Focus Questions

ASPM and catch-curve analysis

Is there a way to combine the ASPM and catch-curve analysis into one diagnostic (e.g. the cohort depletion model)?

Both the ASPM and the CC diagnostic provide information about the absolute abundance and the conflict between information in indices of relative abundance and composition data. Combining these two diagnostics might provide an additive sources of information that would be more valuable than evaluating each diagnostic in isolation. Maunder et al. () provided a flow chart for combining the ASPM and the R0 profile diagnostics to provide weights for an ensemble model. A similar approach might be useful for combining the ASPM and CC diagnostics.

An alternative approach would be to combine the composition data and the index of relative abundance using a depletion estimator on each cohort (or the method of Clark 2022). This would involve using the composition data from the index to turn the aggregated index into an index by age and then convert the aggregated catch into total catch-at-age. This could be done by using the age composition data or an approximation by converting the length composition into age using the growth equation and associated variation of length-at-age (in a reverse way and possibly using the integrated models estimated numbers at age and selectivity).

Does fixing the recruitment deviates at the integrated model estimates (in addition to selectivity) simply transfer all the information from the integrated model?

Several extensions of the ASPM involved different treatments of recruitment. The traditional ASPM diagnostic assumes deterministic recruitment that is either constant or follows the stock-recruitment relationship. ASPM-Rdev estimates annual recruitment deviates, but still not fitting to the composition data so the only information of temporal variation in recruitment comes from the index of relative abundance. The third approach, ASPM-Rfix, fixes the annual recruitment deviates at the values estimated by the integrated model. The integrated model uses the composition data to provide information on the annual recruitment. Therefore, the ASPM-Rfix uses the composition data to some degree. It is possible that fixing the recruitment deviates (in addition to the selectivity) essentially forces the ASPM to a specific estimate of R0 and may not be very informative.

R0 profile

Should the Rdev sum to zero penalty be used?

Stock Synthesis has an option to penalize the recruitment deviates so that they sum to zero. This is to ensure that the R0 (or the determinist stock-recruitment relationship) represents the average recruitment. However, the recruitment deviate penalty already forces the recruitment deviates to sum to one to some extent. When the R0 is fixed at a value different from the MLE, the model can

compensate by making the average of the recruitment deviates different from zero so that the product of the deviate and R0 stays the same no matter what value the R0 is fixed at. Therefore, since all the recruitment deviates will be adjusted from 0 by the ratio ln(R0_fixed/R0), the recruitment deviate penalty has a large influence on the R0 profile and including the penalty to sum to zero will have even a greater impact. It is unclear if the sum to zero penalty should be used or whether doing it with and without the penalty might provide additional information.

R0*exp(Rdev1) = R0_fixed*exp(Rdev2) so Rdev2 = Rdev1*ln(R0/R0_fixed)

What does the profile for the recruitment deviate penalty tell us?

When the R0 is fixed at a value different from the MLE, the model can compensate by making the average of the recruitment deviates different from zero so that the product of the deviate and R0 stays the same no matter what value the R0 is fixed at. Therefore, since all the recruitment deviates will be adjusted from 0 by the ratio ln(R0_fixed/R0), the recruitment deviate penalty has a large influence on the R0 profile and is often the most influential component of the R0 profile diagnostic. Given that the Rdev penalty is expected to have an influential profile, it is not clear what it tells us about model misspecification.

Would integrating out the recruitment deviates provide more useful likelihood component profiles?

Most stock assessments use the penalized likelihood approach to model temporal variation in recruitment as an approximation to the random effects/state-space approach. Integrating out the recruitment deviates may produce a different R0 component profile, particularly for the recruitment deviate penalty component. It is unknown if the R0 profile would be more informative if the recruitment deviates would be integrated out.

Residuals and effective sample size

Should the variance of the likelihood be based solely on the random sampling error?

The Law of conflicting data suggest that the random sampling error should be estimated from the data and used in the likelihood function. Any difference in the effective sample size for composition data (our other measure of the residuals) and the random sampling variance indicates the presence of model misspecification. However, it is not possible to eliminate all model misspecification, making changes to one model component may account for the misspecification of another model component, and hypothesis tests and model selection requires adequate accounting for the total variance. A good practice might be to fix the likelihood variance based in the random sampling variance, estimate an additional variance, for which the goal is to minimize it by fixing model misspecification.

What proportion of the correlation in (composition) residuals come from sampling and from model misspecification?

On of the issues identified in modelling composition data is the correlation in residuals that is over and above the multinomial variance (Francis ref). It is though that much of this correlation in residuals is due to model misspecification, but some may be inherent in the sampling. Few studies have investigated this correlation (see Fisch et al. xxx for an exception). If most of the correlation is from model misspecification or if the random sampling error correlation can be estimated, measuring the correlation in residuals may provide a way to identify model misspecification (see Fisch et al. presentations).

Retrospective analysis

Can Mohn's Rho be used as a criteria for identifying that a model is misspecified?

The most commonly used metric for evaluating retrospective analysis is Mohn's Rho. This quantity has been used to identify model misspecification. Classification as a strong retrospective can be done in multiple ways. Hurtado-Ferro et al. (2015) used simulation to set values for Mohn's rho according to life history characteristics. Brooks and Legault (2016) compare the rho-adjusted values of spawning stock biomass and fishing mortality rate to the confidence interval of the terminal year in the assessment. Miller and Legault (2017) used a parametric bootstrap to estimate the uncertainty of Mohn's rho. Further research is needed on the statistical properties of Mohn's rho and its ability to characterize model misspecification.

Is rho-adjustment, multiplying terminal year estimates by 1/(1+rho), an appropriate response to a strong retrospective pattern?

Rho-adjustment has been found to improve management advice compared to ignoring a strong retrospective pattern (Brooks and Legault, 2016; Wiedenmann and Jensen, 2018; Wiedenmann and Jensen, 2019). However, the rho-adjustment creates a discontinuous time-series in the terminal year. The rho-adjustment moves the spawning stock biomass and fishing mortality rate in the terminal year in the same direction as multiple fixes to the data that eliminate the retrospective pattern, but often not as far (Legault, 2020). The rho-adjustment was similarly found to be not large enough to eliminate overfishing in empirical stock assessment (Brooks and Legault, 2016; Wiedenmann and Jensen, 2018; Wiedenmann and Jensen, 2019).

How else can models be changed when a strong retrospective pattern is found?

A simpler model that ignores some of the data can be used. However, recent work by the Index-Based Methods Research Track in the Northeast found that none of the simpler methods consistently performed better than rho-adjustment in simulation studies. More complex models can be tried, but overparameterization becomes a concern. Fixes to the data can be used, but identifying the correct source of the retrospective pattern is difficult (ICES Methods Group) and using the wrong fix can lead to poor management advice (Szuwalski et al., 2018). State-space models have shown promise with smaller

retrospective patterns compared to traditional statistical catch-at-age models (ICES Methods Group), but can still exhibit retrospective patterns (Perretti et al., 2020).

Hind casting

Is there a management strategy that relates closely to what we observer (and can then test predictions for)?

Hind casting evaluates the model's ability to predict observed data in, for example, a one-step ahead approach. This is a very useful if the observed data is directly related to the management objective, but management quantities (e.g. depletion level relative to that associated with MAY) are usually quite different from the observed data (catch, relative indices of abundance, or catch composition). It might be useful to modify management quantities and objectives to be more closely related to the observations. For example, management could be setting catch under a given (e.g. historically observed) effort level or the catch that would increase the relative index by a certain percentage.

Bayesian model checking

Why isn't the likelihood function used more often as the discrepancy measure?

Bayesian posterior predictive checks are based on setting a discrepancy measurer to compare the simulated data with the actual data. The discrepancy measures can be designed or different purposes. Since the likelihood functions are specifically designed to represent the sampling error in a particular data type, then, intuitively, they would make good discrepancy measures, but they are not commonly used.

I believe part of this has to do with historical use: for instance, we might use maximum likelihood to fit a model, but require a secondary procedure (e.g. chi-square or other test) to tell us whether the fit is very good or not. This may have led to a predilection against using likelihood-based quantities as discrepancy functions. But they certainly can (and have) been used as omnibus discrepancy functions. See for example Table 2 in Conn et al. (2018).

Conn, P.B., Johnson, D.S., Williams, P.J., Melin, S.R. and Hooten, M.B., 2018. A guide to Bayesian model checking for ecologists. Ecological Monographs, 88(4), pp.526-542.example Table 2 in Conn et al. (2018).

How do you calibrate p-values in either the Bayesian or frequentist approach?

Both approaches for calculating p-values (Bayesian and the frequentist approach of Besbeas & Morgan [2014]) examine the probability that observed data produce discrepancy statistics similar to those produced by data simulated from the model. However, in order to properly interpret a p-value as the probability of a type I error, they need to be uniformly distributed on (0,1). Because observed data are used twice in calculation of simulation-based p-values (once to fit the model, and once to calculate a tail probability), they have the unfortunate property that they often *do not* have a uniform(0,1) distribution. This suggests the possibility that one may need to calibrate Bayesian p-values (or

frequentist analogues) in order to properly interpret them. There are several possible procedure for doing this, which are conceptually similar to management strategy evaluations in fisheries stock assessment. We might, for instance, conduct a simulation study with the model we are using, and evaluate the distribution of p-values one might expect when data are simulated under the correct model (that is, when the data generating model and estimation model are the same). Besbeas & Morgan have done this using different discrepancy functions, showing that the Freeman-Tukey discrepancy better approximates a uniform distribution. Alternatively, had they estimated a density function for simulated p-values (with cdf F(p)), a calibrated p-value for a particular dataset may have generated as $F(p_{obs})$, where p_{obs} is a naïve p-value for the observed data. By my reading of the paper they don't actually do any explicit calibration, but it could be done. For further reading on calibration in a Bayesian setting, see Dey et al. (1998) and Hjort et al. (2006).

Dey, D.K., Gelfand, A.E., Swartz, T.B. and Vlachos, P.K., 1998. A simulation-intensive approach for checking hierarchical models. Test, 7(2), pp.325-346.

Hjort, N.L., Dahl, F.A. and Steinbakk, G.H., 2006. Post-processing posterior predictive p values. Journal of the American Statistical Association, 101(475), pp.1157-1174.

How do Bayesian or frequentist p-values work under weighted (or iteratively reweighted) likelihoods typical of fisheries stock assessment?

To my knowledge, this question is yet to be evaluated and is ripe for future research. I suspect performance of these approaches will likely vary between assessment models, between different data sets being modeled (e.g., fits to survey indices vs. fits to age or length compositions), and between different likelihood weighting schemes. Because of how much uncertainty there is, it would likely be advantageous to conduct simulation studies to examine typical ranges of p-values, and if necessary calibrate Bayesian p-values)

Process error (Random effects/state-space models) diagnostics

Can we test between having error on the state versus temporal variation on a particular process (e.g. M)?

Simple State-space models have a single error term on the state that relates to all processes (e.g. M, F, and recruitment) that change the state from one time period to the next and assume that the error is iid and the same for each age, except for the recruitment age. However, the error might be mainly from one of these processes or from several processes and theses processes may change over time. Therefore, the variance of the combined error may change over time and also among ages. More sophisticated state-space models model temporally correlated fishing mortality and/or variation in other processes. Diagnostic tests would be useful to determine if temporal variation in the state or if temporal variation should be modelled in the individual processes and whether the temporal variation is being modelled correctly.

Are there diagnostics that can identify when the assumptions about the state-error and its distribution are misspecified (e.g. should it be iid or AR1).

State-space models generally assume that the error is iid normal or autocorrelated, but the error may follow some other distribution and may vary by time or age. Standard residual diagnostics could be used on the residuals of the estimated state and the deterministic model prediction to evaluate violation of the distributional assumptions. These residuals could be plotted and analyzed by time or age to further evaluate the adequacy of the assumptions.

Are there diagnostics that can determine if modelling temporal variation in the parameters of functional forms is appropriate or if semi- or non- parametric should be used?

Temporal variation in selectivity is often modelled in Stock Synthesis by allowing the parameters of selectivity curves to vary over time using time blocks or temporal deviations. However, there are alternative approaches to model temporal variation in selectivity using semi- or non-parametric approaches. For example, selectivity could be modelled using a 2D (time and age) AR1 process. Standard time series diagnostics could be used for the non-parametric approaches. However, comparing residuals from the functional forms is more complicated since there are the temporal deviates in the parameters and also the temporal deviates in the resulting selectivity at age. Stock assessment applications typically focus more on the residuals of the fit to the composition data to diagnose misfit selectivity. An alternative approach might be the e-empirical selectivity diagnostic that compares the empirical selectivity (catch-at-age/length divided by predicted numbers at age/length) with the estimated selectivity, which could be modified to look at temporal variation in selectivity.

Diagnostics in applications (e.g. Stock Synthesis)

Is it possible to develop code to apply the more sophisticated diagnostics to all the possible models that can be developed in the general stock assessment programs.

Some of the diagnostics used for fishery stock assessment models are quite complicated (e.g. ASPM, retrospective analysis, hindcasting) making them more difficult to automate. However, there has already been substantial success in developing the code for r4ss and diags4ss to be applicable to a wide range of models that can be developed in SS.

Simulation testing

Is there a machine learning technique that can be used to generate a decision tree for an ensemble of diagnostics to identify model misspecification and the appropriate fixes?

Simulation analysis can be used to determine the results of different diagnostics under different model misspecifications. This can also be repeated with simultaneous multiple model misspecifications to represent what might occur in an actual application. The analysis will likely require simulating many scenarios and the evaluation of these scenarios to develop a set of rules to interpret the diagnostics and fix the model. Machine leaning techniques could be used to analyses the simulation results.

Should the simulated scenarios include combinations of different model misspecifications?

In real applications there may be multiple model misspecifications occurring at the same time. These multiple misspecifications may complicate the interpretation of the diagnostics. One model misspecifications may counter the effects of the other so that the diagnostics look OK or they may interact to make the diagnostic look like something else is misspecified. Therefore, simulations with multiple misspecications are needed to interpret the diagnostics.

Automation

Is it possible to automate the acceptance-rejection of models for use with large ensembles?

Evaluating all diagnostics in detail would be possible with the base case approach historically used in fisheries stock assessment, but may not be possible with the contemporary trend to use a large ensemble of models. With a large number of models, it would be beneficial to have automatic criteria to decide if a model passes the diagnostics. This would require developing a measure and a rejection level. If a large enough set of models were evaluated, including the correctly specified model, then it may automatically identify and fix model misspecification (i.e. reject all misspecified models and select the correctly specified model). There may be multiple models that pass the diagnostics tests and therefore the correctly specified model may not be determined, but a set of possible models might result that represent the uncertainty.

General

Is a quantitative criteria for passing or failing a diagnostic needed?

In a perfect world, the suit of diagnostic tests would be used to accept or reject models from an ensemble leaving the correctly specified model remaining. In the case that many model are tested, a quantitative criteria would be needed. A quantitative criteria would also be important to ensure that approach is more objective, transparent, and less ad hock.

Should diagnostics be used to eliminate models, weight models, or identify and fix model misspecification?

Traditionally, diagnostics be used to eliminate models that violate assumptions, but it is preferable to use them to identify and fix model misspecification. More recently, diagnostics have been used to weight models because all models violate the diagnostics to some degree and fit to data (e.g. AIC) is not appropriate to weight models. This is not a desirable situation and it is preferable to eliminate model misspecification, or at least have an ensemble of models that pass the diagnostics.

What diagnostics should be defaults in assessment reports?

It is useful to have a default set of diagnostics to include in assessment reports that can help reviewers and stake holders determine the quality of the stock assessments. Alternatively, the diagnostics could be

available in some other format (e.g. the r4ss html browser based system) that makes them easy to navigate. It's premature to provide a default set of diagnostics, but the set of residual plots in r4ss are a good start. Other default diagnostics might include the ASPM, R0 profile, empirical selectivity, retrospective analysis, and hindcasting.

Do good diagnostics results imply well estimate management quantities?

The diagnostics are not directly related to management quantities, but to the model itself. They can be used to evaluate components of the model or the model in general, but not the management quantities. We have to assume that if the model is a good model (i.e. is passes the diagnostics) then the management quantities are well estimated. Essentially we are extrapolating outside the range of the diagnostics based on the assumptions of the model. This can be viewed similar to using a von Bertalanffy growth model to extrapolate outside the range of the age-length data by assuming the model is correct.

How do you know you changed the right process when you address a diagnostic issue?

When there is a residual pattern (or other diagnostic issue) some changes will easily fix the problem (e.g. the aging error or temporal variation in some process), but it does not necessarily mean that you fixed the right thing. i.e. you may have solved the symptom but not the cause. So it is good to have a rationale for the change rather than just knowing that the change will fix the issue.