CSIRO
MARINE RESEARCH

An approach for assessing the compatibility between a stock assessment and fishery independent indices of juvenile abundance

M Bravington
W N Venables
P Toscas

Prepared for the CCSBT 5th Meeting of the Stock Assessment Group (SAG5)
and the 9th Meeting of the Scientific Committee (SC9)
6-11 and 13-16 September 2004, Seogwipo KAL Hotel in Jeju, Korea

Contents

Abstract

Fishery-independent survey indices (aerial and acoustic) for juvenile SBT are available from 1993 to 2000. Because many of those cohorts have now recruited to the longline fisheries, it is becoming possible to examine whether the surveys and the stock assessment are giving consistent signals. This information is important when deciding how/whether to incorporate the surveys into assessment, and how much improvement there could be in assessment precision if more survey data becomes available in future. Because there are major uncertainties (i.e. wide confidence intervals) in the assessment itself, as well as in the survey indices, the "obvious" methods of comparison are likely to give misleading results. This paper suggests an appropriate statistical approach for making the comparison, and comments on the implications of different possible outcomes. We are currently coding and testing the approach, developing software which could be applied easily to either survey. Preliminary results for the aerial survey are expected by November 2004.

Introduction

If an aerial and/or acoustic survey could give reliable and reasonably precise estimates of relative abundance for young SBT (aged 1-4), then the survey(s) would be potentially very valuable for SBT management. But surveys are expensive, and the value to management has to be considered against the cost. Since there is now 10 years' worth of survey estimates, and there has been enough time for at least some of the cohorts to have entered the non-GAB fishery, it is time to move towards a systematic evaluation of the potential contribution of aerial/acoustic surveys to management.

The first question is whether the survey is giving a signal about recruitment that is consistent with the signal suggested by the rest of the stock assessment. There are reasonable *a priori* grounds for questioning whether the aerial and acoustic surveys will measure a consistent proportion of the juvenile stock over time— basically, there could be problems if a substantial and *varying* proportion of 1-4 year-olds spend their summers near South Africa rather than in the GAB. At the same time, though, the assessment itself is far from precise, and is also subject to various potential biasses. In this document, we outline a statistical approach that could be applied to examine consistency. We also outline some of the questions that would need further work, after the results of a consistency analysis are available. Note that an "inconsistent" verdict would not in itself necessarily mean that a survey was useless for assessment; we discuss this further at the end of the paper. We are currently coding and testing the approach, developing software which could be applied easily to either survey.

Preliminary results for the aerial survey are expected by November 2004.

Throughout, "survey" means either the acoustic or aerial survey (or any fishery-independent survey, in general). Where comments are specific to one type of survey, this is stated explicitly. Conceptually, the only difference between the aerial and acoustic surveys is that the acoustic survey only measures one age class whereas the aerial survey measures a mixture of three, so that a slight simplification is possible for the acoustic survey.

A statistical framework for consistency

We assume that in each year's survey season, there is some true number of fish of surveyable age in the whole ocean, but that this quantity can't be observed directly. The assessment (which is actually run some years later, but that's irrelevant) gives an estimate *with measurement error* of this true number. The AS gives an estimate of a possibly different number, also with measurement error. If the AS is really not consistent with the assessment (i.e. if the AS is measuring the wrong thing) then, when enough years of data have been collected, the set of differences between the AS estimates and the assessment estimates will be too big (i.e. will have too large a variance) to be explained by measurement errors alone. The discrepancy between the underlying quantity measured by the AS and the underlying quantity measured by the assessment can be called the "process error"; it could arise, for example, from inter-annual fluctuations in the proportion of juvenile fish choosing to go to the GAB. If the AS is fully consistent, then the true process error is zero.

By comparing the differences between point estimates with the known measurement errors, we can estimate the process error variance, and place confidence limits on it. A non-zero estimate would suggest some inconsistency; the implications are discussed later.

This framework can cope with both the aerial survey and the acoustic survey. Mathematical details are given in the Appendix. For clarity and simplicity, each survey series should be compared individually to the assessment, rather than trying a simultaneous three-way comparison between aerial, acoustic, and assessment. In the case of the aerial survey, the survey measures a mixture of three age classes in unknown proportions, so extra parameters need to be estimated for the relative availabilities by age in the GAB. The key issue is to get estimates of the measurement error (co)variances matrices across the years, Ω_S for the AS and Ω_R for the assessment. Note that these will in general be non-diagonal matrices, because of shared parameters such as catchability (assessment) or spotter ability (aerial survey) which affect estimates across several years. Approaches are discussed in the next section.

Before launching into the maths, it is important to say why it would be a *bad* idea to simply look at the correlation between the assessment point estimates and the AS point estimates. Correlation is simply addressing the wrong question: the right question is "are the differences in the point estimates too large to be explained by known sources of noise?". (Note also that it's not obvious how to do "correlations" for the aerial survey anyway, since the aerial survey gives an index which is a mix of three age classes in unknown proportions). It would also not be sensible to check whether, for example, a regression coefficient between the assessment and the survey is "statistically significant". For one thing, the implicit relationship considered by the significance test does not account for measurement error in the assessment; and for another, if there is little variability in recruitment over the time period examined, then the statistical power will inevitably be low and an insignificant result has little to tell us about consistency.

Getting covariance matrices for surveys and assessment

It is not important to get highly precise estimates of uncertainty for this consistency exercise, but it is important to avoid major biasses.

The aerial survey estimates are constructed from a likelihood-based statistical model, so standard techniques (inversion of the Hessian; delta method) can be used to estimate Ω_S. (This has already been done on a year-by-year basis to estimate the CVs.) Presumably something similar can be done for the acoustic survey.

The assessment is more difficult, and getting a reasonable estimate of Ω_R has turned out to be harder than developing the consistency check itself. For one thing, there are numerous versions of "the assessment" and we cannot hope to accommodate all the model uncertainty. For another, the objective functions used in SBT assessments do not correspond to well-founded statistical likelihoods, even approximately; the scaling of some terms is in doubt, and some random effects are treated as if they were parameters. In the long term, it is desirable to move towards a more statistical likelihood-based approach for the assessment, but this is easier said than done. For now, we effectively have to choose one particular version of the objective function, and work with that one.

The objective function can be thought of as the basis for a set of hopefully unbiassed estimating equations (namely the derivatives with respect to the unknown parameters and random effects) that lead to estimates of not just parameters but also random effects— including the time series of recruitments. There are a couple of techniques for extracting variance estimates from estimating equations in the non-random effect case, and we are investigating a couple of ways of extending to SBT. To get the basic uncertainty, we intend to try a Bayesian bootstrap (reweighted fit using random non-integer weights) using year-within-datasource as the resampling unit. This level of bootstrapping will capture some but not all sources of uncertainty— it should cope with "effective sample size" issues for length frequencies, for example, but probably not with an unmodelled long-term trend in selectivity. However, if we can demonstrate consistency *without* having included all sources of uncertainty that are present in the assessment (i.e. when we assume an unrealistically small amount of measurement error), then we can be sure that there would be no inconsistency if we were able to use the real, larger, measurement error.

There are some pitfalls when bootstrapping random-effect models; naive approaches can under-estimate some types of uncertainty. It seems to be necessary to combine some of the structural information about variances that is embedded in the nominal "Hessian", with empirical information about variances obtained from the bootstrap. We are currently using simulations of simplified stock assessment models to investigate several ways of making bootstrap-based corrections to parts of the "Hessian". A Bayesian-bootstrap-ready version of the SBT assessment has already been developed, so once we have settled on a bootstrap-correction approach, we will be able to get an estimate of Ω_R quite quickly.

Consistent or inconsistent: then what?

1. If the verdict is "consistent" (point estimate of process error variance = 0 or very small), then we could go on to examine how much precision would be added to an assessment which included a AS, and could evaluate the likely impacts of changing the AS design (e.g. to a cheaper but less precise design).

   Since we can never prove consistency— we can only disprove it— it might also be desirable to conduct a sensitivity analysis by calculating an upper confidence limit for process error variance, and proceeding as in step 3 below.

2. If the verdict is "inconsistent", then all is not necessarily lost. The next question is whether there is any suggestion that the discrepancies between the AS and the assessment are systematic rather than random— e.g. if the AS tended to flatten peaks and troughs in recruitment relative to the assessment. To examine this, we could look at "residuals" from the consistency model, and also consider elaborations of the model in which the relationship between the two series is not strictly proportional. If there are indications of systematically different results, then a fuller investigation (and hopefully more data) would be needed— note that there might be problems with the assessment as well as with the AS.

3. If it appears that there is an inconsistency (non-zero process error), but there is no indication of a systematic effect, then there may not be a problem. It is quite common in fisheries science for the "internal" estimate of uncertainty in a AS (calculated from the variability in the AS data alone) to be lower than the external estimate (calculated by looking at residuals from the survey in the context of a full assessment). Standard practice e.g. in ICES is simply to ignore the "internal" estimate and work only with the "external" estimate, calculated as part of the assessment. This is not really a good idea because it becomes impossible to evaluate the impact of changing the survey design. It is also statistically unnecessary. Instead, we could explicitly allow for survey measurement error as well as process error in a combined assessment, and investigate the gain in precision in much the same way as for the consistent case. Changes in the survey design would affect the measurement error variance but not the process error variance.

Appendix: A mathematical framework for the comparison

The basic statistical idea is to set up a something similar to a factor analysis, where the "latent variable" is true recruitment, but with an extra random effect to describe the proportion of young fish *available* to the survey in a given year. We estimate the variance of that random effect— if the point estimate is zero, then there's certainly no evidence of inconsistency.

In year t, the assessment delivers estimates $\hat{r}_{at}$ of "recruitment", i.e. numbers at age a, which are unbiassed estimates of true numbers-at-age N_{at}, the latter being unobservable. The $\hat{R}_{at}$ are random variables that are unbiassed but subject to measurement errors correlated across ages and years, so that

$$\begin{aligned} \mathbb{E}\left[\hat{R}_{at}\right] &= N_{at} \\ \mathbb{V}\left[\hat{R}|N\right] &= \Omega_R \end{aligned}$$

The survey delivers an estimate s_t which is a composite index across a set of ages $\mathbb{A}$. For the acoustic survey $\mathbb{A} = \{1\}$, while for the aerial survey $\mathbb{A} = \{2, 3, 4\}$. Here we deal with the more general case when more than one age is present, the single-age version being a simpler special case. We assume that

$$
\begin{aligned}
\mathbb{E}\left[S_t|N_t,\eta_t\right] &= \sum_{a\in\mathbb{A}} q_a N_{at} + \eta_t \\
\mathbb{V}\left[S|N_t,\eta\right] &= \Omega_S \\
\mathbb{V}\left[\eta\right] &= \sigma^2 I
\end{aligned}
$$

where η is process error and q_a describes the availability of age a fish to the survey; thus q combines both "sightability" and the average proportion of fish aged a that choose to go to the GAB. Ω_R and Ω_S are known non-diagonal positive definite matrices, and σ^2 and the q_a's are unknown parameters. Here, the process error is assumed to be an effect that varies from year to year, but that affects all cohorts equally; other formulations are possible, and are mentioned at the end.

The "data" we actually see are the $\hat{R}_{at}$ and the S_t (the $\hat{R}$ are not strictly data, but are acting as sufficient statistics for all the data that goes into the assessment process). If we treat $\hat{R}$ and S as Gaussian conditional on N, it is easy to write down an approximate conditional likelihood. This does not really help because the N_{at} are unobservable. However, a marginal likelihood which involves only the parameters q and σ^2 can be obtained by assuming that N has some vague prior distribution, and integrating across the possible values of N. If $\theta \overset{def}{=} \left\{q,\sigma^2\right\}$, we have

$$
f\left(s,\hat{r}|\theta\right) = \int f\left(s|n,\eta\right) f\left(\hat{r}|n\right) f\left(n\right) f\left(\eta\right) dn d\eta
$$

since S and $\hat{R}$ are conditionally independent given N. It is not obvious what form to use for the prior on N (for example: should the prior include the deterministic constraints imposed by mortality, given that the assessment already does this?) but as long as the assessment is providing far more information about N than the prior is, then the prior can be treated as constant for a first-order approximation.

We can then write

$$
\begin{aligned}
&f\left(s,\hat{r}|\theta\right) \approx K\left(\sigma^2\right)^{-T/2} \times \\
&\quad \int \exp\left\{-\frac{1}{2}\left(s-q'n-\eta\right)'\Omega_S^{-1}\left(s-q'n-\eta\right) - \frac{1}{2}\left(\hat{r}-n\right)'\Omega_R^{-1}\left(\hat{r}-n\right) - \frac{1}{2\sigma^2}\eta'\eta\right\} dn d\eta
\end{aligned}
$$

where a dash $'$ means transpose, T is the number of years of survey, the constant of proportionality K depends only on the known parameters Ω_S and Ω_R, and notation has been abused liberally throughout. The integral can be evaluated exactly by standard Gaussian manipulations. This provides a likelihood from which σ^2 and q can be estimated by maximization. The whole procedure is reasonably simple to program in R.

If enough data are available (which seems unlikely for now), different random-effects formulations could be considered, for example where there is a cohort-specific constant proportion that goes to the GAB in summer. This just changes the covariance structure of η. Models with non-proportional relationships can be accommodated by including some fixed-effect intercepts.