CCSBT-MP/0304/08

CSIRO
MARINE RESEARCH

Synthesising Performance of Candidate Management Procedures Across Different Operating Model Scenarios

Tom Polacheck
Dale Kolody

Prepared for the CCSBT 2nd Management Procedure Workshop
7-9 and 13-14 April 2003, Queenstown, New Zealand, CCSBT-MP/0304/08

Table of Contents

Abstract 1
Introduction 1
Primary Performance Indicators in the Synthesis Process 1
The Relativity of the Performance Metric for Different Scenarios 2
Approaches for Synthesising Results Across Operating Model Scenarios 3
  The Weighted Statistic Approach 3
  The Robust Performance Approach 5
Literature Cited 7

Abstract
Simulation testing of candidate management procedures for southern bluefin tuna (SBT) will be conducted across a range operating model scenarios. Problems and two possible approaches that might be considered in synthesizing results across the range of scenarios tested are discussed.

Introduction
The agreed process for the development of a management procedure for southern bluefin tuna (SBT) adopted by the CCSBT involves the simulation testing of the performance of candidate management procedures (Anon 2002a). The simulation testing is aimed at ensuring that a procedure will provide "adequate" performance across the range of uncertainty about the past and future dynamics of the SBT stock and fishery and associated available data. Simulation testing is to be conducted across a range of scenarios representing different aspects of the underlying uncertainty. Within any given scenario, in most cases it is relatively straightforward to compare the performance of different management procedures by comparing the various performance statistics. However, when results from a number of different scenario are being compared, a arriving at a synthesis that would enable relative rankings can be complex – particularly when large uncertainty exists about the underlying stock dynamics compounded by relatively short time horizons in a recovery stock situation. This paper presents a short discussion of some of the problems and possible approaches that might be considered in synthesizing results from across a range of scenarios.

Primary Performance Indicators in the Synthesis Process
At the 2002 CCSBT SAG meeting (Anon 2002b), a set of 8 performance indicators was selected as a minimum to be used in for evaluation of candidate management procedures. These can be grouped into four basic types:

1. Stock status ratios of biomass (e.g. SSB_{2007}/ SSB_{2002}; SSB_{2022}/ SSB_{2002}; $SSB_{2020}/ SSB_{1980;}$ NB_{2022}/ NB_{1992});
2. Total average catch (over a 5 or 20 year time horizon);
3. Stability of catches (defined as AAV) and
4. Distribution of catches among fisheries (defined as average ration of surface to total catch).

It is anticipated that these eight indicators will form the principle basis for the comparative performance of different candidate management procedures. In the current paper, the primary focus of discussion will be with respect to the trade-off in performance between the stock status and total average catch indicators. This is because it is across these two dimensions that there is the greatest potential flexibility and differences among procedures or variants within a single procedure.

Moreover, whatever are defined as "acceptable" range of behaviour with respect to the stability of catches can easily incorporated into any procedure by placing limits on the maximum percentage change in the TAC in any given year and/or by changing the interval between changes in the TAC. Thus, differences in the relative performance of candidate management procedures with respect to stability of catches is unlikely to be a primary factor in the overall evaluation of different procedures performance across a range of operating model scenarios.

Similarly, any candidate management procedure can be tuned to provide a different distribution of catches among fisheries. In addition, the Commission has provided no guidance on what might constitute an agreed appropriate metric for judging performance relative to the distribution of catches among fisheries. In the absence of any agreed metric, fine tuning the performance of any candidate management procedure with respect to the other performance indicators by varying the distribution on catches among fisheries will not be a very productive since similar fine tuning would be possible within any procedure. As such, we would suggest that the primary basis for comparing differences in performance will need to be, at least in the first instance, the stock status/catch trade-off with catch distributions among fisheries held constant. If there was a consensus within the Commission on an agreed metric by which the distribution of catches among fisheries should be evaluated then exploration of performance in this dimension could be explored. However, it is not anticipated that this would have a large effect on the relative performance of different rules in terms of stock status and catches as long as they were allowed to be tuned similarly with respect to the catch distribution among fisheries. Thus, unless large differences rules could in fact be demonstrated when candidate procedures were tuned similarly, performance with respect to the distribution of catches among fisheries is not likely to be the primary factor in attempting to provide an overall evaluation of candidate management procedures which synthesizes their performance across different operating model scenarios.

The Relativity of the Performance Metric for Different Scenarios

The initial set of operating models being used in the simulation testing of candidate management procedures for SBT cover a wide range of productivity levels for the stock (e.g. as reflected in the different steepness values in the stock recruitment curve and the different natural mortality levels). Those selected were meant to reflect the underlying certainty in this aspect of the dynamics of the SBT stock. While some refinement and possible reduction in the range may occur in the selection of the scenarios to be used in the final testing, the final set is still likely to encompass a wide range of scenarios. Given the depleted status of the SBT stock and the agreed rebuilding objective and target timeframe for the stock (i.e. the rebuilding of the spawning stock to the 1980 levels by the 2020), comparing the performance of candidate management procedures across the full range of scenarios in terms of the absolute value of the defined indicators becomes problematical.

Thus, for some of the higher productivity scenarios, the CCSBT rebuilding objective can be achieved in conjunction with substantial increases in the average catch over the next 20 years. In contrast for the lower productivity scenarios, the rebuilding objective cannot be achieved even if catches were to be reduced to zero. Thus, any management procedure would be judged as not providing acceptable performance if the 2020 rebuilding target were to be a minimum criteria to be achieved in all scenarios. Yet, it would still seem appropriate to expect a management procedure to achieve rebuilding for scenarios with high productivity. Similarly, unless catches on average are reduced below their current level, the spawning stock will be lower than its current level in low productivity scenarios, while any average reduction in catch will result in the spawning stock substantially over-shooting the rebuilding target. Thus, what is either “ideal” or “acceptable” performance in terms of the indicators that have been defined for a management procedure will be scenario dependent. In particular, consideration

needs to be given to what are realistic and acceptable performance criteria in the case of low productivity scenarios.

The need to consider different criteria in evaluating the performance indicators for different scenarios indicates that the mean or median (or similar measures of central tendency) of the defined performance indicators across the range of scenarios will not provide a very useful or interpretable measure of a management procedure. Where large differences in productivity exist, the feedback/learning characteristics of a procedure would seem to be the most important. In other words, the ideal procedure would be able to "learn" whether in fact the stock was or was not productive and respond appropriately to achieve "acceptable" rebuilding within the constraints of the productivity of a given scenario. In short, a highly desirable feature of a candidate management procedure would be one that on would appropriately increase catches on average in highly productive scenarios and reduce catches in low productive scenarios. Clearly, this is an ideal and realistic expectations on what is achievable need to be recognized (i.e. the last 50 years of historical data have not been sufficient to allow us to determine the productivity so the ability of a management procedure to resolve this in the next 20 years can be expected to be limited). Nevertheless, the extent to which procedures can distinguish or "learn" what is the relative productivity of the stock should govern the extent of sensitivity to the choice of scenarios using in the testing process. If a high degree of "learning" can be achieve, the evaluation and ultimate recommendations on candidate management procedures will not be highly dependent upon the specific set of scenarios chosen. If little or no learning can be achieved, the evaluation process will tend to highly dependent upon the choice of scenarios and the process used to synthesize the results across these scenarios.

Approaches for Synthesising Results Across Operating Model Scenarios

There are potentially a large number of approaches that could be used to attempt to synthesise performance results across different operating models to provide an overall basis for comparison and evaluation. Two general are discussed here.

The Weighted Statistic Approach

The weighted statistic approach is based upon calculating weighted summary statistics of the performance indicators for each candidate management being considered and using the weighted summary statistics as a basis for comparison and evaluation. The weights used in the calculation of the statistics are meant to reflect the relative overall "likelihood" or "plausibility" of that scenario. Ideally, the weight assigned to any specific scenario would reflect a combination of a prior probability about underlying model or hypothesis represented by that scenario and a measure of the fit or likelihood of the observed data given the specific scenario – an essence an informal, if not fully rigorous Bayesian approach. For example, one could calculate the weighted means and coefficient of variations for each of the performance statistics. Then the mean and approximate confidence intervals could be used to provide measures of average performance and risk of SSB or catches being below their current levels.
Advantages to this approach include that it provides a way to include a wide range of scenarios in the overall evaluation process, but ensures that the evaluation and recommendations with respect to the selection of a management procedure is not unduly influence by improbable or low plausibility scenarios. By examining the relationship between stock status and catch related performance indicators, the

approach can provide a direct measure of the trade-off in these two objectives, which can be used to evaluate how different candidate management procedures make this trade-off. The approach can also be used to provide a measure of "risk" that be used as the direct basis for selecting and tuning a management procedure (e.g. ensuring that there is less the a 5% probability that the spawning stock in twenty years will be less than the current level, etc.)

Despite the above advantages, there are a number of problems in the actual implementation of this approach that make its actual utilization problematical and make it difficult to arrive at consensus. These include:

- The assignment of weights
- The selection of the set of scenarios to include in the evaluation process
- The definition and interpretation of the summary statistics

Assignment of weights

The weight statistics and conclusions drawn from them are likely to be sensitive to the specific numerical values assigned to the weights, while the process of assigning weights can have a large subjective component and is not straightforward. Past experience in the CCSBT demonstrate the difficulty in developing transparent procedures for the assignment of weights and on achieving consensus on the actual weights to use in synthesising stock assessment results across a range of scenarios (which is analogous to the problem of assigning weights to different operating model scenarios). Problems can exist because differences in experience among the scientific committee can lead to different perceptions about the relative plausibility or prior belief to assign to different scenarios. In addition, it is difficult to develop joint priors for various dimensions of the uncertainty space being considered. Priors are normally set for individual parameters or dimensions of the uncertainty space and the overall prior is then determined as the product of the priors set for each combination of parameter values. Difficulties occur because various parameters are not independent and some joint combinations may be highly unlikely. In addition, there is often little or no basis for setting variance and structural related priors (e.g. the degree of smoothness in selectivity curves, the frequency that change in selectivity occur, etc). The value given to these types of priors can have a large effect on the operating model dynamics.

The procedure of assigning weights also needs to adequately account for how well any specific operating model scenario is consistent with the existing observed data. It is important to have some of updating on the input priors to ensure that the relative weights assigned to different scenario reflect the relative support in the data for them – in other words to ensure that scenarios with "poor" fits to the data are not considered highly plausible. To the extent that the objective function used for condition the scenarios is considered to constitute an appropriate likelihood function then the value of the objective function for any given scenario provides a natural weighting that can be used to update prior input weights. However, the components of the objective function that related to data used in conditioning the SBT operating model can only be considered as pseudo. In particular, the fit and differences in likelihood values for different scenarios are sensitive to the effective sample sizes assumed for the multinomial component for the "observed" length and age frequency data. In addition, treating the objective function as a true likelihood function for assigning relative plausibility is conditional on the distributional and structural

assumptions of the model. Lack of fit issues (e.g. trends in residuals) have been a concern in past SBT stock assessments. Such lack of fit issues are not usually properly accounted in the objective functions as these assume that the errors are independent and identically distributed. This can result in a scenario being assigned a relatively high weight based on the value of the objective function, although based on lack of fit considerations it might be considered relatively implausible.

Selection of the set of scenarios to include in the evaluation process

If a weighted statistical approach is used, the specific selection of scenarios becomes critical. Since the weighted results are meant to represent a quantitative summary of performance over the range of uncertainty having a balanced and fully cross set of scenarios becomes important, particularly when prior weights are assigned separately to each dimension of the uncertainty space considered. Otherwise the resulting weighted statistics that are calculated will not be representative of the overall uncertainty space, but will be biased towards that part of the space which is more fully represented. In addition, if the number of dimensions and actual range of the uncertainty space within each dimension is not limited to a relatively small numbers, the number of scenarios that need to be tested will be large to achieve a balance set. This limits the extent to which results from individuals scenarios can be examined and computing constraints may also be encountered.

The definition and interpretation of the summary statistics

The definition of appropriate and meaningful weighted statistics that provides a readily interpretable metric for comparing candidate decision rules is difficult. As discussed above, the interpretation of performance statistics (at least those the minimum set define for the initial stage SBT testing) need to be considered within the context of a particular scenario. For example, the weighted average of recovery ratios across scenarios does not necessarily provide a meaningful guide to overall performance. The distribution may be multi-modal for the scenarios being used or have a long tail for some scenarios. In which case, a weighted mean provides a poor guide as to what might be expected. Thus, a candidate procedure may obtain a value near 1.0 for the weighted mean ratio of. SSB in 2020 to that of 1980, but this may be achieved by having excessively high values for the highly productive scenarios and very low values for the low productive scenarios.

The Robust Performance Approach

The robust performance approach is based upon the concept that a primary goal of the testing procedure is to ensure that a management procedure will provide robust performance across the range of major uncertainties in the stock and fishery, particularly with respect to stock status indicators. In this approach, criteria for "adequate" or "acceptable" levels for critical performance indicators are developed in relationship to management specified objectives. For each scenario, candidate management procedures can then be evaluated as to whether they meet these defined levels. As discussed above, the defined levels of "adequate" or "acceptable" may need to be different depending on the productivity of the stock. Procedures that meet these levels across all scenarios could be considered robust, while ones that do not would be considered unacceptable. If several candidate procedures were judged to be "robust", then a secondary evaluation phase involving "best" performance in terms of catch could be used. In this case, determining the relative ranking of different procedures

within scenarios and comparison of these ranks may provide a relatively simple and robust comparison process.

One possible example of how this might work would be to define "acceptable" performance as being having a 95% chance of achieving the CCSBT rebuilding objectives for scenarios in which the rebuilding target could be achieved under a constant catch of 75% of the current level. For scenarios in which the rebuilding target could not be achieved under a constant catch scenario of 75%, might be defined to ensuring with 95% confidence that the spawning stock in 2O years is increasing and that it has increased by at least 20%. (Note these definitions of "acceptable" are merely meant as examples and would need to be based on discussion and agreement at the Scientific Committee and Commission level). Any management procedure that meets these criteria would be considered robust. Comparisons and recommendations among candidate management procedures that met these criteria could then be based on ranking of these procedures in terms of total catch within a scenario. If procedures had consistent ranks in terms of total catch, this secondary evaluation process would be straightforward. If this catch ranking of different procedures varied greatly among the different procedures, then additional criteria would need to be defined based (e.g. mean rank, the procedure which tended to have the lowest average minimum catch, etc).

This approach also is not without problems. However, the problems may be more tractable if in fact a primary focus for the management procedure is considered to be to ensure achieve robust (in contrast to optimal) performance in the face of uncertainty about the dynamics of the stock and the SBT fishery. Problems include:

- Defining criteria for robustness;
- Hard boundaries or procedures which just fall outside acceptable range for one scenario or a few scenarios;
- Selection of scenarios.

Defining criteria for robustness

Agreeing on some robustness criteria may be relatively straightforward. For example, that the spawning stock in 20 years should be at least as large as it is currently. Defining agreed minimal rebuilding targets in relationship to the potential productivity embedded in different scenarios entails some implicit, but non-quantifiable judgement on risk and requires guidance from management in terms of agreed objectives.

Hard boundaries

One problem with evaluation approaches that are based on "yes/no" criteria of acceptability is that the actual boundaries have an element of arbitrariness and it may not be desirable to reject candidate management procedure that just misses a boundary for one scenario (particularly if the procedure exhibits really good performance in general). However, based on the experience with testing management procedures for SBT so far (e.g. Polacheck et al 2003), this is not likely to be a major problem. This is because the performance all of decisions rules tested to date has a relatively smooth response in their performance in response to changing the value of a tuning parameter. Thus, it appears that any specific candidate management procedure that just misses having acceptable performance for one scenario could be retuned to meet the criteria

by adjusting the values of the tuning parameters without any large change in its overall general performance.

Selection of scenarios

The selection of scenarios will also be a critical issue with this approach. While the emphasis on robustness, means that it is not necessary to ensure that a complete and balanced set of scenarios are tested, there may be difficulty in the determination of what scenarios have a reasonable enough degree of plausibility that they should be included. For some uncertainties, this may not be a critical issue as candidate procedures will be robust and yield relatively similar performance over the full range of possible uncertainties. However, for others, the trade-offs involved to ensure robustness will be sensitive to the range of uncertainties considered plausible – particularly with respect to uncertainties about the general productivity of the stock (e.g. steepness, depensation, regime shifts, etc). Nevertheless, it may be easier to arrive on the range of uncertainty that have sufficient plausibility for which there is a need to ensure then to achieve consensus on the actually relative plausibility over the full range.

Literature Cited

Anon. 2002a. Report of the First Meeting of Management Procedure Workshop , 3-4 & 6-8 March, Tokyo, Japan. CCSBT.

Anon. 2002b. Report of the third Stock Assessment Group Meeting. 3-7 September, 2003. Canberra, Australia. CCSBT.

Polacheck, T., D. Ricard, P. Eveson, M. Basson and D. Kolody. 2003a. Results from Initial Testing of Some Candidate Management Procedures for Southern Bluefin Tuna. CCSBT-MP/0304/5.