SUGGESTED SEQUENTIAL STRUCTURE FOR SIMULATION TESTING OF ASSESSMENT METHODS

Doug Butterworth

Many past exercises involving simulation testing of assessment methods have, in my view, proved less useful than one might have hoped for the reason (again in my view) that they have been based on a range of idealised scenarios. Unsurprisingly there are seldom generic results which are valid across a wide range of circumstances, so that it can become difficult to use such results to specify the circumstances under which some result/advice applies, and hence to know whether it is pertinent to the assessment situation with which one might be faced with a particular stock.

The first and broad principle underlying the structure suggested below is that simulation testing should rather be based on actual situations, i.e. real resources and their associated data. Certainly then, at the end of a simulation testing exercise based upon a particular stock, one has results valid for that stock at least. But the further hope is that as examples of such testing exercises grow, a pattern of results will develop that will allow generic conclusions to be drawn, and hence then inferences made concerning (say) the best assessment approach to apply for yet another stock without again having to take that stock through this same simulation exercise.

Arising out of this principle there follows the concept that simulations should be "conditioned" on the situation believed to apply in respect of the stock concerned. This concept arises out of what has become standard practice in the simulation testing of Management Procedures (MPs) for setting catch limits for whale stocks as conducted in the IWC's Scientific Committee. MPs are intended to be robust against uncertainty, but there is no point in requiring robustness against uncertainties known not to apply for a particular stock. It is from this that the concept of "conditioning" of simulation tests arose: that the different population dynamics models used in testing MPs for a particular stock for robustness against uncertainty should all be required to be consistent with known information for that stock, e.g. a time series of past catches (assuming that to be well determined).

For MP testing, the IWC approach (for any particular model structure assumed for a stock, e.g. a specific value for natural mortality) is to fit a population model based on that structure to yield one specific plausible reflection of the true underlying dynamics for that stock. Given those fixed underlying dynamics, a series of pseudo datasets is then created by generating observations (abundance indices, catch-at-age values, etc.) of the same form and number as the real data, which could have arisen given those dynamics, with the distribution functions used to generate the errors/residuals for those pseudo datasets being as estimated in the original fit of the population model to the actual data.

For the IWC SC in an MP context, it is those pseudo datasets that are used as the basis to develop the simulation tests: for each trial the MP is tested against a range of alternative scenarios for the dynamics which have been obtained by fitting the population model for the model structure concerned (under the same time series of known historical catches) to each pseudo dataset in turn. Here it is suggested that **pseudo datasets generated in this same way** (conditioned on an estimate of the underlying situation of the stock concerned that is

provided by the fit of the original assessment model) **could provide a basis for simulation testing of assessment methods.**

One would not work with only one model structure and assessment procedure to generate such pseudo datasets. Clearly alternative structure/assessment method combinations can be used to estimate alternative underlying dynamics that would also constitute alternative plausible descriptions of the resource's situation. These too can be used to generate further sets of pseudo datasets in the same way. An investigation of assessment method performance should involve not only tests against pseudo datasets generated from the same structure and model, but also from defensible alternatives similarly consistent with the available information, as each could represent the actual underlying situation.

Observation error

In the context considered here, observation error refers to mechanisms that do not change the underlying stock trajectory. Thus, for example, a residual generated from a survey sampling error distribution about the value expected given the underlying true abundance reflects an observation error. In contrast, a mechanism that leads to a change in the population trajectory (or its age structure), such as an alternative deviation about the stock recruitment function, or a variation in the selectivity at age for the fishery which would modify the splits of historic catches into ages, is considered process error.

As a first step in this process of simulation testing of assessment models, it is suggested that pseudo datasets involve the addition of observation errors only when generating pseudo data. There are two reasons for this:

a) simplicity at the initial stage of a complex exercise; and
b) ease of the comparison exercise for estimates obtained when applying assessment methods to simulated pseudo datasets (developed from a particular structure/model combination); if these datasets include only observation error, there remains only one underlying true value for any quantity of interest (e.g. current resource biomass, or an F_{msy} TAC) against which to compare results from the different assessment methods.

Process error

In the IWC situation, where whale populations are generally assumed to have fairly slow and steady dynamics so that observation error dominates process error, process error has seldom been considered when generating MP trials because there has seemed to be no great need to include this. It is in any case problematic to do so, because if the underlying resource abundance differs from one pseudo dataset generation process to the next (e.g. through differing fluctuations in recruitment) upon what does one condition? For example, does one still condition on the historic catch series? However, if earlier recruitment in a particular year was lower than estimated under the original assessment based on the actual data, it may perhaps not even be possible to have taken that historic catch made that year without causing the extinction of the stock for the simulation in question. One could perhaps condition on the historic fishing mortality *F* rather than the historic catch each year, but then one is testing against scenarios that didn't actually happen and so can't actually reflect a possible reality, contrary to the conditioning concept.

The IWC SC has extended it MP testing process to include process error (essentially recruitment or natural mortality fluctuations) on two occasions, but for the whale population concerned their size was sufficiently small that the problem of extinction either did not arise, or arose so infrequently that the odd simulation where it did could simply be omitted without introducing more than negligible bias into the data generation process.

That, however, does not necessarily apply to typical fish stocks (except perhaps to long-lived ones, but there too recruitment may be highly sporadic), so a different approach is required if process error is to be introduced. The one suggested here could be applied whether the model includes random effects or is fully Bayesian. It requires effecting the integration concerned through an MCMC process, which creates equally likely scenarios (trajectories etc.), all of which are fully consistent with aspects of the real situation (such as historical catch series – though even for those one might wish to introduce the possibility of uncertainty which can be handled under this same approach), so that it respects the conditioning principle. The resultant pseudo dataset would then consist of n such MCMC realisations of the underlying dynamics, for each of which m realisations of observation error would be generated, giving nm pseudo datasets.

The difficulty that then arises is that the true value of certain quantities of interest (e.g. current abundance) will not be invariant across all such datasets, so that statistics measuring estimation performance will need to be based upon the differences between estimated and true values in circumstances where *both* vary from pseudo dataset to pseudo dataset. If these true values do not vary too much (say ~ 10%), that might not prove too problematic when it comes to interpreting results, but care may need to be taken in more extreme situations where that level of true variability is perhaps an order of magnitude greater.