G Model
FISH-4554; No. of Pages 13

ARTICLE IN PRESS

Fisheries Research xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Fisheries Research

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

Full length article

Can diagnostic tests help identify model misspecification in integrated stock assessments?

Felipe Carvalho[a,b,*], André E. Punt[c], Yi-Jay Chang[d], Mark N. Maunder[e,f], Kevin R. Piner[g]

[a] University of Hawaii, Joint Institute for Marine and Atmospheric Research, Honolulu, HI 96822, USA
[b] NOAA, National Marine Fisheries Service, Pacific Islands Fisheries Science Center, Honolulu, HI 96818, USA
[c] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195, USA
[d] Institute of Oceanography, National Taiwan University, Taipei 10617, Taiwan
[e] Inter-American Tropical Tuna Commission, La Jolla Shores Drive, La Jolla, CA 92037, USA
[f] Center for the Advancement of Population Assessment Methodology, Scripps Institution of Oceanography, La Jolla, USA
[g] NOAA Fisheries, Southwest Fisheries Science Center, La Jolla Shores Dr, La Jolla, CA 92037, USA

ARTICLE INFO

Article history:
Received 14 June 2016
Received in revised form
15 September 2016
Accepted 16 September 2016
Handled by Prof. George A. Rose
Available online xxx

Keywords:
Integrated stock assessments
Simulation
Model diagnostics

ABSTRACT

A variety of data types can be included in contemporary integrated stock assessments to simultaneously provide information on all estimated parameters. Conflicts between data, which are often a symptom of model misspecification and evident as model misfit, can affect the estimates of important parameters and derived quantities. Unfortunately, there are few standard diagnostic tools available for integrated stock assessment models that can provide the analyst with all the information needed to determine if there is substantial model misspecification. In this study, we use simulation methods to evaluate the ability of commonly-used and recently-proposed diagnostic tests to detect model misspecification in the observation model process (i.e., the incorrect form for survey selectivity), systems dynamics (i.e., incorrect assumed values for steepness of the stock-recruitment relationship and natural mortality), and incorrect data weighting. The diagnostic tests evaluated here were: i) residuals analysis (SDNR and runs test); ii) retrospective analysis; iii) the R_0 likelihood component profile; iv) the age-structured production model (ASPM); and v) catch-curve analysis (CCA). The efficacy of the diagnostic tests depended on whether the misspecification was in the observation or systems dynamics model. Residual analyses were easily the best detector of misspecification of the observation model while the ASPM test was the only good diagnostic for detecting misspecification of system dynamics model. Retrospective analysis and the R_0 likelihood component profile infrequently detected misspecified models, and CCA had a high probability of rejecting correctly-specified models. Finally, applying multiple carefully selected diagnostics can increase the power to detect misspecification without substantially increasing the probability of falsely concluding there is misspecification when the model is correctly specified.

1. Introduction

The advantages of ‘integrated’ assessments are numerous, and include the ability to combine many data sources to estimate important population dynamics processes such as growth, natural mortality, fishing mortality and movement simultaneously (Doubleday, 1976; Fournier and Archibald, 1982; Maunder and Punt, 2013; Punt et al., 2013). This is made possible by summing the log-likelihoods from each data component (e.g., abundance indices, size-composition, tagging) into a single total log-likelihood. Another advantage of integrated assessments is that they allow the sensitivity to dataset choice to be evaluated and hence conflicts among datasets and model misspecification to be identified (Maunder and Punt, 2013). The Stock Synthesis (SS) assessment framework (Methot and Wetzel, 2013) is one of the most well-known examples of an integrated model, and has been applied in a wide variety of fish assessments globally (Wetzel and Punt, 2011; Methot and Wetzel, 2013).

However, simultaneously analyzing multiple data sources can lead to conflicts among the data sources, especially between size-composition data and indices of relative abundance (Francis, 2011; Ichinokawaa et al., 2014; Wang et al., 2014; Lee et al., 2014). Most recently, Maunder and Piner (2015) stated that conflicts between

* Corresponding author at: University of Hawaii, Joint Institute for Marine and Atmospheric Research, Honolulu, HI 96822, USA.
E-mail address: felipe.carvalho@noaa.gov (F. Carvalho).

http://dx.doi.org/10.1016/j.fishres.2016.09.018
0165-7836/© 2016 Published by Elsevier B.V.

data sources arise due to: 1) random sampling error, 2) misspecification of the observation model (i.e., the model processes relating the population dynamics or states to data), and 3) misspecification of the system dynamics model (i.e., the population dynamics model). Analysts often down weight some of the data sources when confronted with conflicting data sources (e.g., Harle et al., 2015; Kell et al., 2014). However, this is not necessarily appropriate because it may not resolve the model misspecification (Wang et al., 2015). Deroba and Schueller (2013) and Lee et al. (2014) have shown that model misspecification can substantially bias assessment outcomes, affecting, in particular, parameter estimates, and determination of stock status. For example, assuming that the selectivity of a fishery is asymptotic when it is in fact dome-shaped can substantially bias estimates of absolute abundance (Wang et al., 2009). Alternative model structures can be explored to identify inconsistencies and hence form the basis to justify down weighting some data sources, as well as an indication of what component of the model structure is misspecified (Maunder and Piner, 2015). Francis (2011) recommends prioritizing indices of relative abundance, assuming that these data are representative of changes in stock abundance. However, age- and size-composition data can be more informative about the level of fishing mortality and biomass when the index is uninformative (i.e., there is no contrast in abundance levels) and/or is of poor quality (e.g., high sampling error or the index is not proportional to abundance). Although size- and age-composition data may provide substantial information on absolute abundance, the prioritization of indices of relative abundance is recommended because even slight model misspecifications can have a large impact on the information about absolute abundance contained in compositional data (Maunder and Piner, 2015; Lee et al., 2014; Wang et al., 2014).

There is little guidance and few objective criteria to determine how to best summarize the results of integrated assessments, determine if the model fits the data adequately, and if the model is well specified. Moreover, it is very difficult to easily evaluate convergence or identify problematic areas given the large number of estimable parameters in these assessments (Harley and Maunder, 2003). Applying classical model diagnostic tools in integrated stock assessments requires further investigation and possible refinement before good practice recommendations can be made. Some of the most common or recently proposed diagnostic tests to be used with integrated stock assessments include:

- *Residual analysis*. Analysis of residuals is perhaps the most common way to determine a model's goodness-of-fit (Cox and Snell, 1968). Residuals are examined for patterns to evaluate whether the model assumptions have been met (e.g., Wang et al., 2009). Many statistics exist to evaluate the residuals for desirable properties. One way is to calculate, for each abundance index, the standard deviation of the normalized (or standardized) residuals divided by the sampling (or assumed) standard deviation (SDNR) (Breen et al., 2003; Francis, 2011). The SDNR is a measure of the fit to the data that is independent of the number of data points. A relatively good model fit will be characterized by smaller residuals (i.e. close to zero) and a SDNR close to 1. Francis (2011) notes that it is also necessary to conduct a visual examination between observed and predicted values to be sure that the fit is good even when SDNR values are not much greater than 1. A non-random pattern of residuals may indicate that some heteroscedasticity is present, or there is some leftover serial correlation (serial correlation in sampling/observation error or model misspecification). Several well-known nonparametric tests for randomness in a time-series include: the runs test, the sign test, the runs up and down test, the Mann-Kendall test, and Bartel's rank test (Gibbons and Chakraborti, 1992). In this study, we used the runs test to evaluate whether residuals are random over time, because this test has been used to diagnose fits to indices and other data components in assessment models (e.g. SEDAR 40, 2015).
- *Retrospective analysis.* Retrospective analysis is another diagnostic approach widely used in stock assessment to evaluate the reliability of parameter and reference point estimates (Cadigan and Farrell, 2005; Hurtado-Ferro et al., 2014). Retrospective analysis involves fitting a stock assessment model to the full dataset, and the same model is then fitted to truncated datasets where the data for the most recent years have been sequentially removed. Retrospective analysis usually assumes that the estimates of historical abundance from the current assessment that uses all the data are more accurate than the estimates of "current" abundance from assessments that ignore recent data, therefore revealing possible bias of model predictions. In stock assessment, the "*e*; ρ"*e*; statistic proposed by Mohn (1999) is commonly used to evaluate the severity of retrospective patterns (Deroba, 2014). This statistic measures the average of relative difference between an estimated quantity from an assessment (e.g., biomass in final year) with a reduced time-series and the same quantity estimated from an assessment using the full time-series. According Hurtado-Ferro et al. (2014), retrospective patterns generally arise from two main causes: time-varying processes unaccounted for in the assessment (i.e., model misspecification), or incomplete data.
- *R_0likelihood component profile.* Negative log-likelihoods of various data components for a profiled parameter (e.g., virgin recruitment) have been used as a diagnostic to evaluate the influence of each data component on estimates of model parameters and outputs (e.g., Maunder, 1998; Maunder and Starr, 2001; Francis, 2011; Lee et al., 2014; Ichinokawaa et al., 2014; Maunder and Piner, 2015). Wang et al. (2014) proposed an extension of R_0 (virgin recruitment) likelihood profiling to diagnose stock assessment models with misspecified selectivity. Their method consists of constructing a R_0 profile for data components simulated without error from a known stock assessment model. The R_0 profile from the known stock assessment model is assumed to represent the "true" information content of each data component. Any differences in subsequent models from the R_0 profile originated from the known stock assessment model are presumed to indicate conflict in the data or model misspecification. However, this diagnostic has not been used extensively or evaluated, and more research is needed before it can be recommended.
- *Age-structured production model.* In some integrated stock assessments the index of abundance provides almost no information on population scale. Consequently, the estimates of the model outputs rely almost completely on the size- and age-composition data and model structure. Maunder and Piner (2015) proposed a diagnostic tool that can be used to evaluate the information content of data about absolute abundance and assess whether the model is correctly specified. This diagnostic consists of comparing the results of an age-structured production model (ASPM) to those from a model estimating all of the model parameters and fitting to all the data (e.g., an integrated analysis). It is inferred that a production function is apparent in the data when the catch data explain indices with good contrast (e.g., declining and increasing trends), therefore providing evidence that the index is a reasonable proxy of stock trend. If the ASPM cannot mimic the index, then either the stock is recruitment-driven, catch levels have not been high enough to have a detectable impact on the population, the model is incorrect, or the index of relative abundance is uncertain or not proportional to abundance. Similar to the R_0 likelihood component profile, this diagnostic has only begun to be implemented, and its utility remains unknown.
- *Catch-curve analysis*. Most of the information on absolute abundance will come from the compositional data if the index of abundance provides little or no information on population scale.

Please cite this article in press as: Carvalho, F., et al., Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. (2016), http://dx.doi.org/10.1016/j.fishres.2016.09.018