in integrated stock assessments

André E. Punt[a,b,*]

[a] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195-5020, USA
[b] CSIRO Oceans and Atmosphere Flagship, Castray Esplanade, Hobart, Tasmania, Australia

ARTICLE INFO

Article history:
Received 31 August 2015
Received in revised form
11 December 2015
Accepted 14 December 2015
Available online xxx

Keywords:
Age-structured stock assessment methods
Conditional age-at-length data
Data weighting
Length-composition data
Stock Synthesis
Simulation
Spatial structure

ABSTRACT

The results of fishery stock assessments based on the integrated analysis paradigm can be sensitive to the values for the factors used to weight each of the data types included in the objective function minimized to obtain the estimates of the parameters of the model. These assessments generally include relative abundance index data, length-composition information and conditional age-at-length data, and algorithms have been developed to select weighting factors for each of these data types. This paper introduces methods for weighting conditional age-at-length data that extend an approach developed by Francis (2011) to weight age- and length-composition data. Simulation based on single-zone and two-zone operating models are used to compare five tuning methods that are constructed as combinations of methods to weight each data type. The single-zone operating models allow evaluation of the tuning methods in terms of their ability to provide unbiased estimates of management-related quantities and the correct data weights in the absence of model mis-specification, while the two-zone operating models allow the impacts of model mis-specification on the performance of tuning methods to be explored. The results of assessments are sensitive to data weighting, but the choice of method for data weighting is most consequential when there is model mis-specification. Overall, the results indicate that arithmetic averaging of effective sampling sample sizes from the McAllister and Ianelli (1997) approach is inferior to other methods, and the new method for computing effective sample sizes for conditional age-at-length data seems most appropriate.

1. Introduction

Increasingly, the scientific advice given to fishery managers is based on the results of integrated population models (Newman et al., 2014). Integrated population models separate the development of the model of the population dynamics from that of the relationship between the data and the model, and allowance can be made for error in both the population dynamics and the observations (Maunder and Punt, 2013). Integrated population models (or integrated analyses) have been used in fisheries for decades, the earliest examples of the method in fisheries being Doubleday (1976), Fournier and Archibald (1982) and Deriso et al. (1985). Use of integrated analysis has been common in fisheries because there are often many data types (e.g., age-and-growth information, catch-rates, survey indices of abundance, catch-at-age data), each of which can provide information about some, but not all, of the parameters or processes that govern the dynamics of exploited fish and invertebrate populations.

Many integrated analysis assessments are now conducted using one of three packages that include generalized estimation frameworks (Stock Synthesis (Methot and Wetzel, 2013), MULTIFAN-CL (Fournier et al., 1998; Hampton and Fournier, 2001), and CASAL (Bull et al., 2005)), although several other, but less generally-applicable, packages have been developed.

One of the key advantages of integrated analysis is the ability to use multiple data sources to estimate the current abundance, trend in abundance, and productivity of populations. However, it is not uncommon for data sources to be in conflict with each other to some extent. Thus, each data type (and each data point within each data type) needs to be assigned a weight. In principle, this weight should relate to the deviation between the data point and its expected value. However, it is not straightforward to objectively select weights, and history reveals that data weighting is influential (e.g., Richards, 1991).

The selection of weights for compositional data (length-composition data, age-composition data, and conditional age-at-length data) is perhaps the most challenging aspect of selecting weights for data (although selecting the extent of variation in pro-

* Corresponding author at: School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195-5020, USA. Fax: +1 206 685 7471.
E-mail address: aepunt@uw.edu

http://dx.doi.org/10.1016/j.fishres.2015.12.006
0165-7836/

Table 1
Notation and derived symbols.

Symbol	Description	Derivation
(a) Length-composition data		
$p_{y,L}$	Observed proportion of animals in length-class L during year y	Input data
$\hat{p}_{y,L}$	Predicted proportion of animals in length-class L during year y	Model prediction
N_y	Input effective sample size for the length-composition data for year y	Input data
n_y	Number of years with length-composition data	Input data
$\bar{L}_L$	Mid-point of length-class L	Input data
E_y	McAllister-Ianelli effective sample size for the length data for year y	$\sum_L \hat{p}_{y,L}(1-\hat{p}_{y,L}) / \sum_L (p_{y,L}-\hat{p}_{y,L})^2$
$\bar{L}_y$	Observed mean length of the catch during year y	$\sum_L \bar{L}_L p_{y,L}$
$\hat{\bar{L}}_y$	Predicted mean length of the catch during year y	$\sum_L \bar{L}_L \hat{p}_{y,L}$
$SE(\hat{\bar{L}}_y)$	Predicted standard error of the mean length of the catch for year y	$\sqrt{\sum_L \hat{p}_{y,L}(\bar{L}_L-\hat{\bar{L}}_y)^2} / \sqrt{N_y}$
(b) Conditional age-at-length data		
$p_{y,L,a}$	Observed proportion of animals in length-class L during year y that are of age a	Input data
$\hat{p}_{y,L,a}$	Predicted proportion of animals in length-class L during year y that are of age a	Model prediction
$N_{y,L}$	Input effective sample size for the conditional age-at-length data for year y and length-class L	Input data
$n_{y,L}$	Number of combinations of years and length-classes with conditional age-at-length data	Input data
$w_{y,L}$	Proportion of the conditional age-at-length data for year y that is in length-class L	$N_{y,L} / \sum_{L'} N_{y,L'}$
$E_{y,L}$	McAllister-Ianelli effective sample size for the conditional age-at-length data for year y and length-class L	$\sum_a \hat{p}_{y,L,a}(1-\hat{p}_{y,L,a}) / \sum_a (p_{y,L,a}-\hat{p}_{y,L,a})^2$
$\bar{a}_{y,L}$	Observed mean age of the catch for year y and length-class L[a]	$\sum_a (a+0.5) p_{y,L,a}$
$\hat{\bar{a}}_{y,L}$	Predicted mean age of the catch for year y and length-class L[a]	$\sum_a (a+0.5) \hat{p}_{y,L,a}$
$SE(\hat{\bar{a}}_{y,L})$	Predicted standard error of the mean age of the catch for year y and length-class L[a]	$\sqrt{\sum_a \hat{p}_{y,L,a}\left((a+0.5)-\hat{\bar{a}}_{y,L}\right)^2} / \sqrt{N_{y,L}}$
$\bar{a}_y$	Observed mean age of the age-length key for year y	$\sum_L w_{y,L} \bar{a}_{y,L}$
$\hat{\bar{a}}_y$	Predicted mean age of the age-length key for year y	$\sum_L w_{y,L} \hat{\bar{a}}_{y,L}$
$SE(\hat{\bar{a}}_y)$	Predicted standard error of the mean age of the catch for year y	$\sqrt{\sum_L (w_{y,L})^2 SE(\hat{\bar{a}}_{y,L})^2}$

[a] The +0.5 is introduced to account the fact that fisheries occur throughout the year.

cess error is often a close second). Until recently, the approach for weighting compositional data was often to apply the following iterative approach:

(a) the values for the weights for the composition data (usually a factor that multiplies some input or initial effective sample size e.g., the number of fish, hauls or even trips sampled) are set;
(b) the population dynamics model is fitted to the data;
(c) the method of McAllister and Ianelli (1997) is used to calculate an overdispersion factor for the composition data, and the residual variances for the indices of abundance are set to the mean square errors;
(d) the values for the weights (the extent of overdispersion for the computational data and the residual variances for the indices of abundance) are replaced by the calculated overdispersion factors (often separately by year) and mean square errors respectively;
(e) steps (b–d) are applied until convergence occurs.

This approach can be criticized for several reasons, including that it may not converge to a sensible result (e.g., resulting in weights of zero or unrealistically high weights for some data types or years within series), but particularly because it fails to account for positive correlation in residuals between adjacent age- or length-classes (Francis, 2011). The likelihood function for compositional data, usually the multinomial, assumes that residuals should be negatively correlated and there should be no "runs" of positive or negative residuals. However, it is commonly the case that there are "runs" of residuals (e.g., Whitten and Punt, 2014). Basing the weighting for the compositional data on the method of McAllister and Ianelli (1997), which assumes that residuals are independent, may lead to over-weighting (compared to other data sources in the assessment) of the compositional data. In principle, a likelihood function could be selected that allows for "runs" of Pearson residuals (such as the multivariate normal). However, it is more common, and computationally easier, to downweight the compositional data using the approach outlined above.

Francis (2011) provided an alternative way to weight age- and length-composition data that accounts for the positive correlation between the residuals (and generally leads to lower weights for such data). However, Francis (2011) did not provide a way to weight conditional age-at-length data. This data type (essentially an age-at-length key) provides key information on growth as well as year-class strength in integrated stock assessments. Unfortunately, "Francis weighting", as originally conceived, was developed for compositional data types that are vectors of numbers whereas an annual age-length key is a matrix.