Resource Management Committee **ICES CM 2002/D:01**

REPORT OF THE

Working Group on Methods on Fish Stock Assessments

ICES Headquarters
3–7 December 2001

International Council for the Exploration of the Sea

Conseil International pour l'Exploration de la Mer

Palægade 2–4 DK–1261 Copenhagen K Denmark

TABLE OF CONTENTS

Section **Page**

1 INTRODUCTION

1.1 Participants

Frans van Beek	Netherlands
Noel Cadigan	Canada
José Castro	Spain
Chris Darby	UK (England & Wales)
Rafael Duarte	Portugal
Rob Fryer	UK (Scotland)
Kristin Guldbrandsen Frøysa	Norway
Holger Hovgaard	Denmark
Tore Jakobsen	Norway
Sigurdur Tor Jonsson	Iceland
Knut Korsbrekke	Norway
Stéphanie Mahévas	France
Benoit Mesnil	France
Coby Needle	UK (Scotland)
Carl O'Brien (Chair)	UK (England & Wales)
Dankert Skagen	Norway
Stuart Reeves	Denmark
Antonio Vázquez	Spain
Henrik Sparholt	ICES

1.2 Terms of reference

A **Working Group on Methods on Fish Stock Assessments** [WGMG] (Chair: C. O'Brien, UK) met at ICES Headquarters, Copenhagen, Denmark from 3–7 December 2001 to:

a. develop diagnostics and testing procedures for the evaluation of methods used for producing stock assessments, short-term forecasts and medium-term projections;

b. apply such testing procedures to the methods routinely used by ICES at present, such testing should pay particular attention to:

   i. bias detection and correction;
   ii. the form of error distributions in stock-recruit relationships taking into account input from SGPRISM;
   iii. other concerns that may be raised by ACFM from time to time based on input from assessment working groups;

c. identify strengths and weaknesses in the methods and propose modifications to assessment models or new models as appropriate;

d. use its diagnostic and testing procedures in order to evaluate the performance of new methodological proposals;

e. present its results in a form that can be readily implemented in the assessments, e.g. through the development of computer software.

WGMG will report for the attention of the Resource Management and Living Resources Committees and ACFM.

1.3 Scientific justification for re-establishing the working group

The Working Group on Methods of Fish Stock Assessment last met in February 1995 (ICES, 1995).

Prior to the present meeting in 2001, ICES has lacked an active forum for developing new methods and investigating properties of fish stock assessment methods. ACFM has discussed this problem and concluded that there is a strong need for regular meetings of the Working Group on Methods on Fish Stock Assessments. The Group shall work with the estimation and projection procedures in a statistical context. The Working Group will concentrate on methodological procedures and will not evaluate any case studies in any detail. However, much focus will be placed on inter-sessional work.

ACFM at its May 2000 meeting wanted to push forward on the quality assurance issues within ICES. These plans involve a Methods Working Group that can serve as the focal point for certifying assessment methods. The proposed

Methods Working Group will help ACFM in this respect by providing background studies for ACFM to aid its decisions on whether to certify a particular method or not.

The Methods Working Group should also, as part of its remit, serve as the ICES' focal point for the discussion of new methods. As a first priority, the Group shall evaluate methods used for producing medium-term projections. The Methods Working Group was set up jointly under the Resource Management Committee (RMC) and ACFM. Membership should involve, but not be restricted to, ACFM members.

ACFM has previously identified a number of immediate concerns for the Methods Working Group to address which are essentially of a methodological nature:

- medium-term projections, definitions and procedures (following on from the EC Concerted Action FAIR PL98-4231);

- assessment in data poor situations; e.g. CPUE reference points and dynamic pool models;

- time series methods (Gudmundsson, 1994);

- technical definitions of biological reference points;

- the mid-year projection problem; e.g. standard procedures;

- identification of bias in assessment and removal of bias from established assessment procedures. This point should be discussed under various headings such as retrospective analysis as a tool to reveal bias, implementation bias in the procedures (time tapers etc.);

- biological reference points for non-standard situations;

- weighting in tuning procedures of indices that do not cover the entire population; e.g. blue whiting (Spanish survey and commercial indices) and weighting in ICA;

- environment factors in assessments; and

- estimation of variance in the assessment procedures.

Within ICES, the prevailing inclination is to seek solutions through a sophistication of models and methods. However, this is unlikely to be sufficient unless similar efforts are made to improve the data used in stock assessments, both the total catch and calibration data sets. The definition of fleets for tuning purposes should be improved, and stricter criteria should be used to select the catch and effort data retained for each fleet (Mesnil WD1). Standardisation of fishing effort to account for vessel characteristics and fishing patterns in space and time should be performed much more systematically than has been the case in the past.

Issues of data quality and model mis-specification, together with advice and guidance on short- and medium-term prognoses are discussed more fully in the main body of this report in Sections 4 through 6.

ACFM will at each of its meetings prepare a document summarising the assessment problems identified by the ICES stock assessment working groups. Based on this compilation, ACFM will make a prioritised list of topics for consideration at a future meeting of the Methods Working Group. This list will be communicated to RMC as input to setting terms of reference (ToR) for the Methods Working Group.

1.4 Structure of the report

Sound assessment methods are a basic requirement for providing advice. However, many stock assessments carried out by ICES show a retrospective bias – either analytically or historically. In the provision of management advice by ACFM, this raises three immediate concerns:

1. **Stock assessments using XSA/ICA, for example, are found to be biased in estimates of F (fishing mortality) and N (population numbers)**

- Bias can be numerically corrected through the use of a bootstrap technique but this does not identify the source and cause of the bias. An investigation of this for XSA and ICA would be informative and may provide a numerical correction to be made. Such numerical studies might indicate the root cause of the problem.

- Models for catch numbers-at-age may introduce statistical bias through the way in which models are defined – log-transformations that are not bias-corrected after fitting when converting back to the non-transformed scale. Can models be defined and fitted so that estimation biases are not introduced?

2. Medium-term projections

- Given that an unbiased assessment of F and N can be obtained, how should one project stock status forward into the medium-term (up to 10 years ahead)?

3. Provision of software and quality control

- SGFADS (ICES, 1999) produced ideas on software standards and the ways in which software should be validated and the quality assured.

The Working Group decided to address these three main concerns during its first meeting but in addition, considered possible consequences of variable data quality on the reliability of stock assessments as this was considered to be equally as important.

The terms of reference (ToR) are addressed within the six main sections of the report. Specifically, ToR a) is addressed within Sections 5 and 6 of the report, ToR b) is addressed within Section 6, ToR c) is addressed within Sections 4-6, ToR d) is addressed in Section 5 and ToR e) is addressed in Section 7.

In Section 2, the background to the problem of bias in stock assessments is discussed in the context of the quality of the ACFM advice that has been provided. Section 3 reviews a number of the working documents and background papers. Section 4 addresses issues of data quality in the context of likely impacts on estimates of stock size; Section 5 addresses issues of model mis-specification and the design of future simulation experiments to investigate hypotheses; and Section 6 provides advice and guidance on short- and medium-term prognoses. In Section 7, the issue of software certification is discussed and current plans for the further development of stock assessment software tools by members of the working group. A compilation of the Working Group's recommendations from the main body of the report is provided in Section 8; together with details of further work needed to be undertaken.

2 BACKGROUND

Many stock assessments carried out by ICES show a retrospective bias, either analytically or historically. In most cases the bias appears to be over-estimation of spawning stock biomass (SSB) and under-estimation of fishing mortality (F). In practice, it will initially not be clear whether one is dealing in stock assessments either with bias or with variability.

In the past, once it has become clear that bias is present, the consequence has been that the advice given by ICES, has generally been overly optimistic. The levels of fishing mortality associated with total allowable catch (TAC) have been higher than the target fishing mortalities that were associated with those that formed the advice of the TAC.

ICES needs an inventory of the frequency and size of the bias in assessments, and also on the direction of the bias and the likely causes; e.g. model mis-specification, data quality. Action should be taken to estimate and document the size of the bias, either to remove the bias from the assessment (if possible) or to establish procedures that ensure that due account of the bias is taken in the prognoses and the ICES' advice.

2.1 Bias in stock assessments

Bias gives a wrong perception of the state of the stock in terms of exploitation and biomass and by implication, a wrong perception of the management action required to meet specific objectives.

2.1.1 Short-term predictions

Short-term predictions are an extension of the stock assessment fitted model and highly important in relation to the ICES' advice and the management decisions based upon TACs. Over-estimation of population numbers-at-age can lead

to the adoption of TACs that were too high in the short-term, and afterwards these TACs appeared to have been taken with a higher F than anticipated at the time of the decision of the TAC. The ICES TAC advice, therefore, when based on an assessment which over-estimates SSB and under-estimate F, has contributed to a further deterioration of some stocks. In the case of North Sea cod, for example, the recently advised reduction in F has been in accordance with the precautionary approach (PA), but when combined with the over-estimated population numbers-at-age has resulted in a TAC which did not reduce F. In some cases the choice of TAC may even have led to an increase in F.

In the short-term prediction for the TAC year, an assumption has to be made about the exploitation in the current year. The choice is between the assumption that a certain catch will be taken (often, the TAC that has been specified for that year) or that the exploitation will continue at a certain level of fishing mortality. Empirically, the assumption of a *status quo* F seems to be the most robust. However, this choice ignores the intended effect of any management measures taken to control the fishery in that year. Using an F constraint regularly results in (unrealistically) high predictions of catch in the intermediate year.

Using an F constraint thus may lower the extent of the over-estimation of the catch volume in the prediction year. Beneficial as this may be to the status of the stocks concerned, from a methodological point of view this is clearly not an ideal situation. Alternatively it can be assumed that a certain catch will be taken in the current year. In this case the fishing mortality is adjusted to match the expected catch. However, when the prediction is based on a biased assessment, this assumption leads to a further increase in the bias. For example, for a stock where SSB is systematically over-estimated and F is under-estimated, downward adjustment of F to match the expected catch leads to a larger over-estimate of the remaining stock and gives an overly optimistic expectation of future exploitation possibilities.

Presently bias is not taken into account in the ICES' advice, but it is mentioned as a relevant factor. How should managers deal with such a comment? In an attempt to take account of a bias in the North Sea plaice assessment, the EC proposed a TAC in 2002 which was 10% lower than the ICES' advice. The 10% was a guesstimate carried out by the Commission. There is a clear need to quantify the extent of the bias, so that advice can be interpreted in an unambiguous fashion.

2.1.2 Medium-term projections

Medium-term trajectories are being used by managers to make a choice between short-term options in relation to medium-term goals. Bias can affect the starting point of the forecasts.

Medium-term prognoses should be relatively stable between years if they are to be useful for management. However, in some cases medium-term prognoses carried out in subsequent years and using standard procedures, show considerable differences. This is difficult to understand and has been criticised by managers. These differences in medium-term prognoses can only be explained if there are substantial changes in the underlying data or in the basic assumptions; e.g. the form of the stock-recruitment relationship. Is it also possible that the results of the models used by ICES are simply very sensitive to some input parameters? In that case it is relevant to know to which parameters the models are most sensitive. Bias in the assessment model can contribute to variation in the medium-term projections between years.

ICES should give guidelines to stock assessment working groups on what data should be used for input to (short- and) medium-term prognoses (F-pattern, weight-at-age, maturity-at-age) and in Section 6 of this report WGMG provides initial guidance. It is important to realise that the medium-term should include the short-term as well!

The short-term part of the medium-term analysis should be consistent with the deterministic short-term predictions! In fact, the medium-term analysis could be used to quantify the uncertainty in the deterministic short-term forecast. It is important that ICES also gives guidelines on how to deal with specific recruitment patterns; e.g. in stocks such as sole, haddock, Norwegian spring spawners and horse mackerel, or the absence of patterns.

2.1.3 Precautionary approach (PA) reference points

PA reference points are defined in such a way that there is a low probability that limit reference points are exceeded. In fact they intend to take account of the uncertainty in the stock assessment. Managing a stock below $\mathbf{F}_{pa}$ should have a high probability that the fishing mortality is actually below $\mathbf{F}_{lim}$. Bias contributes to the uncertainty of the assessment, however it contributes in a certain direction. This has not been taken into account when setting the PA reference points.

2.2 Quality of ACFM advice

ICES' advice about TACs is based on keeping F below a certain level and keeping SSB surviving the *TAC year*, above a certain level, in accordance with the PA reference points. Consistency in the forecasts is therefore crucial for the quality of the ICES' advice, as is a reliable estimate of the uncertainty.

A paper presented to WGMG (Sparholt WA5) considers the consistency in the short-term forecasts made by ICES in the period 1988-2000, for six of the major fish stock assessments undertaken annually:

- North Sea cod;
- North Sea plaice;
- North Sea sole;
- North Sea herring;
- central Baltic (Sub-divisions 25-32) cod; and
- Northeast Arctic cod.

The method investigated focuses on consistency of the SSB forecasted to be surviving the *TAC year*, taking into account the catch actually taken in the *TAC year*. For each assessment year the forecast tables from the ACFM reports are considered; together with a new option which corresponds to the actual catch taken in the forecast year. Generally, the forecasts made by ICES have been rather imprecise, when measured as SSB surviving the *TAC year*. They have been wrong by a factor of two or more in 20 (3 too small by a factor of 2 and 17 too large by a factor of 2) out of 67 assessments for the six stocks considered. In three assessments, the SSB was over-estimated by a factor of about 4. A clear bias towards a tendency to over-estimate SSB is apparent in 5 of the six stocks. Full details of the analyses are reproduced in Appendix A of this report.

2.3 Conclusion

The precise reasons for the poor quality of the forecasts are generally not known but any error or bias in the assessment will tend to increase in the forecast. The advice may be based either on keeping F in the TAC year below the PA level or leaving a certain SSB after the TAC has been taken. The latter option is considerably more sensitive to errors in the assessment. The implications for the nature and quality of the TAC advice are a concern to both ACFM and WGMG.

Bias seems to be the over-riding problem in the provision of ICES' TAC advice. It affects both the short-term and the medium-term prognoses but there is no consensus about the reasons for the biases in current stock assessments.

At this first meeting of WGMG, the group has started on the task of addressing the problem of bias within the stock assessment process; including short-term predictions and medium-term projections.

3 APPLICATIONS AND INVESTIGATIONS

Following the studies by Sinclair *et al.* (1990), ICES stock assessment working groups have come to realise, in retrospect, that some assessments systematically over- or under-estimate recent stock sizes and exploitation rates In some stocks this has led to optimistic forecasts, with realised catches smaller than predicted. The response within ICES has usually been either to manipulate various options available in stock assessment programs or to remove CPUE calibration data for some fleets or ages in the assessment. It is the intention of this working group to explore ways of reducing the problem with methodological solutions that might provide *a posteriori* corrections to the estimates.

However, these efforts might divert attention away from the predominant cause of the problem, which lies in the disparity between the assessment model and the data used for calibration. For example, no account of changes in efficiency over time is taken into account for most instances where CPUE data are utilized. This is a long-standing problem in ICES which has been overly confident in its assumption of constant catchability.

A number of working documents and a selection of relevant background material were circulated prior to, and during, the meeting in order to address the retrospective problem. A full list of working documents and background material presented to the working group is given in Section 9 of this report. On the basis of presentations on these documents and papers, the working group decided to consider three main topic areas, each of which could result in retrospective patterns in the output from sequential stock assessments and forecasts:

- data quality;
- model (mis-) specification; and

– short- and medium-term prognoses.

These topics are reflected in each of the next three sections of the report. Section 4 addresses issues of data quality in the context of likely impacts on estimates of stock size; Section 5 addresses issues of model mis-specification and the design of future simulation experiments to investigate hypotheses; and Section 6 provides advice and guidance on short- and medium-term prognoses.

Working and background documents are cited as appropriate in the relevant sections of this report.

In addition, the working group considered that four working papers were significantly important that they have been reproduced in Appendices A through D.

4 DATA QUALITY

As intimated earlier in Section 2, retrospective discrepancies in estimates of population size and the fishing mortalities obtained in successive years are an unfortunate but common feature of several stock assessments. They generally arise because of a mis-match between the assumptions built-in the methods (e.g. constancy of catchability or constancy of natural mortality over time) and the properties of the data. Although each method may give rise to some sort of bias, comparative assessments of problematic stocks with different methods often indicate qualitatively similar retrospective patterns (e.g. Darby WU2). There are thus good reasons to suspect that data problems are involved in such instances.

The purpose of this section is to review which components of typical assessment data can give rise to retrospective discrepancies, and under likely conditions. Whenever possible, suggestions will be made as to how each cause can be detected and its effects reduced, if not eliminated - bearing in mind that multiple causes affecting the different data are usually involved in real fisheries cases (e.g. Hutton *et al*., 2001).

Only discrepancies in the estimates of current and past states are considered; further problems arising in catch forecasts due to the specification of input data are dealt with in Section 6.

4.1 Catch data

4.1.1 Introduction

Most of the stocks assessed by ICES rely on landings data rather than catch data. The landings are mostly taken from official statistics, but may be adjusted in assessments for various reasons. In most fisheries discarding takes place. Discard data are not included in official statistics, but for a few stocks estimates based on observer programmes are used for assessment purposes.

Other sources of mortality caused by the fishery are slipping (fish released from the gear while still in the sea and dead), ghost fishing (lost fishing gear like gillnets still catching fish) and escapement mortality (mortality of fish escaping through trawl meshes etc.). Slipping is known to have occurred in some pelagic fisheries and will, for assessments, have a similar effect as discarding. Escapement mortality has mostly been discussed in connection with mesh size regulations, and ghost fishing has not been considered to be an assessment problem. These sources of mortality are not discussed further in this report but are merely identified here as possibly occurring in some fisheries.

The errors in the catch data used in stock assessments can largely be ascribed to the following three sources:

- poor discard data,
- over- and under-reporting of landings, and
- quality of biological sampling.

4.1.2 Discards

Changes in discard practices will be influenced by management measures, but minimum landing size, minimum mesh size and by-catch regulations can also be important factors. In some areas (e.g. Iceland, Norway, Russia) discarding is prohibited, in other areas (EU) it is legal. In mixed fisheries, management by single stock TACs can lead to discarding of a species when the TAC for that species has been taken while the mixed fishery continues.

In most cases discarding is size dependent, i.e. the smallest, and usually lowest-priced, fish are discarded. The availability of the fish to the fishing fleet may also affect the discard rate and market mechanisms are usually involved. In some cases these can lead to discarding of larger fish (high-grading). They could also lead to discarding of the less valuable species in catches from mixed fisheries.

In order to be used in assessments, time series of discard estimates are required. Discarding is difficult and costly to quantify because data are generally collected through on-board sampling of fishing vessels. Discard levels may be estimated by designated observers and for some stocks comprehensive time series are available (Stratoudakis *et al*., 1999). There are examples of stocks where sampling programmes have revealed high discard levels (e.g. North Sea cod, Western Baltic cod), but the discard data are presently insufficient for assessment purposes. For many stocks discard sampling has been sporadic or non-existent.

In the ICES area, discarding is included in assessments for some haddock and whiting stocks (IV, VIa and VIIa), northern hake, megrim (VII an VIII) and Bay of Biscay sole. *The inclusion of discard catches is considered to reduce bias and to give more realistic values of fishing mortality and biomass for these stocks but also contributes to the noise in the data* (quotation from Section 1.3.1.2 of ICES, 2002). As these discards typically are small fish, the main effect is to raise the level of estimated recruitment.

All mortality not included in the assessment is a potential source of error and bias. Lack of discard data will mainly be a source of error in recruitment estimates, but will have some influence on the tuning of the assessment if the age groups subject to discarding are included in the tuning. This influence is, however, small if the discards are mainly from the recruiting year-class.

In general, it can be concluded that lack of data on discarding of recruits or pre-recruits will have little effect on the retrospective pattern, but could have a considerable impact on the projections. If discards comprise larger individuals, on the other hand, retrospective patterns may occur (Mohn, 1999).

4.1.3 Mis-reporting

Mis-reporting of catches may occur by area and by species. Mis-reporting by area (often implying mis-reporting by stock) usually occurs when the quota for a species is taken in one area but not in another. This type of mis-reporting is sometimes detectable and adjusted for in assessments when the evidence is considered sufficient.

Mis-reporting by species in mixed fisheries can occur when quota is taken for one of the species, but the evidence is usually not sufficient to warrant adjustment of catch figures.

Substantial *black landings* (i.e. unreported landings) are sometimes suspected, but there is usually only anecdotal evidence. North-East Arctic cod is one exception where estimated/assumed black landings were used in the assessment, but there have also been estimates included in assessments of stocks in the North Sea and other areas.

The levels of mis-reporting are probably closely linked to management measures. If quotas are not restrictive there would appear to be less incentive for mis-reporting. There are, however, documented cases where tax evasion has been the motive for black landings, but these are single incidents that do not reveal the size of the problem.

Because reduced TACs usually reflect a stock decline, mis-reporting may tend to increase when the stock declines and decrease if the stock should increase again. It is quite conceivable that this could lead to a retrospective pattern in the assessment which would differ between declining and increasing stock situations.

Mis-reporting in any form is a serious problem in assessments because it is generally very difficult to ascertain how large the problem is, and being illegal mis-reporting cannot be estimated by traditional scientific methods. Because the extent and pattern of mis-reporting is usually poorly known, it is also difficult to assess the potential impact on a given assessment. However, examples are presented in ICES (1997) and Shelton & Lilly (1998).

4.1.4 Biological sampling

For many stocks sampling is adequate. Biological sampling of catches mostly takes place on the landings. However, sampling is often inadequate and may well not be fully representative of the landings.

In some case there are problems with age reading. Errors in age reading can lead to errors and bias in the age composition of the catches (Reeves, 2001). Especially if age reading procedures differ between countries or have changed over time, the assessment could be severely affected (ICES, 2001d).

Biological sampling may also be used as a basis for separating the allocated catches from stocks or even species that are not reported separately in official catch statistics and again is a potential source of error in the estimated catch.

4.2 Tuning data

4.2.1 Survey series used for 'tuning'

Three types of surveys are used by ICES stock assessment working groups to tune VPA-type models. These are bottom trawl surveys, acoustic surveys and egg/larvae surveys. The two first types give information on stock numbers-at-age whilst the third type gives information on SSB; i.e. not age disaggregated.

Generally, surveys are assumed to give unbiased estimates of the portion of the stock available to the survey, but with large variation due to the relative small sample size.

Survey tuning series can give a retrospective pattern in assessments if:

- there is a trend over time in catchability;
- there is a sudden jump in catchability from one period to the next; and
- there is an extreme value in the time series, with either very low or very high catch rates.

Trends in catchability in surveys are always difficult to reveal because trends in residuals from VPA models can also be due to problems with the VPA stock estimates derived from other sources of information. A sudden jump in catchability and a single outlier can sometimes be seen in the residuals, although not before it is too late (in terms of obtaining assessments for advice), i.e. after some years when the VPA has converged.

The possible causes for features in surveys which give rise to changes in catchability are listed below.

Changes to surveys

Generally, there has been an improvement in standardisation of surveys over the years. Also instruments to monitor the geometry of the trawl while fishing, have been improved and allow for better and continuous monitoring of trawl performance. Although these improvements mainly lead to less variation in performance, it is possible that continuous monitoring of the survey gear can lead to an increase in catchability estimates' because there are fewer hauls where the gear is not working properly. However, in some circumstances improvements to surveys can lead to a decrease in catchability (e.g. due to a decrease in mesh size used).

For the North Sea IBTS survey it has been shown that there can be rather large differences in fishing power even for vessels that claim to follow the same procedures with respect to the gear used and the general survey procedure applied (Simmonds & Rivoirard, 2000).

It is important to evaluate whether a change in vessel might explain any retrospective pattern. Shifts and replacement of ships over time potentially introduce a risk for changes in fishing power. For instance the replacement of the R/V Eldjarn (with very low catch rates) with the R/V G.O. Sars (with close to average catch rates) may have produced an increase in q. R/V Eldjarn was used during 1983-1993, and accounted for about 10% of the total number of hauls. In general its fishing power has only been half of the average level and thus the replacement has meant an increase in IBTS index q of about 5%. However, compared to the size of the retrospective errors observed in the assessments, this is not likely to be the cause of the problem.

There is a potential for variations in catchability for acoustic surveys caused, for example, by change of vessels with different noise profiles and different sensitivity to weather conditions. Conducting acoustic surveys is relatively complicated due to advanced technical equipment and software. Acoustic surveys rely quite heavily on trawl samples to get a representative picture of size distribution. Changes in echosounder equipment and keel-mounted transducers have also been shown to produce sudden shifts in catchability.

Overall changes to survey design lead generally to a jump in q in the time series if conversion factors are not applied. Such conversion factors are, however, difficult to estimate with the required precision. Often when attempts are made, the factors obtained are so uncertain that no conversion is implemented. However, **assessment procedures could model surveys separately without the need to calculate correct conversion factors**. The intention behind this is that one should treat any survey after a significant change as a new survey.

Factors not controlled through survey design

There are additional sources of survey variability in addition to the sampling error. Such variability could be rather short-lived and would have to be treated as *noise* together with the sampling error. However, if the underlying cause of the variation persists over 2 years or more this could cause trends. Examples of such variability that could induce trends are:

1. Survey trawls are size selective. This is a potential source of a trend in q if the length at age has a trend over time. Large variations in growth are more typical for stocks in boreal systems with larger variations in environmental conditions. There are also indications that the state (condition) of the fish affects the catchability through varying swimming capacity (speed, endurance). Swimming speed and or endurance is also effected by variations in temperature (e.g. reduced swimming capacity in very cold water).

2. Many demersal and most pelagic species have vertical migrations following a diurnal pattern. This can lead to a lack of complete coverage and will, if the pattern of vertical migrations evolves over years, produce trends in survey indices. This is because a varying fraction of the stock is not available for the demersal trawl. Also a varying fraction of a stock in the "dead-zone" close to the bottom will affect acoustic survey indices (Aglen, 1994). Such trends in the catching efficiency of the trawl could also be related to overall changes in visibility (increased visibility increases escapement).

3. Another effect related to vertical positioning is the tilt angle of herring when compensating for lack of buoyancy. A trend in the distribution of tilt angles can lead to strong trends in acoustic indices unless corrected for with tilt angle dependent target strength (Huse & Ona, 1996).

4. Weather conditions are a known source for bias. Air bubbles caused by strong winds will effectively dampen echo signals. There is currently no standardised method of correcting for this. Increased wave heights will reduce the catching efficiency of a trawl. Several consecutive years with *worse* than normal weather conditions could cause a trend in the catchabilities of both acoustic and bottom trawl surveys.

5. For different reasons, some surveys have incomplete coverage of the spatial distribution of the stock. Changes in the geographical distribution of a stock relative to the survey area could lead to higher or lower survey indices. Changes in geographical distribution can be caused by changes in migratory patterns or by changes when commercial fleets are fishing more heavily in some areas (local depletion) or both. Changes that evolve over several years would then lead to trends in the survey indices.

6. The catching efficiency of a trawl is potentially density dependent. This has been indicated by Godø *et al.* (1999). They showed that single fish have a different behaviour relative to a sampling trawl than a group of fish entering the opening of the trawl (cod, haddock and plaice). This work shows at least that the catching efficiency will depend on the degree to which fish form schools or patches.

7. Surveys on early life stages of fish suffer from problems related to a very high and variable natural mortality in these early life stages. If there is a change over time in early life mortality it could lead to a retrospective pattern.

4.2.2 Commercial CPUE data

Catch-per-unit-effort (CPUE) data can, in principle, be a measure of abundance over the whole stock distribution. However, this may not be the case with commercial data that are measures of local densities in discrete places where fishermen have chosen to trawl. To a lesser extent, spatial heterogeneity may be a problem for surveys, a topic which was extensively discussed during the 1989 meeting of this working group (c.f. ICES, 1993a).

The issue of the use and misuse of commercial catch-and-effort (CPUE) data as indices of abundance has been extensively debated in the fisheries literature for decades. It is of relevance to stock assessments using both (tuned) age-based and surplus production methods. Previous meetings of this Working Group have repeatedly called attention to the

necessity of using appropriate measures of fishing effort to derive indices of abundance, such as in 1989 (ICES, 1993a, p. 174): *The quality of the results given by all tuning techniques depends directly on the quality of the relationships between fishing mortality and fishing effort, or (and often equivalently) between stock abundance and CPUE or survey indices. It is thus crucial to use a satisfactory definition of the fishing effort and the best possible indices of abundance. Such attempts are a necessary complement to the efforts developed throughout the years to get the most efficient tuning techniques*.

A potential reason for retrospective patterns that immediately comes to mind is a drift in catchability due to continuous gains in efficiency, which are an expected feature of a commercial activity. It is an obvious violation of the basic assumption of constant catchability imbedded in most tuning methods such as XSA, ICA, ADAPT, etc. The process whereby such trends, when not explicitly accounted for, lead to the type of retrospective patterns seen recently for many stocks (i.e. over-estimation of stock sizes and under-estimation of fishing mortality) is intuitively easy to understand: the catch rates, when based on nominal effort, indicate higher abundance than reality. Simulations using artificial data perturbed by introducing a trend in catchability and then analysed using XSA (ICES 1997, Darby WU1, Duarte WA6) or ICA (Skagen WS4) confirmed that this is a sufficient reason for the observed patterns of over-estimation of stock size. The simulations conducted by Mohn (1999) indicate that a similar pattern may also arise when there is an abrupt shift in catchability somewhere in the time series, rather than a steady trend. Changes in catchability may arise from changes in fishing strategies, whether spontaneous or management induced, as well as from technological improvements.

There are good reasons to suspect that this process is one of the major causes of the retrospective patterns obtained by many ICES stock assessment working groups: in general, the tuning data used by these groups are based on nominal effort and for rather broadly defined fleet categories, whereas it is very likely that the efficiency of the vessels within such fleets has evolved over the period considered in tuning (typically 10-20 years). In addition, fleet data may represent a significant part of the international catch which may induce correlations.

Past attempts at estimating an *efficiency creep* parameter within the tuning method have proved rather frustrating and have been discouraged by previous Methods Working Group. Even though methods have evolved since then, it is still likely that the information content of typical assessment data is insufficient or too noisy to provide reliable estimates of such a parameter internally. Mohn (1999) suggests a moving window approach to detect such trends but shows that his estimated relative q trend (ERQT) correction does not work well when changes in q are involved. **This Group therefore recommends that effort data be corrected for changes in efficiency by specific analyses prior to setting up the tuning data for assessment.**

A generally recommended approach to estimate efficiency parameters and standardise effort is to use Linear Models (ICES 1984, Hilborn & Walters 1992). The family of Generalized Linear Models is even preferable since it provides distributions that may be more appropriate for CPUE data (e.g. ICES 1992). Mahévas (WD2) gives an example of the use of generalized linear models for a fishery where changes in efficiency have taken place due to the gradual introduction of a new gear in the fleet category used for tuning. This work is based on individual vessels' CPUE per trip and per rectangle from log-books, complemented with interviews of skippers providing detailed information on their gear, equipment, strategy, etc. over the period. This analysis clearly shows that there are measurable differences in efficiency among the vessels, even though it proved difficult to associate them with any change in specific characteristics. Taking account of the vessels' strategy in each trip improved the fit of the model. The estimated efficiency parameters can be used to standardise fishing effort across vessels and across years, a process which is facilitated in the latter case if true indices of abundance are available.

It is appreciated that such analyses involve a significant amount of work and require retrieving detailed information by vessel and trip for a number of past years. A useful first step would be to reconsider the definition of the "fleets" used for tuning which currently include vessels that may be too heterogeneous. Ideally, only the catch and effort data from individual vessel-trips whose characteristics, gear and fishing patterns are well-known and have been consistent over the period should be considered in any tuning fleet. By the way, this would facilitate the detection and elimination of outliers. ***A priori*** **standardisation of effort with, for example, tonnage or engine power, but such an approach may not be appropriate in all cases and should be used only when justified on the basis of analyses** such as discussed above; i.e. when differences in efficiency are demonstrated to be significantly related to such factors.

Commercial tuning series used in ICES assessments are often based on national fleets that may not cover the whole stock area. This in itself is a problem, compounded when there are changes in the distribution of the population because the indices from the various fleets will exhibit conflicting signals. However, the conditions under which such changes may give rise to a retrospective pattern in assessments are unclear and may depend on whether the changes are permanent or transitory and on the weight each fleet receives in the tuning.

4.3 Natural mortality

Natural mortality, M, is usually assumed constant in the standard assessment models, ignoring its possible changes over time. It has been demonstrated that changes in M with time cause a retrospective pattern in sequential population analysis when not accounted for (Mohn, 1999). Using a fixed value at the beginning of the series and an M decreasing with increasing age, biomass appears to be over-estimated in the terminal year of each retrospective data window, while the average F is considerably under-estimated. However, in some cases it is possible to detect and remove these trends by correction equations (Mohn, 1999).

Analyses of simulated data were carried out with separable models (ICA-like models), where a larger simulated value than the base-line M led to an under-estimate of F and an over-estimate of SSB (Skagen WS4). In the case where an increasing trend with time in the true M is simulated, the F and the SSB in the last year approach the true values. Thus, in this case the retrospective bias is actually a bias in the past, while the last year's assessment is correct.

Multi-species interactions caused by fish predator-prey relationships are an important component of natural mortality. Most often the predation mortality is found for the youngest age group and unaccounted trends in predation mortality may therefore bias the estimates of the younger ages. However, for long lived species the effect on SSB or reference F will usually be limited. Large scale disease outbreaks (e.g *Ichtyophonus* affecting herring) or environmental changes (e.g. cold winters affecting sole survival) may lead to significant mortality changes. However, such changes are expected to occur *at random,* not likely producing consistent trends.

4.4 Conclusions

Retrospective analyses are regularly carried out by stock assessment working groups and are used, together with Quality Control Diagrams, to trace problems in assessments. Retrospective discrepancies have an adverse impact on the quality and credibility of ICES' advice, notably when stock abundance is systematically over-estimated which implies that exploitation rates are under-estimated. It is thus imperative that the cause of such discrepancies be identified and redressed. The review in this section indicates that several data-related problems may be involved, but failures of the tuning data, whether from surveys or from commercial data, to comply with the constant catchability assumption of the prevailing methods and incomplete accounting of catches are important causes of the type of retrospective discrepancies observed recently. For the former cause, thorough analyses of the abundance indices should be conducted routinely and appropriate correction factors estimated prior to use in tuning. Problems with survey data are usually due to changes to the survey protocol that are not corrected for. Most factors can be controlled by scientists conducting surveys and, if this is properly done, surveys will usually produce far more consistent results than CPUE series.

A message to assessment working groups is that they should favour fewer data of good quality (as evaluated independently of the assessment model) instead of large quantities of data of unknown properties.

However, **retrospective patterns should not be taken as the only diagnostic of problems in assessments. Consideration should also be given to all other assessment diagnostics**. Consistently wrong estimates of historical stock size may be obtained when a fraction of the catch is not accounted for or when catches and CPUE series have correlated mis-reporting (Darby WU1). This point is further addressed in Section 5.3.1.

4.5 A final word

Hyperstability is one of the best and worst features of a fishery. For the fishermen, it means not suffering decreases in CPUE as abundance changes. For the manager, it offers the terror of a stock declining without any change in CPUE to tell him trouble is afoot. Some of the major fishery collapses in the world have been ascribed to hyperstability. (quotation from Hilborn & Walters, 1992, p.188). Recent problems with several ICES' stock assessments indicate that our ability to monitor the state of those stocks and to give timely warnings (and advice) is deficient, with the consequences sounding like some of the sad words in this quotation. The future of some fisheries is now endangered, as is the credibility of ICES' advice. There is thus an urgent need to redress this situation.

The prevailing inclination is to seek solutions through a sophistication of models and methods. However, this is likely to lead nowhere unless similar efforts are made to improve the data used in assessment models, and notably the effort data used to derive tuning indices. **The definition of fleets for tuning purposes should be improved, and stricter criteria should be used to select the catch and effort data retained for each fleet.** Standardisation of fishing effort with account of vessel characteristics and fishing patterns in space and time should be performed much more systematically than has been done so far. ICES may now be paying for its lack of interest in effort data and their standardisation compared to what is done elsewhere in the world!

5 NUMERICAL AND STATISTICAL ASPECTS OF FISHERIES MODELS

5.1 Introduction

The retrospective problem has been recognised as widespread and serious. The reasons why this problem appears are not fully known, as described in more detail in Section 2 and investigated from a data perspective in Section 4. There is a general understanding that trends in tuning index catchability, when used in models that assume constancy, can cause this effect. Several working documents presented to the meeting (Reeves & Hovgård WA1; van Beek & Pastoors WA4; Sparholt WA5; Duarte *et al*. WA6; Mesnil WD1; Darby WS2; Skagen WS4; Darby WU1; Darby WU2), as well as recently published studies in the literature (e.g. Mohn, 1999), point in this direction. It has been clearly demonstrated that the problem is more complex, however, and that e.g. trends or shifts in natural mortality, discards and mis-reporting, mis-specification of selection and catchability at age can contribute to the problem, sometimes in a quite complex way. There are indications that attempting to estimate more parameters than the information in the available data allows for, may produce deviations from the true state of the stock that may persist for several years, mimicking assessment bias. There may also be cases where the present estimate of the stock trajectory is biased, whilst those in the past may have been right (ICES, 1997; Mohn, 1999; Skagen WS4; Darby WU1)

The working group recognises a need for rather extensive studies to understand more clearly the kinds of stock dynamics, data structures and model specifications that produce retrospective patterns, how they can be diagnosed and possible remedies. In particular, **WGMG recognised the need to:**

- **understand the mechanisms which create inconsistencies in the perception of the development of the stock**
- **investigate the sensitivity to various causes of the retrospective bias of different model formulations, and the extent to which such causes can be accounted for in model structures**
- **develop diagnostics, and to understand the extent to which problems can be revealed by the diagnostic tools.**

In order to achieve these aims, the working group concludes that it is necessary to do extensive studies on data sets with known properties where both the true state of the stock and the flaws in the data are fully known. Such data sets are needed to study possible causes and mechanisms. There is also a need to know the true state of the stock when exploring diagnostics, to identify cases where problems are not detected and cases where the diagnostics indicate problems that are not real. It may be useful to study selected real cases in addition, but given the complexity of the problem, it seems unlikely that all aspects of retrospective patterns can be revealed by studying cases where the true state of the stock is unknown. Many such data sets have been published previously, but they are generally made for specific purposes and do not fully cover the needs identified here. Therefore, in Section 5.2 the working group has outlined specifications for software to generate simulated data with known properties, which allows for a wide range of properties and problems, both in the population and in the data derived from the population. Section 5.2.2 presents results from the use of a simple simulation to investigate the estimation bias inherent in XSA.

The working group also considered the development of diagnostics which may be useful in recognising various problems associated with the retrospective problem, but which may also be more generally useful in evaluating assessments. Some possible approaches are discussed in Section 5.3.

5.2 Simulated assessment data

It would be useful to have a computer generator for stock assessment data to be available for modelling purposes, as identified in the previous Section 5.1. This is useful to facilitate the evaluation of causes of the retrospective problems previously discussed.

In the following Section 5.2.1, the working group has provided a specification for computer software to generate simulated data with known properties that could allow for the investigation of a wide range of problems, both in the population and in the data derived from the assumed population model. Section 5.2.2 presents results from the use of a simple simulation to investigate the estimation bias inherent in XSA. This is merely illustrative and there is no *a priori* reason to suppose that XSA is any worse than other stock assessment methods based upon catch-at-age data. All methods need to be investigated. The EU Concerted Action FAIR PL98-4231 *Evaluation and Comparison of Methods for Estimating Uncertainty in Harvesting Fish from Natural Populations* (Patterson *et al*., 2000) showed that all VPA-based methods had similar levels of estimation bias and could be grouped accordingly. All non-linear models, and VPA calibration models are non-linear, suffer from estimation bias but there are established statistical techniques to adjust inferences for such bias. This is further explored in Section 5.2.3.

5.2.1 Specification of a data generator

Ideally, the data generator would have a separate module for generating a simulated population and another for deriving assessment input data from the population. Each set of data generated must be precisely documented so that it can be reproduced unambiguously.

The simulated population would include:

- a matrix of population numbers-at-age
- a matrix of weights-at-age
- a matrix of maturity-at-age

From the simulated population, the following input data for an assessment are derived:

- catch numbers-at-age
- discard numbers-at-age
- survey indices-at-age
- CPUE-at-age for commercial catches (catch numbers-at-age per effort measure)
- weights-at-age
- maturity-at-age

Year and age vectors

All input age-year matrices should be defined by a combination of year vectors and age vectors. The way in which calculations are made is specified below. It should be possible to add stochastic terms both to the functions defining the population and to the derived assessment data. The seed for the random number generator must be input, to ensure that even a stochastic data set can be reproduced exactly.

Each year or age vector (Y_y or A_a for short) is in principle defined as a function of time/age and of the state of the stock. Each function should be programed as a separate subroutine and there should be an assembly of relevant functions. The format of these functions should be standardized as far as possible. Typically, the following parameters will be needed:

- general shape: uniform, increasing, decreasing, step, top-hat, logistic, ogive, SSB related, weight-at-age related, von Bertalanffy, etc.
- parameters of the shape function, when needed
- random effects, type of: none, normal, log-normal
- parameter of random effects, when needed
- autocorrelation: parameter of the autocorrelation

Model parameters

The following parameters will be needed to define a set of data – a) general parameters, b) parameters that define the population, c) parameters to derive tuning data, and d) parameters that modify data derived from the population.

a) general parameters:

- initial year
- last year
- initial age
- last age
- plus group or not – YES/NO (default YES)
- number of CPUE series – n (default 1)
- number of survey index series – n (default 1)
- years between spawning and recruitment
- seed for random numbers – a number

b) parameters that define the population:

- population numbers in the initial year in numbers at age – an age vector

- annual recruitment in numbers at the initial age – a year vector
- fishing mortality – In general, fishing mortality is the product of a year effect and an age effect. Both should be allowed to vary with time. Natural mortality – product of an age vector and year vector, as for fishing mortality
- mean weight at age – product of an age vector and year vector; the age vector can be the von Bertalanffy function and the year vector a parameter in this function.
- maturity ogive – product of an age vector and year vector.
- distributions and parameters for stochastic terms

c) parameters to derive tuning data:

- survey catchability (n) – product of an age vector and year vector
- effort – a year vector
- distributions and parameters for stochastic terms

d) parameters that modify data derived from the population:

- discards (proportion discarded) – as for fishing mortality
- distributions and parameters for stochastic terms

The assessment input data may be truncated to a specified precision level.

Calculations

The order in which the program proceeds could be as follows:

1 – Input parameters for the run.
2 – Natural mortality calculated: $M_{(a,t)} = Y_t \times A_a$
3 – Fishing mortality calculated: $F_{(a,t)} = Y_t \times A_a$
4 – For each year, from initial year to last year:

- mean weight at age calculated
- number at initial age calculated
- numbers at age calculated: $N_{(a+1,\, t+1)} = N_{(a,t)} \, EXP(-M_{(a,t)} - F_{(a,t)})$
- SSB calculated

5 – Catches at age calculated
6 – Survey CPUE (n) calculated
7 – Noise: catch at age and CPUE results are modified with noise according to the specific parameters.
8 – Truncation

Implementation

The data simulation program should be implemented in such a way that it is easy to define properties, and to add alternative functions as necessary. This implies that a strict modular design is necessary. The output should be to files that can be used directly as input to assessment programs, like the 'Lowestoft format'.

Many simulated data sets have been published previously, but they are generally made for specific purposes and do not fully cover the needs identified by WGMG. For example, Patterson *et al*. (2000) performed a series of analytical experiments on sequential population assessment and forecasting models. One of the experiments examined the bias in estimates of stock size and exploitation levels using simulated data. Darby (WU2) has extended their study using simulated data in an attempt to examine the estimation bias inherent in XSA. The details of the simulated data and the experiments conducted are presented in the next Section 5.2.2. in an attempt to illustrate the utility of simulation studies to the investigation of problems in fisheries.

5.2.2 Preliminary results of a simulation study into XSA estimation bias

Darby (WU2) has analysed the magnitude of the estimation bias in the estimates of population abundance and exploitation estimated by the Extended Survivors Analysis (XSA) algorithm (Shepherd, 1999). The study extends work

carried out within the EU Concerted Action FAIR PL98-4231 *Evaluation and Comparison of Methods for Estimating Uncertainty in Harvesting Fish from Natural Populations* (Patterson *et al*., 2000).

Patterson *et al*. (2000) performed a series of analytical experiments on sequential population assessment and forecasting models. One of the experiments examined the bias in estimates of stock size and exploitation levels using simulated data. The simulation study used exact catch at age data and one survey with a controlled level of uncertainty in the catch –per-unit-effort estimates. The results of the study indicated that when using non-linear minimisation of VPA-based methods, there was an over-estimation bias in the estimates of SSB and the TAC forecast but estimates of $F_{0.1}$ were unbiased.

The EU study examined one scenario for the level of uncertainty in the catch-per-unit-effort data, the coefficients of variation are listed in the text table below.

Ages	1, 2	3, 4	5, 6	7, 8	9, 10	11, 12	13 - 15
C.V	0.75	0.7	0.6	0.5	0.6	0.7	0.75

Darby (WU2) has extended the study, examining the estimation bias inherent within XSA at a range of levels of uncertainty. The dynamics of the simulated population are described in Restrepo *et al*. (2000, BU3).

Observations used for model fitting

A single fleet exploits the stock with a fixed selection pattern over time.

S_a = 0.05, 0.1, 0.3, 0.7, 0.9, for a=1 to 5 and S_a=1.0 for older ages,

Catchability was held constant and there were no stochastic components to the fleet dynamics. For all years and ages, survey CPUE were generated for the start of the year. The indices had log-normal errors with a constant CV over all ages at the levels 0.01, 0.1, 0.3, 0.5, 0.7 and 1.0. The simulated population model was used to generate 1000 replicate CPUE data sets.

One catch-at-age data set, without measurement error, was generated with ages 1-15 and no a plus group.

XSA model structure

Three XSA assessment model structures were fitted to each of the replicate data sets. Each model formulation was based on the structure of the underlying simulated data. The first estimated all cohort terminal population numbers with the only model constraint being that catchability at age 15 was equal to that at age 14. The second and third model structures introduced two commonly used constraints that allow a reduction in the number of parameters estimated by the model, namely F shrinkage and a catchability *plateau* at a younger age. Within the F shrinkage model, the terminal population estimates at the oldest ages were derived from the average fishing mortality estimated for the five younger ages in the same year. This resulted in an XSA structure which is very similar to that of the most commonly used ADAPT (Gavaris, 1988) formulation.

Performance statistics

Although the individual assessment realizations generated a large number of outputs, the analysis of results was limited to the following statistics:

- population numbers at each age in the final year,
- SSB at start of year 26,
- $F_{0.1}$ (calculated using the estimated selectivity for year 25), and
- TAC in year 26 corresponding to the estimated $F_{0.1}$.

Computation of the TAC in year 26 required a projection of recruits, which was fixed at zero for simplicity.

Results

The estimated bias in the results of the replicated assessments is summarised in Tables 5.2.2.1 – 5.2.2.3 and Figures 5.2.2.1 – 5.2.2.3. Bias in the assessment estimates is calculated as

Proportional bias = (XSA estimate – true value) / true value

Within each series of summary tables the bias in the assessment estimates is presented for each level of uncertainty. Table (a) presents the proportional bias in the arithmetic mean of the XSA estimates of final year population numbers at age. Table (b) presents the proportional bias in the arithmetic mean of the SSB and $F_{0.1}$ estimates for year 25 and the TAC forecast for year 26. Tables (c) and (d) present the proportional bias in the median of the XSA estimates.

Figures 5.2.2.1 – 5.2.2.3 illustrate the bias in the average and median of the estimates. Figure (a) presents the proportional bias in the average of the XSA estimates at age; Figure (b) the proportional bias in the average of the management metrics. Figures (c) and (d) present the proportional bias in the median of the XSA estimated values.

The tables and figures illustrate that XSA has an over estimation bias in both population numbers at age, SSB and the forecast TAC at $F_{0.1}$. The over-estimation bias increases with the magnitude of the coefficient of variation of the CPUE data, linearly for $F_{0.1}$, non-linearly for population numbers at age, SSB and the forecast TAC.

Figure 5.2.2.1, the model structure in which all of the cohort terminal populations are estimated, illustrates that the non-linear increase in the expected value bias of the population estimates at age is greater for the youngest and the oldest assessment ages. At the youngest ages there have been fewer observations taken from the cohort for estimation of the abundance. At the oldest ages, there is greater uncertainty due to the reduced level of convergence within the VPA equations. The expected value bias is less than 10% for all ages except age 1 with CV up to 50%. At 70% CV's the bias is generally less than 15%. The median of the expected values is biased to a lesser extent with values less than 10% across all of the CV range.

Figure 5.2.2.2 illustrates that if the constraint of constant exploitation pattern at the oldest ages (F shrinkage) is applied, the convergence of the VPA at the centre of the age range adds structure that substantially reduces the bias at the oldest ages. Only at the youngest ages is the bias large. The bias at ages 4 – 15 is less than 10% even at a CV of 100%. The estimates of $F_{0.1}$ are unbiased, a finding which is consistent with the EU study. Relative to the CV of the tuning data the median of the estimated values can be considered to be un-biased.

Figure 5.2.2.3 shows that a similar reduction in the uncertainty at the oldest ages of the assessment can be achieved by constraining catchability to be constant across the oldest assessment ages. However, the degree to which the bias is reduced is not as great as that achieved by the F constraint.

The reduction in bias is conditional on the inside knowledge of the structure within the model generating the data sets. If an inappropriate constraint had been applied, such as a dome shaped exploitation pattern the bias could have been increased.

Discussion

Cadigan & Myers (2001) and the EU concerted action FAIR PL98-4231 have examined the bias in population numbers at age and fishing mortality derived from simulated data and assessment models. Both studies showed that there is an over-estimation bias in the population abundance at age and SSB. Cadigan & Myers (2001) examined the sensitivity to the assumed error distributions (log-normal and gamma), establishing that the bias was present for both error distribution assumptions and more pronounced under the assumption of log-normal errors. The same authors found that as the magnitude of noise in the CPUE data increases the bias increases.

The simulation study has established that XSA estimation bias increases non-linearly with the coefficient of variation of the CPUE series. Without constraining assumptions, the bias is less than 10% for the majority of ages with CV values below 50%, from 50 – 70% below 15%.

This analysis is considered preliminary because only one level of fishing mortality and stock trend has been examined. Further work will be carried out to examine the influence of variation in these factors. In addition the simulation is based on a single fleet CPUE series.

Given the conditional nature of the results of the analysis, an evaluation of the possible level of bias within assessments can be made. Figure 5.2.2.4a presents the distribution of standard errors of log catchability for each age in 60 fleet or survey calibration series from a sample of 14 assessments. The standard error of the log catchability can be taken as an approximation to the CV of the series. For this distribution 87% of the values lie below a CV of 70%. Figure 5.2.2.4b presents the distribution of standard errors with age. As would be expected, the higher standard errors occur at the youngest and oldest ages of the range. Figure 5.2.2.4c presents the expected value XSA estimation bias plotted against

simulated fleet CV, overlain with the frequency distribution of the assessment estimates of the CPUE series log standard errors. The mode of the distribution would indicate that if the calibration series were used to fit XSA assessments as single series, with no model constraints, the level of bias expected would be less than 5% for all metrics. In most assessments inverse variance weighting of the CPUE series will down weight the contribution of noisy series and should contribute to a reduction of their impact on the bias in the assessment results.

Patterson *et al*. (2000) concluded that estimation bias can be adjusted for by the use of the percentile bias adjustment method (Efron and Tibishrani, 1993). The simulation study will be extended in order to examine the potential for this approach to correct for the XSA estimation bias using this method. In the next Section 5.2.3 there is a discussion of two potentially important causes of bias – model mis-specification, and estimation of parameters in non-linear models.

(a) Expected

	AGE													
CV	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.001	-0.001	-0.001	-0.001
0.1	0.005	0.005	0.003	0.002	0.004	-0.001	0.002	0.002	0.000	0.001	-0.001	0.000	-0.002	0.002
0.3	0.045	0.034	0.023	0.016	0.023	0.006	0.016	0.017	0.014	0.018	0.014	0.024	0.019	0.034
0.5	0.121	0.089	0.058	0.038	0.054	0.024	0.043	0.047	0.038	0.049	0.045	0.063	0.060	0.079
0.7	0.232	0.164	0.108	0.078	0.098	0.059	0.088	0.095	0.085	0.108	0.111	0.145	0.157	0.194
1.0	0.477	0.338	0.232	0.204	0.225	0.166	0.252	0.261	0.285	0.365	0.419	0.503	0.650	0.778

(b)

CV	SSB	TAC	F0.1	Mean Pop
0.01	0.00	0.00	0.00	0.00
0.1	0.00	0.00	0.00	0.00
0.3	0.02	0.02	0.00	0.02
0.5	0.04	0.04	0.01	0.06
0.7	0.09	0.08	0.02	0.12
1.0	0.22	0.24	0.03	0.37

(c) Median

	AGE													
CV	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-0.001	-0.001	-0.001	-0.001
0.1	-0.001	0.001	-0.001	0.000	0.003	-0.004	0.001	0.001	-0.002	-0.002	-0.004	-0.001	-0.007	-0.001
0.3	0.000	0.004	0.001	0.000	0.014	-0.008	0.005	0.010	0.003	0.003	-0.004	0.008	-0.012	0.006
0.5	-0.001	0.017	0.004	0.004	0.027	-0.011	0.007	0.020	0.007	0.017	0.004	0.016	-0.018	0.002
0.7	0.004	0.028	0.012	0.014	0.041	-0.012	0.014	0.034	0.028	0.045	0.020	0.040	-0.001	0.028
1.0	0.018	0.038	0.030	0.033	0.058	0.000	0.038	0.062	0.061	0.073	0.062	0.075	0.037	0.065

(d)

CV	SSB	TAC	F0.1	Mean Pop
0.01	0.00	0.00	0.00	0.00
0.1	0.00	0.00	0.00	0.00
0.3	0.01	0.01	0.00	0.00
0.5	0.03	0.02	0.00	0.01
0.7	0.05	0.05	0.00	0.02
1.0	0.10	0.12	-0.02	0.05

Table 5.2.2.1 The results from a simulation analysis of the bias in estimates of population numbers at age, SSB and F0.1 in the final year of an assessment and a TAC for the year after the final data year at F0.1. The XSA assessment model was specified with no F shrinkage and catchability at the oldest age equal to that at the penultimate age.

(a) Expected

	AGE													
CV	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.3	0.04	0.03	0.02	0.01	0.02	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00
0.5	0.11	0.08	0.05	0.03	0.04	0.01	0.03	0.03	0.02	0.02	0.01	0.02	0.01	0.02
0.7	0.21	0.14	0.09	0.06	0.07	0.03	0.05	0.05	0.03	0.04	0.03	0.04	0.02	0.04
1.0	0.40	0.26	0.16	0.10	0.12	0.07	0.09	0.09	0.06	0.07	0.06	0.07	0.05	0.08

(b)

CV	SSB	TAC	F0.1	Mean Pop
0.01	0.00	0.00	0.00	0.00
0.1	0.00	0.00	0.00	0.00
0.3	0.01	0.01	0.01	0.01
0.5	0.02	0.02	0.01	0.03
0.7	0.05	0.05	0.02	0.07
1.0	0.08	0.08	0.04	0.12

(c) Median

	AGE													
CV	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	-0.01	0.00	-0.01	0.00
0.3	0.00	0.00	0.00	-0.01	0.01	-0.01	0.00	0.00	-0.01	-0.01	-0.02	0.00	-0.02	-0.01
0.5	0.00	0.01	0.00	-0.01	0.02	-0.03	0.00	0.01	-0.01	0.00	-0.03	0.00	-0.03	-0.02
0.7	0.00	0.01	0.00	-0.01	0.02	-0.04	0.00	0.01	-0.02	0.00	-0.04	-0.01	-0.04	-0.03
1.0	0.00	0.01	0.00	-0.01	0.03	-0.05	0.00	0.01	-0.02	0.00	-0.05	-0.01	-0.05	-0.04

(d)

CV	SSB	TAC	F0.1	Mean Pop
0.01	0.000	0.000	0.000	0.00
0.1	-0.001	-0.001	0.003	0.00
0.3	0.002	0.006	0.010	-0.01
0.5	0.014	0.018	0.015	-0.01
0.7	0.028	0.036	0.018	-0.01
1.0	0.053	0.064	0.019	-0.01

Table 5.2.2.2 The results from a simulation analysis of the bias in estimates of population numbers at age, SSB and F0.1 in the final year of an assessment and a TAC for the year after the final data year at F0.1. The XSA assessment model was specified with F shrinkage for the oldest assessment age and catchability at the oldest age equal to that at the penultimate age.

(a) Expected

	AGE													
CV	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.1	0.004	0.005	0.003	0.001	0.004	-0.002	0.001	0.000	-0.001	0.000	-0.003	-0.001	-0.003	0.001
0.3	0.043	0.032	0.021	0.014	0.021	0.004	0.013	0.013	0.009	0.011	0.006	0.013	0.007	0.019
0.5	0.118	0.085	0.054	0.034	0.049	0.019	0.036	0.039	0.028	0.037	0.030	0.043	0.032	0.051
0.7	0.223	0.153	0.097	0.066	0.084	0.045	0.068	0.072	0.058	0.072	0.064	0.084	0.070	0.100
1.0	0.424	0.275	0.174	0.124	0.146	0.096	0.125	0.133	0.114	0.137	0.130	0.160	0.145	0.198

(b)

CV	SSB	TAC	F0.1	Mean Pop
0.01	0.000	0.000	0.000	0.00
0.1	0.000	0.000	0.001	0.00
0.3	0.011	0.011	0.004	0.02
0.5	0.034	0.032	0.013	0.05
0.7	0.066	0.059	0.029	0.09
1.0	0.126	0.067	0.061	0.17

(c) Median

	AGE													
CV	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0.01	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.1	0.000	-0.001	0.000	-0.001	0.004	-0.005	0.000	0.001	-0.002	-0.003	-0.006	-0.003	-0.003	0.000
0.3	0.000	0.002	0.004	-0.002	0.011	-0.010	0.004	0.005	-0.007	-0.004	-0.014	-0.002	-0.004	0.005
0.5	0.000	0.009	0.007	-0.002	0.018	-0.026	0.009	0.016	-0.007	-0.005	-0.016	0.006	-0.002	-0.001
0.7	0.002	0.013	0.013	0.001	0.026	-0.031	0.016	0.023	-0.003	-0.002	-0.005	0.010	0.001	-0.003
1.0	0.005	0.021	0.023	0.007	0.042	-0.030	0.029	0.039	0.004	0.010	0.000	0.017	0.012	0.006

(d)

CV	SSB	TAC	F0.1	Mean Pop
0.01	0.000	0.000	0.000	0.00
0.1	-0.002	0.000	0.000	0.00
0.3	0.005	0.007	0.000	0.00
0.5	0.015	0.027	0.001	0.00
0.7	0.035	0.050	0.007	0.00
1.0	0.075	0.046	0.009	0.01

Table 5.2.2.3 The results from a simulation analysis of the bias in estimates of population numbers at age, SSB and F0.1 in the final year of an assessment and a TAC for the year after the final data year at F0.1. The XSA assessment model was specified with no shrinkage and q constrained for ages greater than age 6 at the value for age 6.

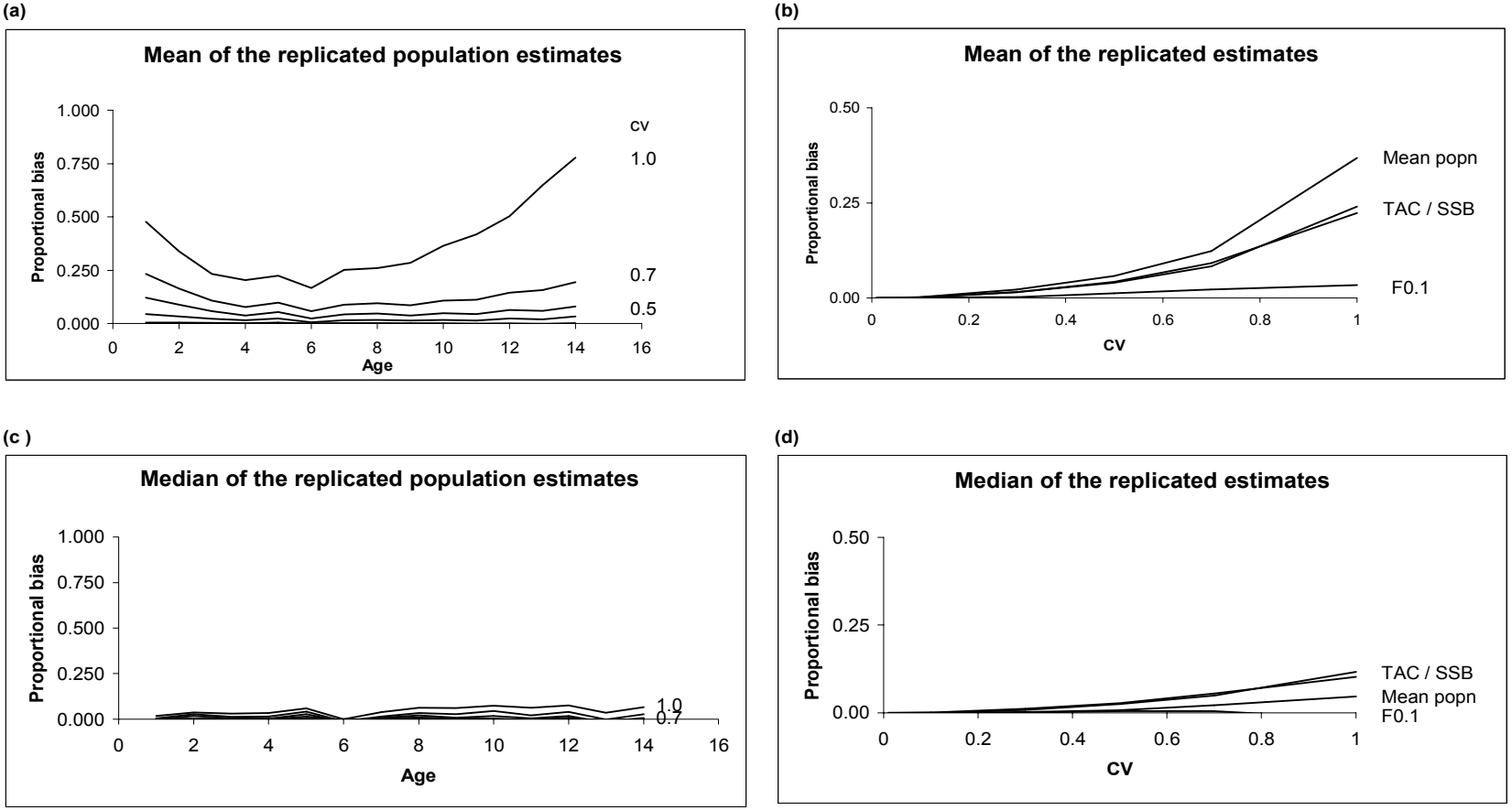


Figure 5.2.2.1 The results from a simulation analysis of the bias in estimates of population numbers at age, SSB and F0.1 in the final year of an assessment and a TAC at F0.1 for the year after the final data year. The XSA assessment model was specified with no F shrinkage and catchability at the oldest age equal to that at the penultimate age. (a) The bias in the estimate of expected value of population numbers at age at increasing coefficients of variation in simulated CPUE calibration data. (b) The bias in the estimates of the expected value of the TAC, SSB, F0.1 and average population at age bias. (c) The bias in the median value of estimated population numbers at age. (d) The bias in the estimates of the estimated value of the TAC, SSB, F0.1 and average population at age bias.

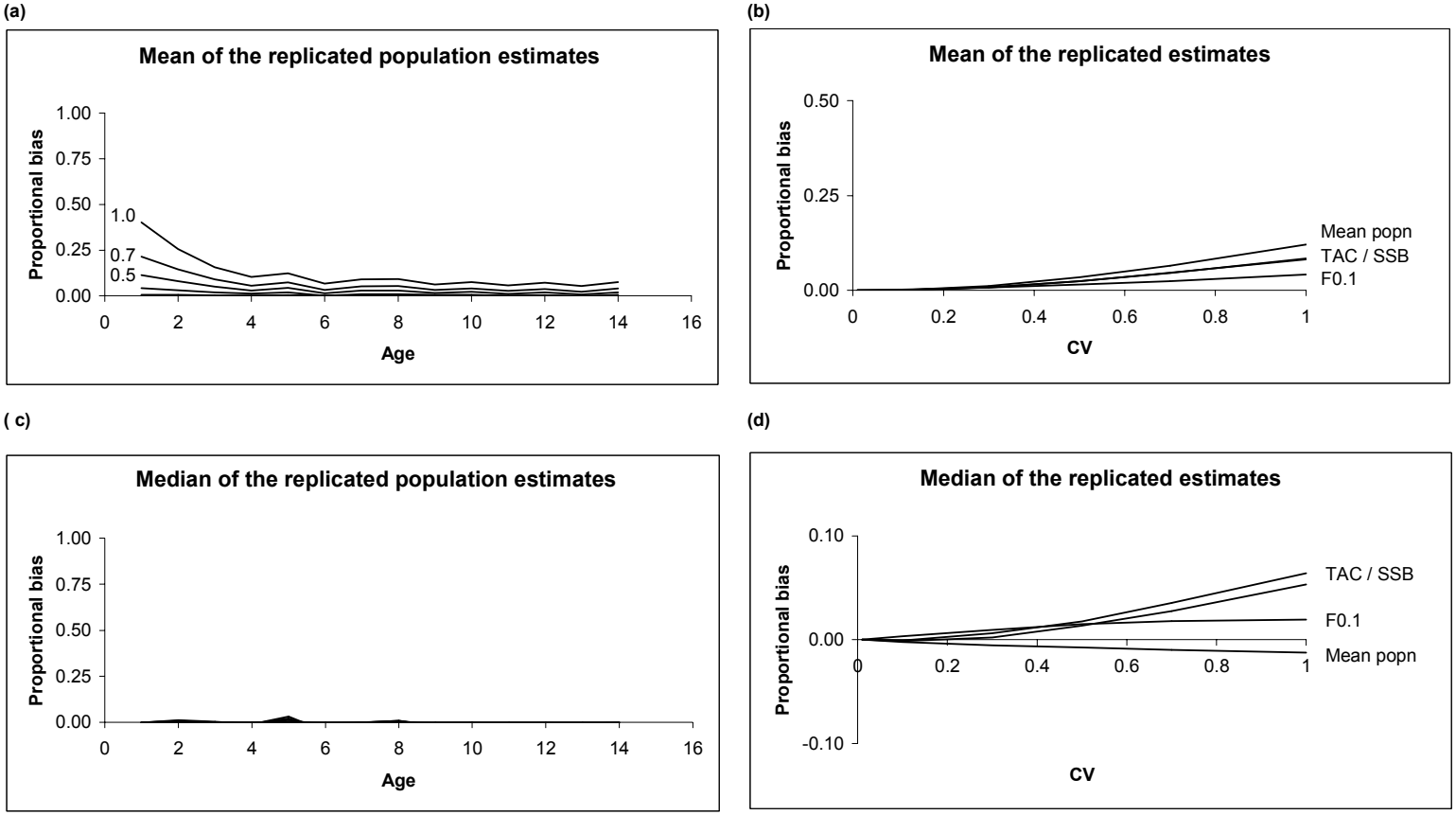


Figure 5.2.2.2 The results from a simulation analysis of the bias in the estimates of population numbers at age, SSB, F0.1 for the final year of an assessment and a TAC at F0.1 for the year after the final data year. The XSA assessment model was specified with F shrinkage for the oldest assessment age and catchability at the oldest age equal to that at the penultimate age. (a) The bias in the estimate of expected value of population numbers at age at increasing coefficients of variation in simulated CPUE calibration data. (b) The bias in the estimates of the expected value of the TAC, SSB, F0.1 and average population at age bias. (c) The bias in the median value of estimated population numbers at age. (d) The bias in the estimates of the estimated value of the TAC, SSB, F0.1 and average population at age bias.

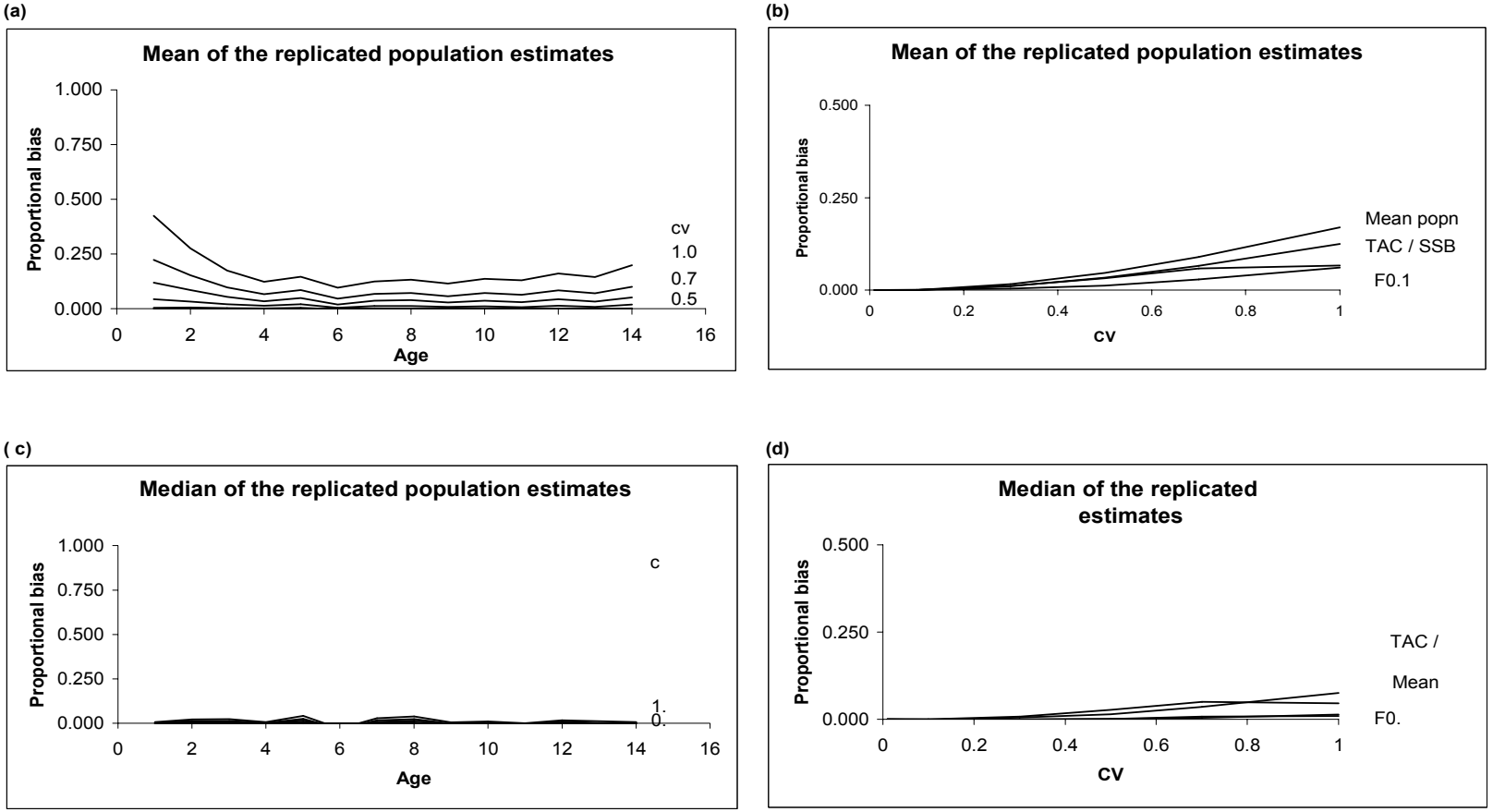
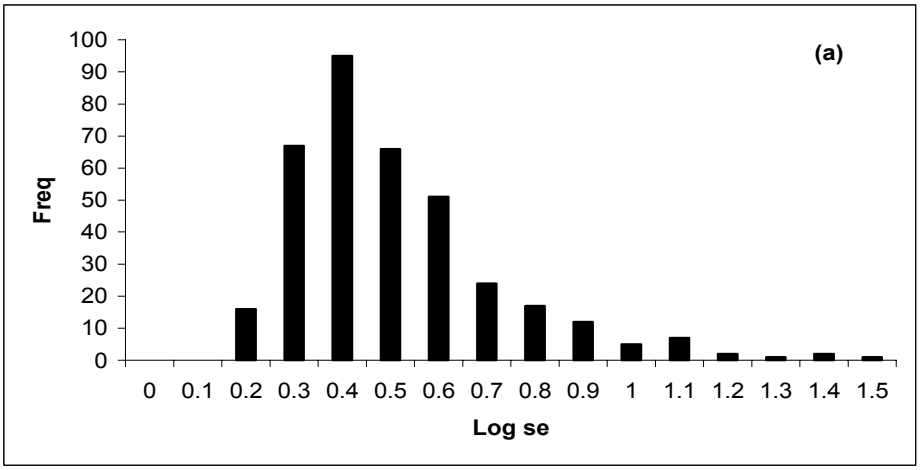
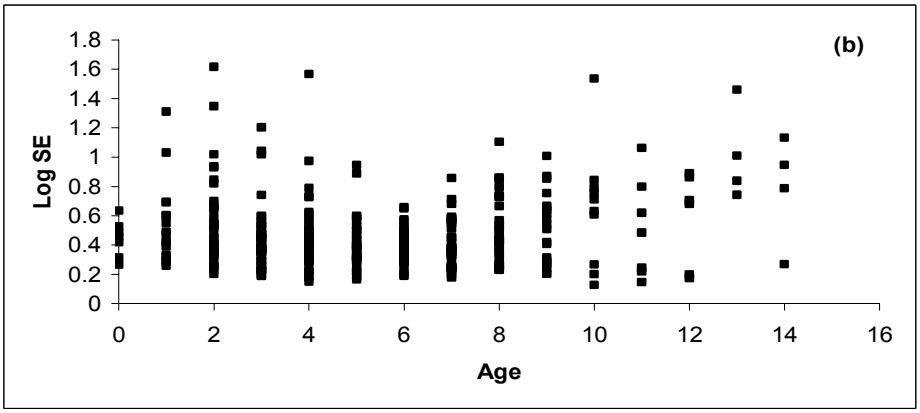


Figure 5.2.2.3 The results from a simulation analysis of the bias in estimates of population numbers at age, SSB and F0.1 in the final year of an assessment and a TAC at F0.1 for the year after the final data year. The XSA assessment model was specified with no F shrinkage and q constrained for ages greater than age 6 at the value for age 6. (a) The bias in the estimate of expected value of population numbers at age at increasing coefficients of variation in simulated CPUE calibration data. (b) The bias in the estimates of the expected value of the TAC, SSB, F0.1 and average population at age bias. (c) The bias in the median value of estimated population numbers at age. (d) The bias in the estimates of the estimated value of the TAC, SSB, F0.1 and average population at age bias.





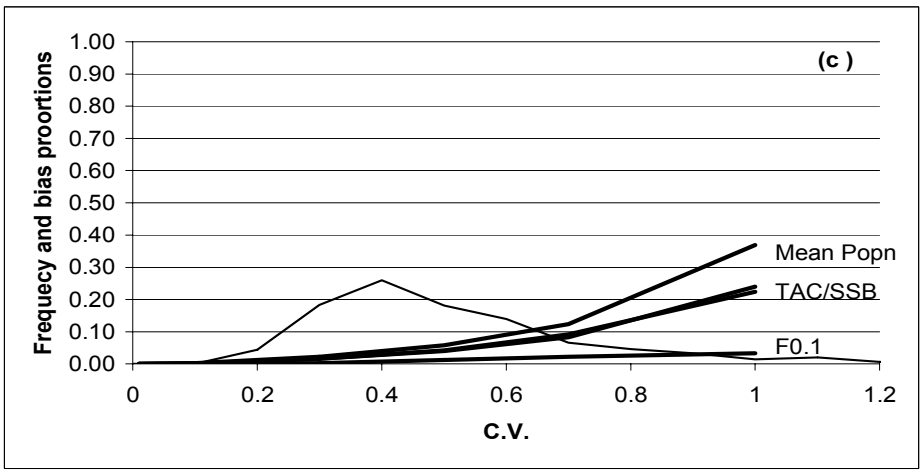


Figure 5.2.2.4

a) The frequency distribution of the estimates of standard errors of log catchability at age taken from 60 fleets fitted within 14 assessments.
b) The distribution of the estimates of the standard errors of log catchability with age.
c) The distribution from 5.2.2.4a plotted with the estimated bias in the XSA assessment estimates of population numbers, $F_{0.1}$, SSB and TAC in the following year.

5.2.3 A general comment on bias

ToR a) refers to: ... *evaluation of methods used for producing stock assessments, short-term forecasts and medium-term projections*. One result of the EU Concerted Action FAIR project PL98-4231 was recognition that it is useful to distinguish the components of a *method*; i.e. structural and distributional model assumptions versus the approach for making statistical inferences. It is most informative to know if differences between *methods* were caused by divergent structural and distributional model assumptions or by application of different techniques for making statistical inference. This is discussed in Gavaris (WS1) and reproduced below.

Regarding Bias

Two potentially important causes of bias are a) model mis-specification, and b) estimation of parameters in non-linear models.

a) model mis-specification:

Some theoretical work has been done on properly characterizing uncertainty when making inferences and a choice of model is involved. Typically this work has focused on special families of models where the collection of models to choose from includes the *correct* model. Fisheries assessment models are simple representations of reality, therefore by definition they are not the *correct* model and the suite of model alternatives does not include a correct one. In fisheries assessment, we hope to apply a model that is a close approximation to reality and one for which the assumptions, if violated, do not have significant consequences for the results. The skill of the assessment scientist is required to select one, or a limited number of plausible models. How to best convey the uncertainty associated with this choice has not been established. One of the uncertainties is associated with a shift, or bias, and another is associated with the effect on dispersion of the estimate. Sensitivity analysis of critical model assumptions is one tool for exploring these uncertainties. Another approach has been to assign *degree of confidence* (frequentist methods) or *priors* (Bayesian methods) to the alternative models and derive the resulting confidence distributions or posteriors for the parameters of interest. Yet another approach has been to summarize the potential risks if a model is assumed when an alternate model is more appropriate.

b) estimation for non-linear models:

Estimation of parameters in non-linear models has received much more attention in the statistical literature. The nature of this estimation bias is relatively well understood and is often classified as either *intrinsic* or *parameter effects* non-linearity. Estimates of bias can be readily obtained through linear approximation or from bootstrap techniques. While intrinsic non-linearity cannot be reduced, parameter effects non-linearity can be through transformation of the parameters. This aspect is only marginally useful for fisheries management advice. For example, in the common calibrated VPA, the transformed parameter logN has virtually no bias while the parameter N has measurable bias. Unfortunately, fisheries managers like to have a point estimate of N and not logN. On the other hand, appropriate confidence statements for N could be based on statistical properties of logN if transformation invariant techniques (percentile bootstrap) are employed. A more straight forward approach though, is to use a technique like the bias corrected percentile bootstrap that automatically accounts for bias if it is there. Correcting for bias in the confidence distribution and in confidence statements can be recommended on a routine basis using tools such as the bias corrected percentile bootstrap. Adjusting for bias in point estimates cannot be categorically recommended. The additional variance due to the bias correction can be significant. However, the bias in calibrated VPA models is relatively well determined with good agreement between linear approximation and bootstrap results. Further, if one desires an estimate of central tendency to go along with bias corrected confidence distributions, the bias adjusted point estimates are more suitable.

Summary

In summary, estimation bias in fisheries assessment models appears to be caused largely by parameter effects non-linearity. While it is useful to eliminate this non-linearity to improve performance of numerical optimization (e.g. ADAPT employs this feature) by transforming parameters, it is of limited practical importance if the transformed parameters are not the parameters of interest for fisheries management decisions.

5.3 Diagnostics

5.3.1 Residual patterns in catchability

Stock assessment working groups regularly examine the catchability residual patterns derived from the fitting of assessment models. Under the assumption of constant catchability, transitional changes in the level of catchability will result in retrospective bias in the population estimates derived from the model. As observed by Sinclair *et. al*. (1991), ICES (1991; 1993b), ICES (1997) and Mohn (1999), retrospective patterns are generally introduced into model results by departures from model assumptions. Examples are discussed elsewhere in this report and include mis-reporting of catch data, discarding, trends in the efficiency of commercial tuning fleets etc. Each of these departures from the model assumptions can also induce patterns within the residuals of log catchability. The difficulty is that the causes are confounded and there is insufficient information to distinguish between them.

The problem is illustrated in Figure 5.3.1.1 in which three scenarios for departures from model assumptions are illustrated using a simulated data set. The simulated population is sampled for catch at age data and CPUE data under age independent constant catchability with random log-normal error induced on the CPUE with a 20% CV. A series of retrospective XSA assessments are carried out on the sampled data and the resulting time series of estimates illustrated. In Figures 5.3.1.1a and 5.3.1.1b the catch at age data has been under reported by 20% for the period 1985 – 1993. In Figures 5.3.1.1c and 5.3.1.1d natural mortality was doubled in the real population for the period 1985 –1993 but assumed constant at the lower level in the assessments. In Figures 5.3.1.1e and 5.3.1.1f the catchability of the fleet was increased by 20% for the period 1985 – 1993. The retrospective and catchability residual time series illustrated in Figures 5.3.1.1a-f, where the catchability figures present information for all fleet ages simultaneously. The simulations reveal that without additional information derived from a source that is external to the assessment, one cannot distinguish between the underlying causes of the residual and retrospective patterns.

The utility of additional information is shown in Figures 5.3.1.1g and 5.3.1.1h. In this series of retrospective assessments, a survey (20% CV) with constant catchability has been used to fit the assessment model in addition to the fleet with increasing efficiency. Catch and natural mortality are the true values. The use of inverse variance weighting ensures that the survey receives the majority of the weight in the analysis, due to its conforming to the assumption of constant catchability, and this removes the majority of the retrospective bias. The addition of a source of information that conforms to the model assumptions has revealed the increase in catchability of the commercial fleet. This scenario has been simulated using a survey generated with equivalent noise levels to the fleet data. In reality levels of noise in the survey and catch data may be too large to detect changes in fleet efficiency.

The contrasting of survey and fleet CPUE series within and outside an assessment in order to estimate trends in fleet catchability has been used in several studies of fleet dynamics (e.g. Harley *et al.*, 2001). The extra information provided by the *certain knowledge* that a survey has constant catchability allows the cause of the retrospective pattern to be disentangled. An extra parameter could have been fitted to the fleet catchability in order to model the increase and solve the bias problem but the justification for the extra parameter is not valid without the additional survey information. Attempting to include an extra trend parameter to explore the possibility of a trend requires a strong signal in the remaining data to determine the parameter, and in addition the information in the tuning series about trends in the population abundance will be spent on estimating that parameter. Therefore, due to the confounding of the causes of the departures from model assumptions in residual diagnostics, model over parameterization is easy to achieve. **It is recommended that modelling data sets in order to detect departures from model assumptions should take place prior to the fitting of an assessment model and using data that are independent of the assessment information.**

Retrospective analysis of catchability residuals

Correlation between the estimates derived from sequential assessments, which are not independent because they are primarily based on the same data, can result in retrospective patterns.

During a simulation analysis with noisy CPUE calibration information it was established that an extreme value in the tails of the distribution of catchability data resulted in retrospective bias in the assessment. This occurred because the high value caused a strong revision of catchability that only gradually recovered towards the *true* value with the addition of new information in subsequent years. The retrospective bias resulted from purely random effects within correlated sequential model estimates.

One diagnostic approach that may be useful is a retrospective plot of log catchability residuals. The plots established that the extreme value had a significant influence on the time series of model's estimates of catchability enabling identification of the origin of the pattern. The plots may provide some insight into why the stock estimates are changing systematically.

Are model estimates biased?

ICES (1997) and Darby (WU1) have examined the effects of bias in the catch at age and CPUE data used to fit assessment models. They have each demonstrated that if catch at age, biased by under-reporting, is fitted within an assessment model with an unbiased survey series, a retrospective pattern in the series of terminal year estimates of biomass and SSB is generated (Figure 5.3.1.2). Note that although the retrospective pattern would be considered to be undesirable and could lead to rejection of the model formulation, the terminal year estimates do characterize the "true" population trajectory.

Correspondingly, the two studies have illustrated that the lack of a retrospective pattern does not indicate that an assessment is unbiased. The complexity of residual analysis is compounded by the common usage of fleet data that is derived from the same source as the total catch at age data and therefore correlated with that information. Fitting assessments to correlated, biased information may result in consistent retrospective patterns that give false perceptions of the stock trajectory. Figure 5.3.1.3 illustrates a simulation in which under-reported catch at age data is fitted in an assessment model with fleet CPUE data that is also under-reported in equal proportions, and an unbiased survey. There is no retrospective pattern in the assessment series which would therefore be considered to provide consistent estimates of catches and population numbers for forecasts. However, the estimates are biased and the only indication of this would be the trend in the survey residuals. The biased fleet-data has no catchability trend, it has a CPUE series that is consistent with the biased population calculated from the under reported catch data.

In both simulations the survey series, which is an unbiased index of the *true* population, is estimated to have an increase in catchability because it is not consistent with the populations reconstructed from the under-reported catch data. **It is recommended that analysis of survey series residuals should be given a high priority during the fitting of assessment models, even if a retrospective pattern is not apparent in the time series of assessment estimates. This analysis should take place over the whole of the available time series, not only for the most recent data. Where surveys show transitions in catchability careful consideration should be given to the underlying cause and the quality of the catch data and any commercial tuning series.**

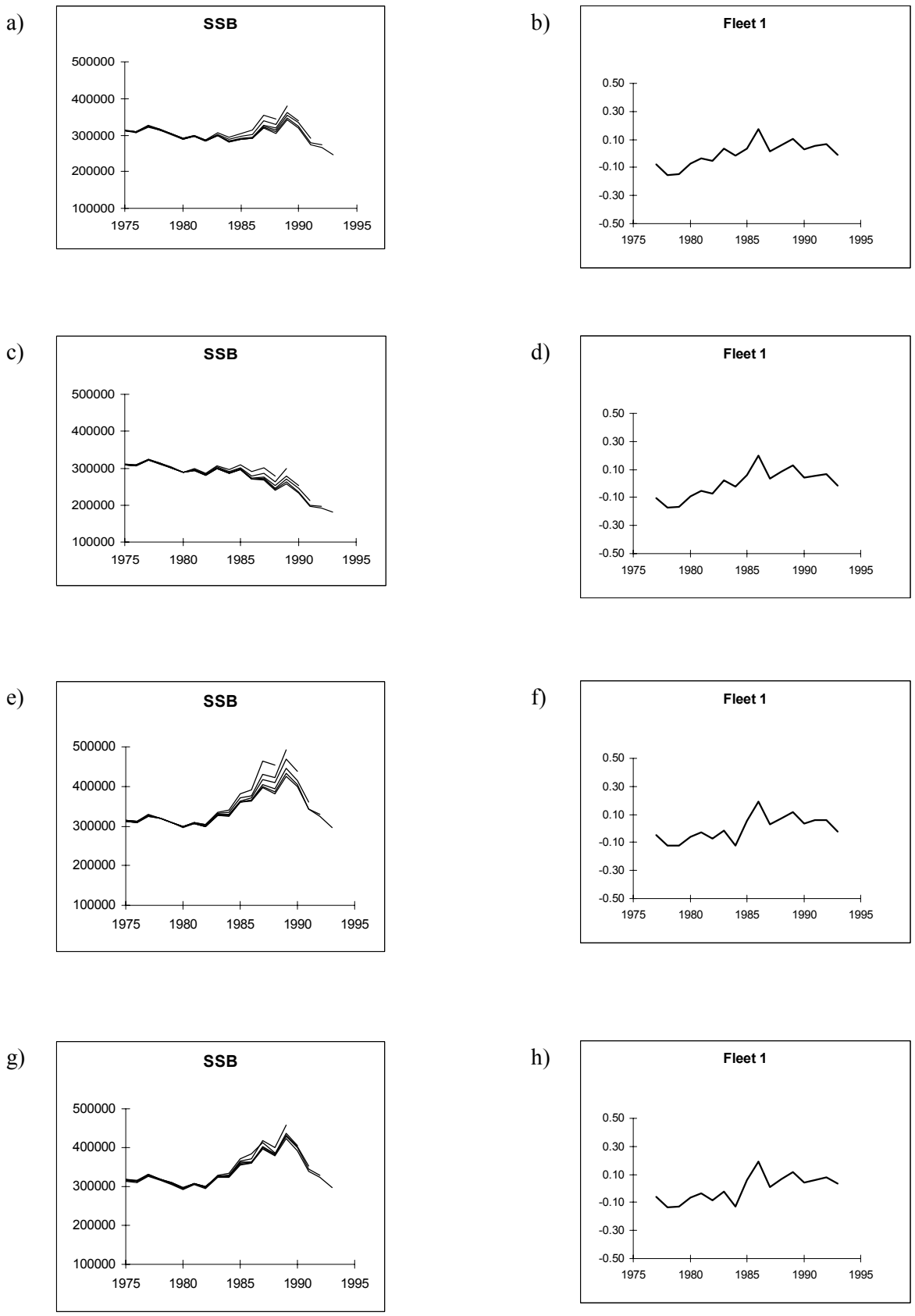


Figure 5.3.1.1 Retrospective patterns and catchability residual patterns generated by violation of assessment model assumptions. (a), (b) under-reporting of the catch data (c), (d) increase in natural mortality, (e),(f) an increase in catchability, (h),(g) the correction of retrospective bias by the introduction of a series with constant catchability.

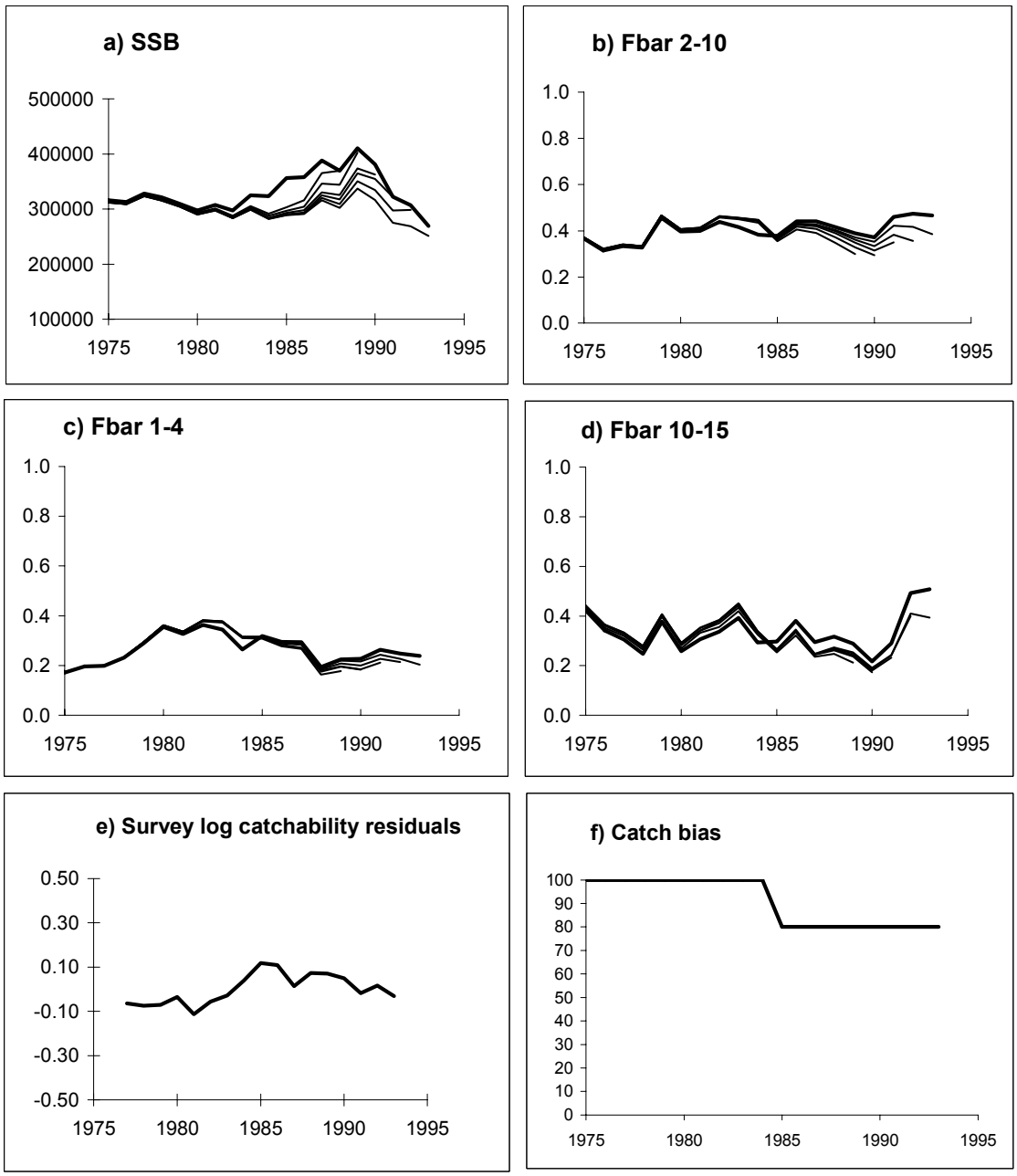


Figure 5.3.1.2 A simulation of the induction of a retrospective bias into a series of estimates from an XSA assessment by a 20% mis-reporting of catch at age data. The solid lines represent the true population trajectory, the fine lines the XSA assessment estimates.

Note that the terminal years of the XSA assessments provide reasonable estimates of the trajectory of the "true" biomass and that the survey, which was generated with constant catchability, appears to have a trend.

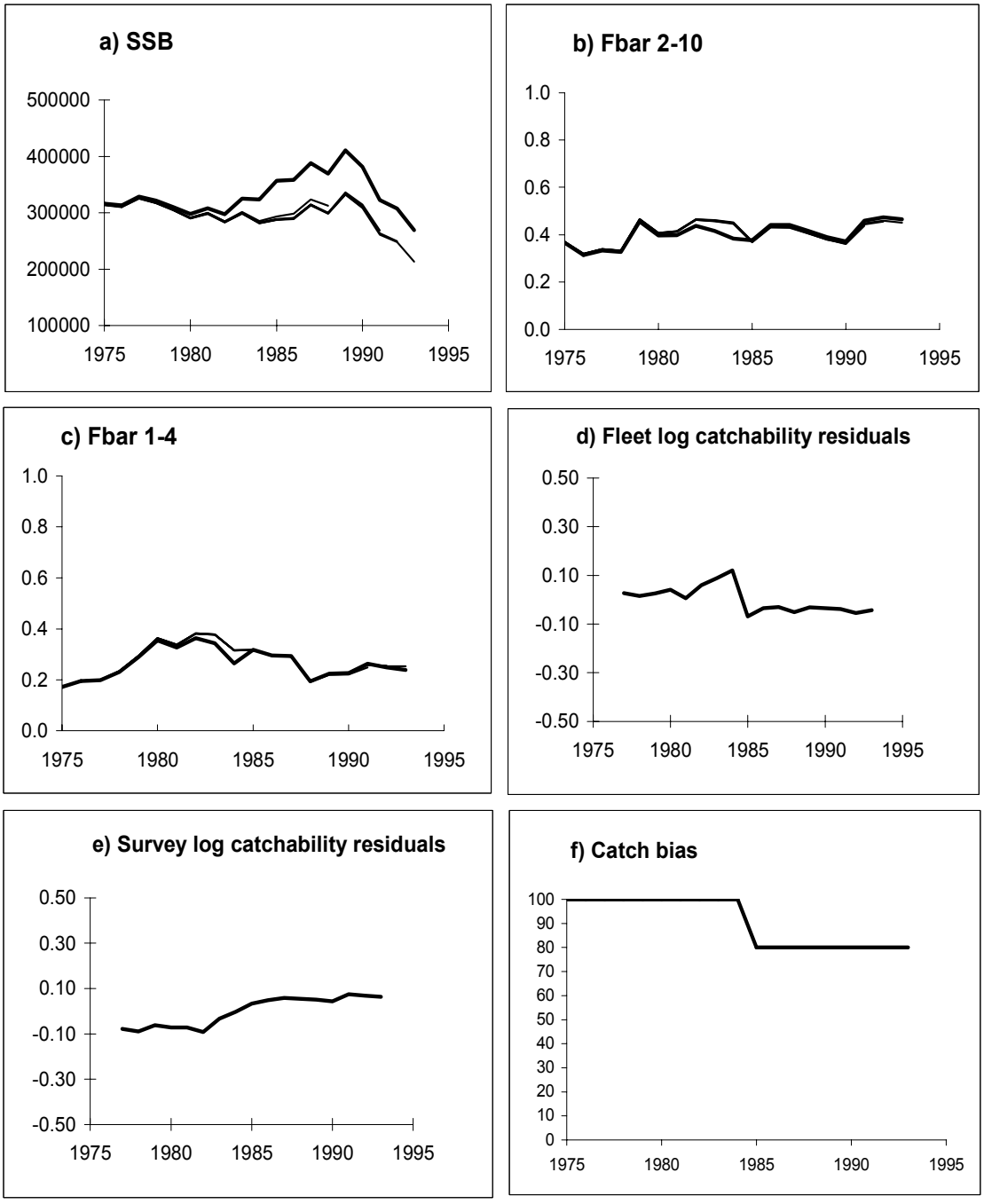


Figure 5.3.1.3 An example of the lack of a retrospective bias in the estimates from an XSA assessment when tuning fleet data is mis-reported with the same proportion as the catch at age data. The solid lines represent the true population trajectory, the fine lines the XSA assessments.

Note the biased fleet catchability has no time series trend in the survey residuals, whereas the survey, which was generated with constant catchability, appears to have a trend.

5.3.2 Local influence diagnostics

A method to estimate the influence of inputs on important assessment model results is described in this Section 5.3.2. This method could be applied to the retrospective problem to increase our understanding of the possible causes. In particular, **the local influence method could be used to find perturbations of VPA inputs that remove or reduce the retrospective problem.** For example, this method can be used to assess whether small adjustments to the catch inputs can remove or greatly reduce retrospective patterns in VPA results. Assessment scientists for the stock can then evaluate the plausibility of the adjustments. By exploring adjustments to a variety of VPA inputs and assumptions, **the method may be useful for identifying a smaller subset of the inputs that are more likely causes of the retrospective problem.**

An important part of any model-based statistical investigation is an assessment of the influence of model components and assumptions on important results such as parameter estimates, functions of parameter estimates, and tests of significance (for example, see Hinrichsen, 2001). Local influence diagnostics involve evaluating the effects of small perturbations to model inputs using simple descriptions of the geometry of the influence surface near the perturbation origin. The diagnostics are based only on unperturbed parameter estimates and are relatively easy to compute. This is important for complex and computationally intensive models because it does not require re-estimation for each perturbation of the model. If the dimension of the perturbation is large, estimating the perturbed model at all possible corners of the influence surface can involve a computationally prohibitive number of optimisations. Local influence diagnostics are particularly useful when the influence graphs are linear; fortunately, this is often the case for VPA's, as demonstrated by Rivard (1989).

Cook (1986) first proposed local influence diagnostics for the likelihood displacement (LD) surface resulting from small perturbations to multiple components of a model. Cook (1986) measured influence using the normalized curvature of the LD influence graph at the perturbation origin. He referred to this as local influence. The curvature of a graph is usually considered a second-order property compared to the slope; however, by construction the LD surface has zero slope at the origin and so the curvature is important. In practical situations there are two problems with the LD local influence approach. The direction of maximum local curvature of the LD influence graph, identified by Cook (1986) as an important diagnostic, is computationally difficult to evaluate for high dimensional (k) perturbations since an eigen-decomposition of a $k \times k$ curvature matrix must be performed. The second problem is that the LD influence measure is focused directly on parameters (θ); however, in VPA's it is functions of θ, not θ itself, that are of interest and used to make decisions. Describing local influence for such models using LD does not lead to demonstrations of model sensitivity that are meaningful to practitioners; that is, LD local influence diagnostics may not clearly reveal whether important model results are sensitive to model assumptions.

The diagnostics in this Section 5.3.2 are more general than those in Cook (1986) in that a wide variety of relevant influence measures can be considered, as well as a wider variety of estimation methods. A description of how to compute generalized local influence diagnostics is presented; together with a demonstration of the method using a recent VPA for the cod stock in NAFO Divisions 3N and 3O. This includes demonstrating that some local influence diagnostics provide useful information about the effect of larger perturbations – these are referred to as global influence. This greatly enhances the utility of local influence diagnostics.

Generalized Local Influence diagnostics

Assume that the problem involves estimating a $p \times 1$ parameter vector θ by maximizing a fit function, $F(\theta)$, that is twice differentiable in θ and yields unique interior parameter estimates. The fit function will depend on data but for simplicity these have been omitted from the notation. The fit function will usually be the kernel of a *true* log-likelihood, but other choices such as a quasi-likelihood may be used. The estimate of θ, denoted as $\hat{\theta}$, is the solution to

$$\dot{F}(\hat{\theta}) = \frac{\partial F(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0.$$

Next perturb k model components with a $k \times 1$ perturbation vector w and study the influence of the perturbation on key model results. Let $g(\theta)$ be such a result; for example, $g(\theta)$ could be the spawning stock biomass in the last year of the VPA. The perturbation w is of the form $w=w(h)=w_0 + h \times d$, where w_0 is the null perturbation, d is a fixed direction vector of length one, and h is a scalar that determines the magnitude of the perturbation. Sometimes $w \varepsilon R^k$ but often w must belong to a subspace of R^k depending on the model and perturbation scheme. Interest is centred on assessing

influence for the perturbed result $g_w = g_w(\hat{\theta}_w)$ which is assumed to be a first-order differentiable function in h and θ. Note that g_w depends on w not only through $\hat{\theta}_w$, which is an important difference between LD and this approach.

Measure the first-order local influence of a perturbation using the slope in the direction d, denoted as $S(d)$, of the influence graph of g_w versus $w(h)$,

$$S(d) = \frac{\partial g_\omega(\hat{\theta}_\omega)}{\partial h}\Big|_{h=0} = d' \frac{\partial g_\omega(\hat{\theta}_\omega)}{\partial \omega}\Big|_{h=0}.$$

Using the chain rule it can be shown that

$$S(d) = d' \left\{ \frac{\partial g_\omega(\hat{\theta})}{\partial \omega}\Big|_{\omega=\omega_o} + \frac{\partial \hat{\theta}'_\omega}{\partial \omega}\Big|_{\omega=\omega_o} \frac{\partial g(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} \right\}.$$

Also, it follows that

$$\frac{\partial \hat{\theta}_\omega}{\partial \omega}\Big|_{\omega=\omega_o} = -\ddot{F}^{-1}\Delta, \tag{1}$$

where F is the Hessian matrix,

$$\ddot{F} = \frac{\partial^2 F(\theta)}{\partial\theta\partial\theta'}\Big|_{\theta=\hat{\theta}}$$

and

$$\Delta = \frac{\partial^2 F_\omega(\theta)}{\partial\theta\partial\omega'}\Big|_{\theta=\hat{\theta},\omega=\omega_o}.$$

Note that (1) is similar to equation (14) in Cook (1986).

For case-weight perturbations the □ matrix is simply the Jacobian matrix. Case-weight perturbations are usually written as $F = \Sigma f_i w_i$. For this perturbation scheme it is easy to show that

$$\Delta = \begin{Bmatrix} \partial f_1/\partial\theta' \\ \cdot \\ \cdot \\ \partial fk/\partial\theta' \end{Bmatrix}.$$

Using (1), a relatively simple formula for $S(d)$ is

$$S(d) = d'\,\dot{g}_o, \tag{2}$$

where

$$\dot{g}_o = \left\{ \frac{\partial g_\omega(\hat{\theta})}{\partial \omega}\Big|_{\omega=\omega_o} - \Delta' \ddot{F}^{-1} \frac{\partial g(\theta)}{\partial \theta}\Big|_{\theta=\hat{\theta}} \right\}.$$

Compute $\dot{g}_0$ without estimating $\hat{\theta}_w$; this is computationally more tractable because it does not involve $\hat{\theta}_w$. When k and p are large, computing $\partial\,\hat{\theta}_w / \partial\, w$ can be very time consuming for complex models like SPA for which estimation of θ often involves nonlinear optimization. All that is required to compute (2) is $\hat{\theta}$ and two derivatives involving g which in many situations can easily be computed analytically or using a single numerical differentiation. This is particularly advantageous when g_w is much faster to evaluate than F_w because many evaluations of F_w are usually required to compute $\hat{\theta}_w$.

S(d) may be used to compute the slope of the influence surface in a variety of directions. For example, if interest centres on the local slope for the perturbation of a single model component then this requires setting the corresponding element of d equal to one and all others to zero. Another interesting direction is the one with maximum local slope, denoted as s_{max}. Perturbations with large absolute elements in s_{max} are also relatively influential. Under the constraint $d'd=1$, the Cauchy-Schwarz inequality can be used to show that the maximum slope is $S(s_{max})=(\dot{g}_o'\dot{g}_o)^{1/2}$ and $s_{max} = \dot{g}_o/S(s_{max})$. Note that s_{max} is invariant to uniform scalar transformations of g or w, and this is a desirable property. Once $\dot{g}_o$ has been computed then s_{max} is easy to obtain; in particular, it does not involve an eigen-analysis of a potentially large curvature matrix.

SPA example

Next is illustrated the utility of the generalized local influence diagnostics using an SPA for cod off the southeast coast of Newfoundland, Canada. The stock is located in NAFO Divisions 3N and 3O (see Figure 5.3.2.1). The data and the SPA methods and results are described in Stansbury *et al*. (1999) which the interested reader should consult for further details.

An electronic version of the document is available at:

http://www.nafo.ca/publications/meetings/SciCoun/1999/resdocs/scrtoc.htm.

This document includes a description of the basic biology of 3NO cod, and of the cod fishery in this region.

The catches used in the 3NO cod SPA cover ages 2-12 for 1959-1998. The catches, in thousands, are given in Table 23 in Stansbury *et al*. (1999). Three surveys were used for estimation. The surveys cover ages 2-10 for 1984-1998. This SPA formulation differs slightly from that of Stansbury *et al*. (1999); however, the estimates of survivors derived are very close to theirs, with the largest absolute relative error being 2.4%. Estimates are not present here. Note that the mean number per tow index for age 2 in 1998 is actually 0.16 and not 2.16 as reported in Table 18 in Stansbury *et al*. (1999). However, the SPA results in Stansbury *et al*. (1999) are based on the correct value, and these values are used in the analysis presented.

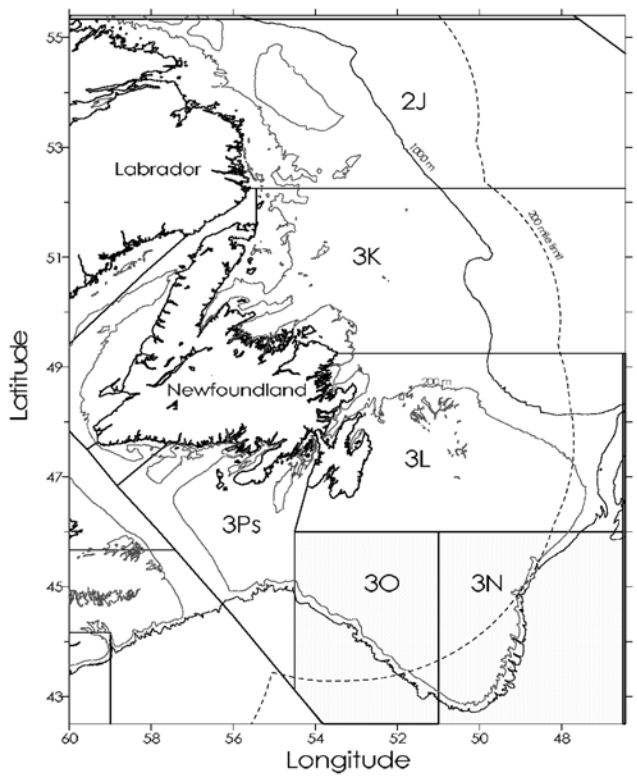


Figure 5.3.2.1. NAFO management boundaries around insular Newfoundland, off the east coast of Canada. The 200 and 1000 meter depth contours are plotted, along with the 200 mile Canadian resource jurisdictional limit.

Influence analyses

There are a variety of assumptions in the 3NO cod SPA that are tenuous to varying degrees. This holds true for most SPA's. For example, the commercial catches for 3NO cod are estimated and their quality has varied over time, due in part to the international nature of the fishery. Uncertainty about M is also a concern. It is useful to explore tenuous assumptions using sensitivity analyses. This is because stock assessment scientists often cannot predict the outcome of a change in an SPA input or model assumption, which makes it difficult to assess which assumptions are most influential in determining important model results. The direction of maximum change for local influence diagnostics is particularly appealing for this purpose since it can be used to assess the worst scenario involving some tenuous assumption. If the effect of this perturbation does not alter our conclusions, then one may feel more confident about these results. There is also a need to understand the sensitivity of our stock-size inferences in order to decide where to best expend effort in improving SPA's by improving inputs or assumptions.

Examples of two influence analyses for 3NO cod are presented. The first analysis involves perturbations to the catch data, while the second involves perturbations to M. Influence is assessed using two measures. The first influence measure uses the estimated total exploitation rate in 1998, which is the last year of catch in the SPA; that is,

$$\text{estimated total exploitation rate (TE)} = \frac{\sum_a C_{a,1998}}{\sum_a \hat{N}_{a,1998}}.$$

This is the fraction of the initial stock in 1998 that is removed by the fishery.

The second influence measure is an estimated change in stock productivity, which is an interesting feature of the 3NO cod SPA. The amount of production, or number of offspring per unit of sexually mature population, is estimated to have declined between the 1960's and the 1980's (see Fig. 17 in Stansbury *et al*., 1999). The greatest differences are the large year classes in 1962-1964 and the low year classes in 1983-1987. This variation in year class strength occurred at approximately equivalent parental population sizes. To examine the sensitivity of the SPA estimate of the trend in productivity, an influence measure.

$$\text{estimated recruitment trend (RT)} = \frac{\sum_{y=1986}^{1990} \hat{N}_{3,y}}{\sum_{y=1965}^{1967} \hat{N}_{3,y}}.$$

is used. Note that the measure of year class strength uses population size three years later. The components of $S(d)$ are computed numerically.

3NO cod SPA sensitivity to catch

The catch perturbations are of the form $C_w = (C+\delta) \times w$. A small offset, $\delta = 0.1$, is added to the catch so that zero catches are also perturbed. The average catch for all ages and years is 2771 (in thousands), so δ is a very small adjustment. The perturbations are of the form $w = 1 + hd$. For an individual catch perturbation $C_{wi} = (C+\delta)(1+h)$. Multiplicative perturbations for catches are used since any differences in the reported catches and the true anthropogenic removals are probably more multiplicative than additive.

The elements of s_{max} are plotted in Figures 5.3.2.2 and 5.3.2.3 for ***TE*** and ***RT***, respectively. The results in Figure 5.3.2.2 suggest that increasing the catch in 1998 increases ***TE***, which is an obvious result, but that increasing the catch prior to 1994 tends to decrease ***TE***. This is a less obvious result, but commonly known among SPA "experts". Changing the catch during 1994-1997 tends to have relatively small effects on TE. Note that changing the catch prior to 1984 has no effect on ***TE*** and all the corresponding elements of s are zero (not shown in Figure 5.3.2.2). This is because the surveys used to estimate SPA parameters only cover the 1984-1998 period (Stansbury *et al*., 1999). Therefore, SPA estimates prior to 1984 are extrapolations, and altering catches in these years has no effect on ***TE***.

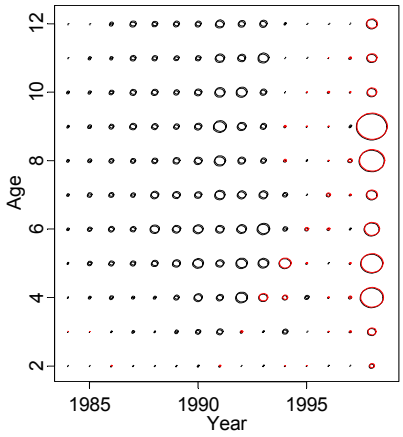


Figure 5.3.2.2. Elements of s_{max} for the total exploitation rate (***TE***) catch influence analysis. $S(s_{max})=44.2\%$ of the estimated 1998 ***TE*** (0.05). The bubble area is proportional to the absolute value of the element, and the color denotes the sign of the element (red +, black -).

The results in Figure 5.3.2.3 suggest that ***RT*** can be increased (towards no trend) by either decreasing the catches of the 1962-1964 cohorts (e.g. age 5 in 1967-1969), or increasing the catches of the 1983-1987 cohorts (e.g. age 5 in 1988-1992). The changes are greater for catches at ages 3-6 than at other ages. To understand the results in Figure 5.3.2.3, recall that SPA computes historic recruitment by essentially adding up catches and other mortality (M). Therefore, decreasing the catches for the 1962-1964 cohorts will decrease the estimated size of these year classes, and this leads to an increase in ***RT***. Similarly, increasing the catches of the 1983-1987 cohorts increases the estimated size of these year classes and also increases ***RT***.

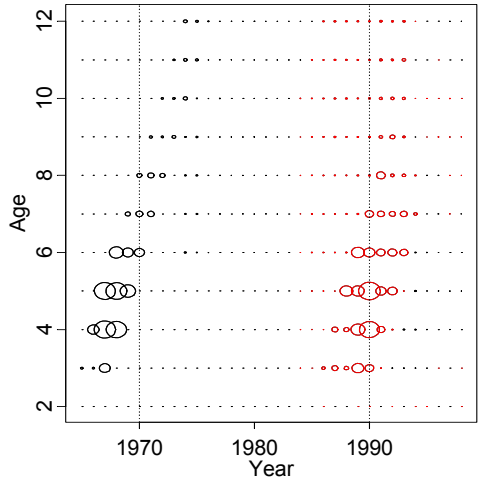


Figure 5.3.2.3. Elements of s_{max} for the recruitment trend (***RT***) catch influence analysis. $S(s_{max})$=41.4% of the estimated (0.05). See Figure 5.3.2.2 for the bubble description.

In Figures 5.3.2.4 and 5.3.2.5, plots are shown of ***TE*** and ***RT*** influence graphs for selected perturbations that were chosen to demonstrate the global properties of the local diagnostics. The direction of maximum local slope, s_{max}, as an *extreme* perturbation. Individual perturbations with large negative and positive slopes, and one with a near zero slope were selected. Figure 5.3.2.4 illustrates that changing catches in the direction of s_{max} produces the largest change in ***TE***, and that the influence graph is almost linear. By construction $S(s_{max})$ is always positive. The ***TE*** local slope for $C_{9,1998}$ is 0.0135, which is slightly more than half of $S(s_{max})$, which is also what Figure 5.3.2.4 suggests. The local slope for $C_{4,1992}$ is -0.0019, and Figure 5.3.2.4 demonstrates that increasing $C_{4,1992}$ always decreases ***TE***, but that the absolute effect of perturbing $C_{4,1992}$ is less than the effect of perturbing $C_{9,1998}$. This is exactly what the local influence diagnostics suggest. Note that the local slope for $C_{11,1994}$ is 2.5 x 10^{-3} which is very small, and $C_{11,1994}$ is not globally influential on ***TE*** either. The ***RT*** local slope is -0.007 for perturbations to $C_{5,1968}$, 0.002 for $C_{6,1990}$, and 0.0002 for $C_{11,1994}$. Figure 5.3.2.5 demonstrates that the ***RT*** local slopes also provide a good summary of the global influence graphs because the relative patterns in local influence also persist globally.

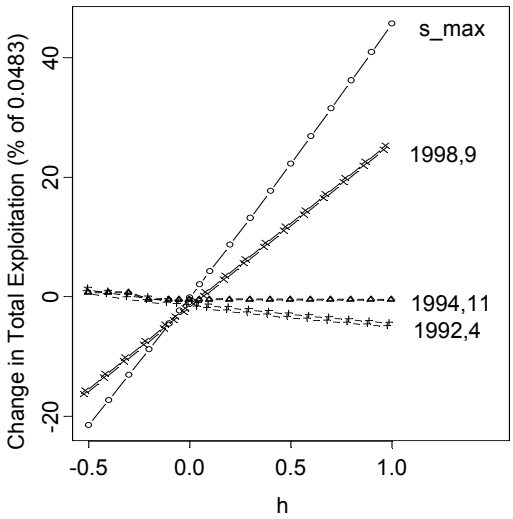


Figure 5.3.2.4. Total exploitation rate (***TE***) displacement for some global catch perturbations. The unperturbed estimate of ***TE*** is given in parentheses. The (year, age) indicate perturbations of individual catches.

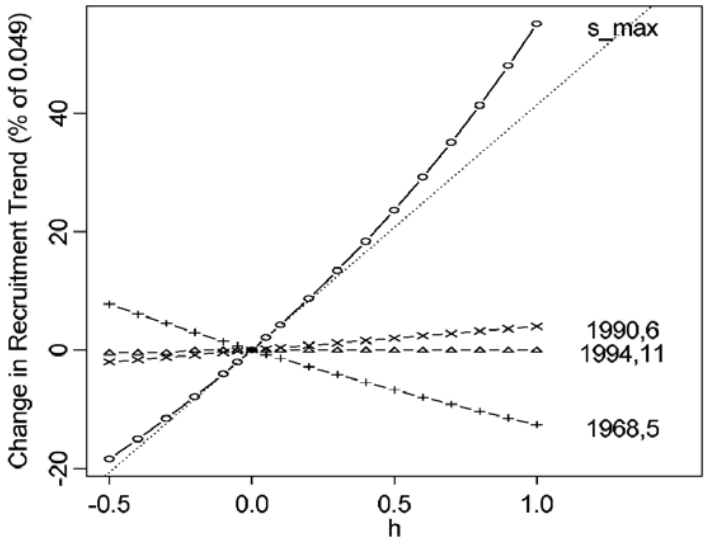


Figure 5.3.2.5. Recruitment trend (***RT***) displacement for some global catch perturbations. Plotted text is similar to Figure 5.3.2.4. The unperturbed estimate of RT is given in parentheses. A straight line with a slope corresponding to $S(s_{max})$ is plotted as ……'s.

3NO cod SPA sensitivity to M

The M perturbations are of the form $M_w = M+w$, where $w=hd$. Additive perturbations to M affect the model in a multiplicative manner. One might expect M perturbations to affect SPA estimates similarly to catch perturbations; however, this is not always the case. The elements of s_{max} for ***TE*** are plotted in Figure 5.3.2.6. Contrary to the results in Figure 5.3.2.2, the results in Figure 5.3.2.6 suggest that increasing M in 1998 decreases ***TE***. The reason is that M in 1998 is used to project (reduce) $N_{a,1998}$'s to the time that the survey occurs. To obtain the same fit to the survey data, the $N_{a,1998}$'s must also increase, which results in a decrease in ***TE***. Another difference is that increasing M in 1995-1997 tends to decrease ***TE*** whereas multiplicative increases to the catch in these years have almost no effect. However, similar to Figure 5.3.2.2, the results in Figure 5.3.2.6 suggest that increasing M prior to 1994 tends to decrease ***TE***.

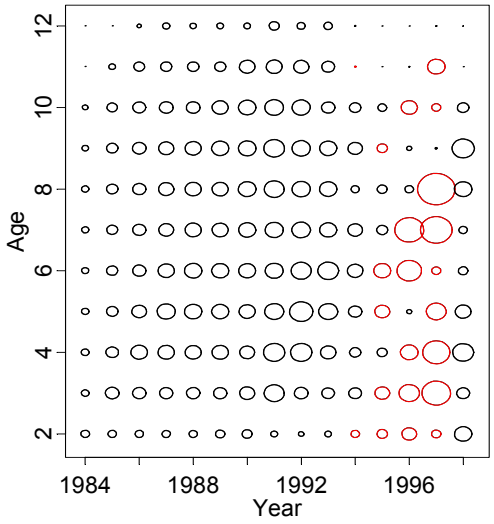
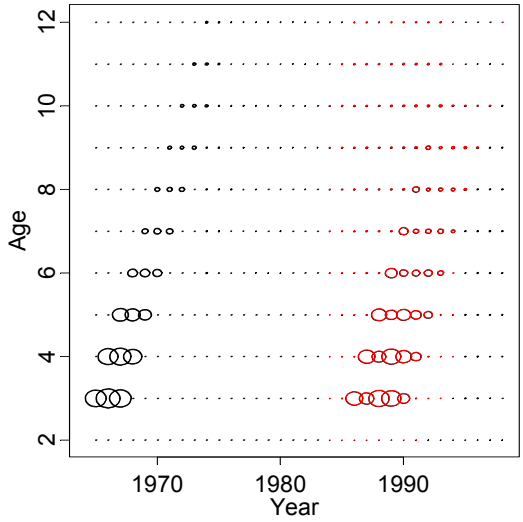


Figure 5.3.2.6. Elements of s_{max} for the total exploitation rate (***TE***) *M* influence analysis. $S(s_{max})=35.6\%$ of the estimated 1998 ***TE*** (0.05). See Figure 5.3.2.2 for the bubble description.



Figu**re 5.3.2.7**. Elements of s_{max} for the recruitment trend (**RT**) *M* influence analysis. $S(s_{ma})=102\%$ of the estimated ***RT*** (0.05). See Figure 5.3.2.2 for the bubble description.

The results in Figure 5.3.2.7 are similar to those in Figure 5.3.2.3, except that *M* at age three for the 1962-1964 and 1983-1987 cohorts tends to be more influential on ***RT*** than the age three catches for these cohorts. Also, $S(s_{max})$ is more than twice the value in the catch influence analysis. This difference is further considered in the discussion later in this Section 5.3.2.

Global influence analyses for ***TE*** and ***RT*** are presented in Figures 5.3.2.8 and 5.3.2.9. The *M* perturbations range from $M_w \varepsilon [0.1,0.5]$. Values outside this range may be unrealistic. The corresponding local slopes for the perturbations in these figures are presented in Table 5.3.2.1. Similar to the catch perturbations, the results in Table 5.3.2.1 provide a useful summary of the influence graphs in Figures 5.3.2.8 and 5.3.2.9.

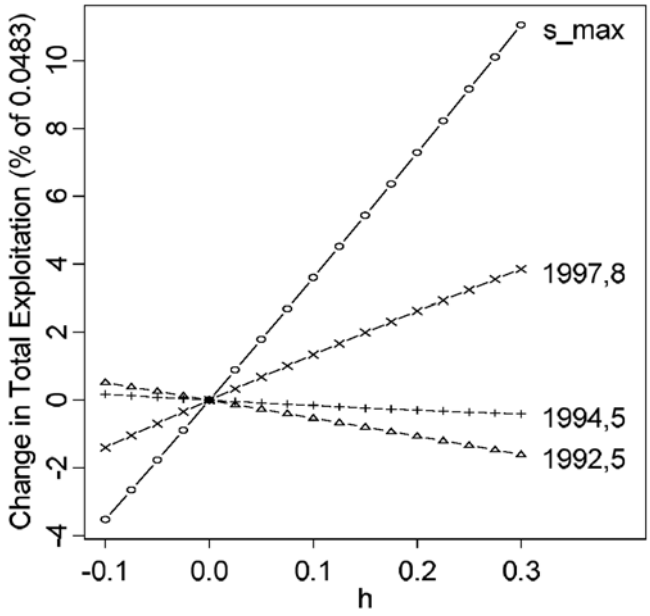


Figure 5.3.2.8. Total exploitation rate (***TE***) displacement for some global *M* perturbations. Plotted text is similar to Figure 5.3.2.4. The unperturbed estimate of ***TE*** is given in parentheses.

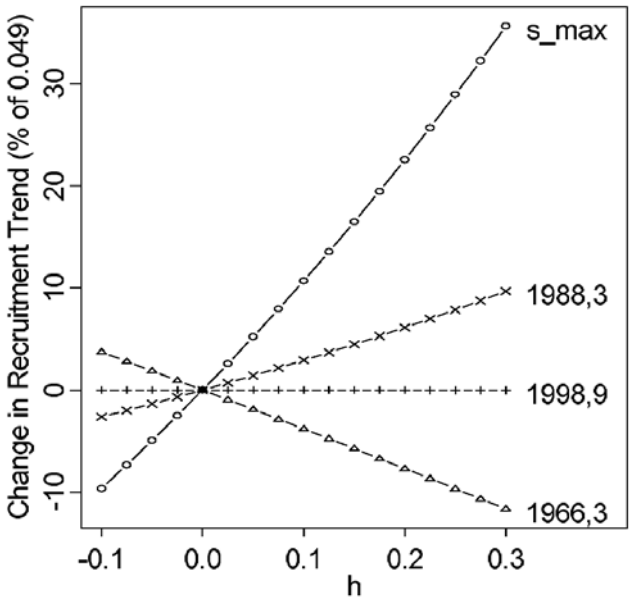


Figure 5.3.2.9. Recruitment trend (***RT***) displacement for some global *M* perturbations. Plotted text is similar to Figure 5.3.2.4. The unperturbed estimate of ***RT*** is given in parentheses.

Table 5.3.2.1 Local slopes for the perturbations in Figures 5.3.2.8 and 5.3.2.9.

TE		**RT**	
Perturbation	Slope	Perturbation	Slope
s_{max}	0. 0172	s_{max}	0. 0500
1997,8	0. 0066	**1988,3**	0. 0136
1994,5	-0. 0008	**1998,9**	< -0.0000
1992,5	-0. 0025	**1966,3**	-0. 0185

Discussion

Results have been presented that facilitate a general assessment of local influence; that is, for general fit (objective) functions and general measures of the effects of perturbations. Our results can be used to construct a practical set of relevant diagnostics by focusing on the statistics that decisions are based on. The SPA example illustrates how generalized local influence diagnostics can lead to a better understanding of a model, and the sensitivity of important model results to input data and assumptions.

The ***TE*** and ***RT*** local influence analyses for the 3NO cod SPA produced results that SPA "experts" may have qualitatively predicted; however, many people who use SPA are not "experts", and presenting influence diagnostics can assist in their understanding of SPA. Even SPA experts may not have predicted some of our results. An empirical assessment of the global properties of the local influence diagnostics for the SPA have been presented for one example stock. The diagnostics have been shown to provide an excellent description of global influence, and the local influence diagnostics can give good approximations of the effect of some global model perturbations.

Other SPA influence analyses have been performed for three fish stocks using different perturbation schemes and many (usually 6) measures of influence. The conclusions reached from these analyses tend to agree with Rivard (1989). However, our methods are a substantial improvement over his because all model components are perturbed simultaneously and only minimal constrains on the local perturbations (e.g. $C_w > 0$) have been employed. Rivard (1989) considered only constrained perturbations such as equal sized relative perturbations to the catch data or part of the catch data, or random perturbations to the catch data. Such perturbations would miss some of the sensitivities that our approach can detect, such as those in Figure 5.3.2.2.

When curvature is present it is more complex to interpret generalized local influence diagnostics, and it is more difficult to assess the effect of non-local perturbations when $||w_0 - w||$ is large. The catch perturbations with ***RT*** influence illustrated this. A quadratic approximation to the graph of $g_w(\hat{\theta}_w)$ versus w using $\partial^2 g_{w(h)}(\hat{\theta}_{w(h)}) / \partial h^2$ may produce better local predictions of the effect of a perturbation. However, this raises questions about the utility of s because it will not necessarily give the direction of maximum change even for perturbations that are close to w_0. Describing influence when substantial curvature is present is a useful area for future research.

There will likely be a tendency for some stock assessment scientists to use the influence analysis methods presented here to assess whether an SPA is more sensitive to catch or M, or some other modelling component. Our results suggest that this can be problematic, since it is only sensible to compare influence diagnostics if the perturbation schemes are comparable. Consider the results from the catch and M ***RT*** influence analyses. For the catch influence analysis $S(s_{max})=0.020$, which is less than half the value for the M influence analysis, where $S(s_{max})=0.050$. Superficially this might suggest that the 3NO cod SPA tends to be more sensitive to M than catch; however, the problem is the scale of the perturbations. Although s_{max} is invariant to uniform scalar transformation of w, $S(s_{max})$ is not. $S(s_{max})$ can be interpreted as the change in g_w when $h=1$. It was thought sensible to look at catch perturbations when $h=1$ because for a single perturbation this implies a doubling of the catch, which is a plausible perturbation. However, M perturbations in excess of $M+0.3$ were not investigated because this would imply an unreported annual mortality in excess of 40% which seems unrealistic. This suggests that the catch and M local slopes are not comparable. The problem of defining comparable perturbation schemes is tantamount to deciding on the magnitude of the errors in inputs that is sensible to consider. Unfortunately, for SPA's little information to use for this purpose is often available.

Nonetheless, the local influence analysis can shed some light on whether an SPA is more sensitive to catch or M. Our results from the ***TE*** influence analysis suggest that ***TE*** in 1998 is more sensitive to catch than M. Catch perturbations in 1998 (see Figure 5.3.2.4) tend to produce greater variations in ***TE*** than any "similar-sized" M perturbations (see Figure 5.3.2.8). Substantially larger M perturbations are required to produce the amount of variation depicted in Figure 5.3.2.4. This leads to the conclusion that in the 3NO cod SPA, uncertainty in the 1998 catch is more important for inferences about total exploitation in 1998 than is uncertainty about M. The local influence analyses may also provide information on whether other SPA outputs are more sensitive to catch or M; however, the results of the ***RT*** investigation suggest that care must be taken in using these analyses for such purposes.

Conclusions

The working group agreed that the local influence diagnostics could provide useful information about model sensitivity. This information is currently not considered in a routine manner. **The working group recommends that influence diagnostics should be developed for routine use within stock assessments, addressing both data and modelling issues. It is further recommended that such methods be applied to specific case studies to examine their potential for analyzing retrospective problems.**

5.3.3 Over-parameterised models

Typically, analytical assessments of fisheries involve complex and computationally intensive models.

Over-parameterised models, such as separable models using only catch data, can give rise to estimates of terminal fishing mortalities or population numbers-at-age, for example, that are unreliable (Skagen WS4). Presenting confidence intervals and confidence regions around the parameter estimates can help to avoid over-interpreting point estimates and to identify over-parameterisation.

Approximate confidence intervals can be constructed in several ways. If the parameter estimates are obtained by maximum likelihood, confidence limits can be obtained by:

- profile likelihood methods: these might be preferable because they make no assumptions about the shape of the likelihood surface, and

- calculating the Hessian of the log-likelihood (at the maximum likelihood estimate),

More generally, bootstrap methods could be used. Extensions of bootstrap techniques may adjust for some types of bias in the parameter estimators.

All these methods also give information about the covariances between the parameter estimates through correlation matrices, profile likelihood regions, and matrices of bootstrap realisations. This information could be used to identify those parameter estimators that are (almost) confounded and to suggest more stable (less correlated) or more parsimonious parameterisations.

Note, however, that Hessian, profile likelihood and bootstrap methods can all give misleading inferences when applied to over-parameterised models. Further, their performance is likely to depend on the model under consideration. It would therefore be sensible to investigate the performance of these inference methods when applied to the over- (highly) parameterised models typically used in stock assessment. One example may be the assessment of NEA mackerel in 1998. The input to the assessment, which is done with ICA, is catch numbers-at-age and a triennial SSB estimate from egg surveys. The outcome of the assessment was extremely sensitive to the weight given to the SSB data, but the diagnostics, including the variance estimate for the terminal F were not alarming (ICES, 1999a).

Comparing the results from standard estimation procedures for parameterised models with those of more robust/resistant techniques would help to detect results that are due to a few highly influential data points. This is particularly relevant to over-parameterised models.

5.3.4 Bounding the scale of the possible causes of the retrospective pattern

Stock assessment working groups should be encouraged to provide bounds for the changes to the assessment data that would be required to remove retrospective patterns, if present. For example, the question: *What is the magnitude of the required changes to the catch, CPUE and/or discard data or the value of natural mortality, that are required to provide consistent estimates from the assessment?* The magnitude of the required changes might give some insight into the more probable causes of bias to be evaluated.

Shelton & Lilly (2000) and Morgan & Brodie (2001) have carried out studies that explored the causes of retrospective patterns by estimating additional parameters for stocks in the NAFO region. Shelton & Lilly (2000) examined why the assessments for the Northern cod stock (2J3KL) failed to provide acceptable reconstructions of the population when data for the 1990s were added to the assessment. The assessment model assumed a constant age-independent instantaneous rate of natural mortality, accurate catch reporting, and a constant catchability for each age. The authors considered it likely that one or more of the model assumptions failed to hold during the time of the Northern cod stock collapse. Studies were carried out to determine the magnitude of the departure from the assumptions required in order to allow the model to fit the data. Information related to changes in natural mortality, fishing activity, and survey catchability was reviewed to evaluate the plausibility of departures from model assumptions of the magnitude estimated. It was concluded that un-reported catch was the most plausible of the main contributing factors to the lack of model fit. However, as the amount of extra catch required considerably exceeded the capacity of the commercial fleets, then factors such as increased natural mortality, and possibly changes in survey catchability, also played a role.

Morgan & Brodie (2001) carried out a similar analysis of the effects of increasing natural mortality during specific years on the consistency of the assessments of American plaice in NAFO Divisions 3LNO. They established that a

change in the assumed natural mortality from 0.2 to 0.53 in the years 1989 – 1996 produced a more consistent assessment.

While these studies do not provide definitive answers to the causes of inconsistencies between model assumptions, they do allow bounds to be ascertained for the required changes to input data and model assumptions. These studies may provide information about the feasibility of various departures in model assumptions which can be carried forward within the advice on the quality of the assessment.

5.4 Conclusions

It seems clear that there is no single cause for the retrospective patterns seen in some stock assessments. In general, the problem is not primarily a problem with the assessment tool, but rather that the assessment model, as it is specified, fails to interpret the assessment data in the appropriate way, because the data represent something that is different from the *a priori* assumptions.

Making retrospective assessments may reveal that the problem has been there in the past, but there is no diagnostic that can confirm that the measures that may have been taken have eliminated the problem. In some cases, current diagnostics can point to the cause of the problem, and additional diagnostics have been recommended within this Section 5 that will add insight to the problem.

One of the important observations from this Group, and other fora, is that it is not automatically obvious that the terminal year estimate from a stock assessment is in error and that the historical reconstruction is accurate. In other words, in any stock assessment, the terminal year estimate may be reasonable but as data are added, subsequent estimates for that year from the VPA historical reconstruction could be erroneous. The conclusion is that it would not be possible to determine if bias adjustment were necessary unless the cause of the inconsistency was at least postulated.

Some causes of the problem, like catchability trends in tuning indices, are readily identified. However, it is also clear that the causes of the problem are not fully understood, and that they may be quite complex. **The WG considers that there is a need for extensive studies with data sets with known properties, and has outlined a framework for making such studies with simulated data**.

The advice at present to assessment scientists that face a retrospective problem would be that if a certain problem with the data is suspected as a mitigating cause of the problem, one should explore how any retrospective pattern may be eliminated by perturbing the data. This will both help to verify the hypothesis about the cause, and allow the appropriate consideration of whether or not the corresponding errors in the data are of a realistic order of magnitude. Influence diagnostics have been proposed by WGMG to investigate the nature of this problem.

6 POPULATION FORECASTING

6.1 Medium-term projection

6.1.1 Introduction and context

Definitions

Medium-term analyses carried out by ICES Working Groups are projections of future yields and SSB over a 5–10 year forecasting period. The projections are based on a large number of simulations, starting from the most recent assessment of the state of the stock, which is projected forward making assumptions on expected recruitment, weight-at-age, maturity-at-age, natural mortality, selection pattern and the exploitation rate in the forecasting period. These simulations can be done for different assumptions about levels of fishing mortality. The obtained set of simulations can be used to define probability profiles of the expected catch and stock.

Purpose

Medium-term analyses were originally designed to estimate the probability of achieving a specified objective if the fishery was continued under certain assumed conditions over a longer period. ICES presently uses such analyses in the formulation of its annual fishery management advice. In general, ICES will refrain from giving advice that is likely to bring SSB below defined precautionary limits within the medium-term. Also if the SSB is below the defined limits, short-term advice is directed to bring the stock above this limit within the medium-term.

The medium-term software that is currently used by Working Groups permits the presentation of a time trajectory of the development of the SSB and expected catch and associated uncertainty (see example in Figure 6.1). Recently, a new presentation was introduced by Robin Cook, in the form of a contour plot showing time trajectories of probabilities that $SSB < B_{pa}$ for different assumptions of fishing mortalities (see example in Figure 6.2).

Fisheries managers generally consider the phrase *medium-term* to imply a 4- to 6-year period, in contrast to ICES Working Groups which assume an 5- or 10-year period. In recent years, decisions on next year's TAC and recovery plans have been based on projections of the expected development of the stock in the first years of the medium-term trajectories as presented by ICES.

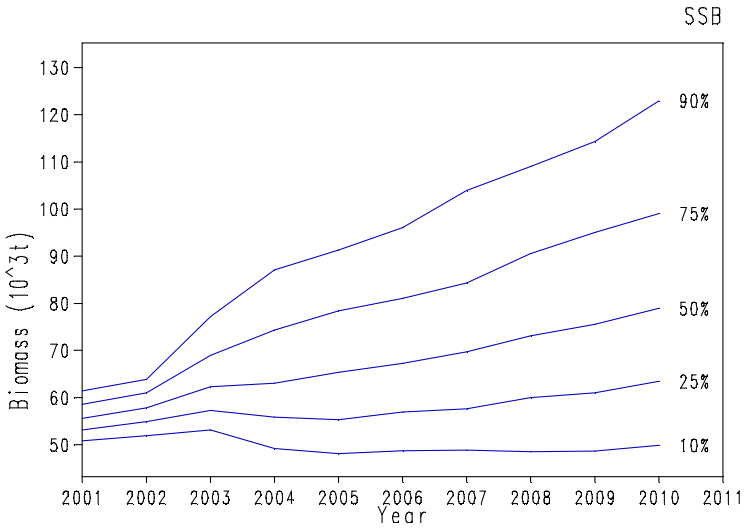


Figure 6.1. Percentiles of medium-term SSB projections for North Sea cod, assuming s*tatus quo F*. Source: ICES (2002).

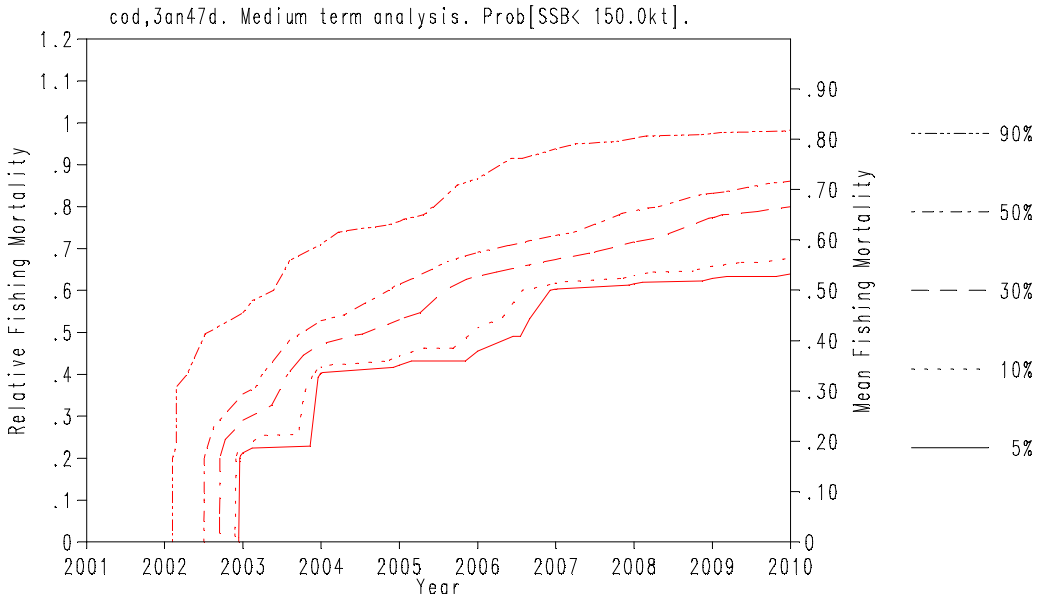


Figure 6.2. Contour plot of the probability that $SSB < B_{pa}$ for a range of *F*-multipliers in medium-term projections for North Sea cod. Source: ICES (2002).

Desirable properties

Short-term forecasts of SSB and landings may vary considerably between subsequent assessments, because these prognoses are very dependent on the present state of the stock and the expected recruitment. On the other hand, medium-term projections should be less dependent on the present situation and more dependent on the imposed biological parameters, assumed selectivity pattern and the choice of a recruitment model. Therefore, in principle they should be more stable.

Medium-term analyses are considered to be useful by managers because they are perceived to give an indication of achievable biomass targets in the medium-term. They also indicate the amount of action to be taken in the fisheries (reduce fishing mortality) in order to achieve an acceptable probability that these targets are met; and frame management perceptions about stock development in the period between short-term forecasts and long-term equilibrium analyses. The measures to be taken would in many cases require considerable action and could have severe consequences to the fishing industry. The basis for justifying these actions and accepting the consequences must be solid and stable. Therefore, medium-term projections would have to be relatively stable from one year to the next. Only when there are clear changes in biological processes (growth, maturity), exploitation pattern (technical measures) or expected recruitment, can significant changes in medium-term prognoses be justified.

Sensitivity to input parameters and assumptions

The results in the first years of a medium-term projection (in its current implementation) are sensitive only to the starting values of the stock, and to the estimates of recruitment in the following years. In general the influence of stock starting values decreases with time and depends on how long these fish remain in the population in the following years. A strong year-class in the starting stock will dominate the population for several years. The duration of the effect of this year-class in the medium-term will mainly depend on fishing mortality (F). If F is high, and the year-class is fished out quickly, the influence of the year-class will reduce quickly. If the input to a medium-term analysis is obtained from a biased assessment, the prognoses will be biased as well in those years where the biased year classes contribute to the predicted landings and stock.

The results in the later years of a medium-term projection are mainly sensitive to the assumed recruitment model. These results may differ depending on the choice of the stock-recruitment model (Shepherd, Beverton-Holt, Ricker, Ockham, etc.) Basing the stock-recruitment model fit on a limited subset of available data, which may be appropriate if there is evidence of a shift in *regime*, or the addition of an extra year to the data from which the relationship is derived may also change this relationship and the results of the medium-term analyses.

The presentation of medium-term projections was considered by the Working Group. In addition to the many methodological problems with these projections, the extreme fractiles are not accurately estimated and absolute levels of probability are not reliably estimated. Specifically, **the extreme percentiles (5^{th} and 95^{th} percentiles) of predicted SSB and catch cannot be considered to be reliable** – the 25^{th} and 75^{th} percentiles are better behaved but their use may be overly restrictive (Patterson *et al.,* 2000).

The present projection programs (WGMTERM, ICP) do not take into account stock specific temporal patterns of recruitment, which may lead to simulation runs with highly unlikely results. The results of the medium-term prognoses are of course dependent on the input values, weight-at-age, maturity-at-age, natural mortality and exploitation pattern. In the present implementation these values are assumed to be constant and no variability is assumed. The main variation in the results originates from variations in recruitment.

Deficiencies in the present implementations and procedures

A major problem is that a bias in the results of the stock assessment is transferred to the medium-term prognoses. This in particular, affects prognoses in the near future.

For some stocks, observations of weight-at-age indicate that they have changed over time, with or without trends. These biological parameters in the model are currently assumed to be constant with no error. This will underestimate the variability of the distributions of possible stock trajectories. Similarly, changes in maturity-at-age are related to changes in growth, and (as for weights-at-age) the present implementation is restricted to constant values over the whole projection time period.

A number of stocks show specific temporal patterns in recruitment, such as good year classes alternating with poor year classes, longer time periods of high recruitment alternating with periods of poor recruitment, frequently peaks in

recruitment, or occasional occurrence of large year classes. The patterns, in which these recruitments appear, determine the range and dynamics of which is specific for that particular stock. The present implementation does not take account for these patterns resulting in occasional unlikely results

On some occasions the results of medium-term projections will be different from the results in the short-term forecasts in the corresponding years. This arises because of differing assumptions in year-class estimates and weights-at-age, and the fact that medium-term projections are stochastic, short-term forecasts deterministic.

Current and on-going work

Recent work on medium–term projections has taken place under the auspices of an EC Concerted Action (PL98-4231), and of the ICES Study Group on the Incorporation of Process Information in Stock-Recruitment Models (SGPRISM). As part of the former, Patterson *et al.* (2000) tested the performance of three methods used within ICES to make medium-term projections and hence probabilistic statements about the likely outcomes of fishery management actions. Using data for a large number of stocks, they ran retrospective assessments removing the most recent eight years of data, then projected forward for five years, and then compared the distribution of the outcomes from the projections with *the truth* as estimated by an assessment using all data. They found that all methods tended to under-estimate the true uncertainty in the system, and further that some methods had a tendency to over-estimate the resultant SSB.

In relation to this result, Patterson *et al.* (2000) note that : "*We have not yet explored why the estimated uncertainty fails adequately to characterise real uncertainty in future stock sizes. This is a complex issue, as there is a large number of conditioning assumptions underlying each stock assessment, and it may be the case that a different conditioning assumptions may be inappropriate in different cases. It is also possible that factors exogenous to single-species assessment models may not be represented adequately in the stochastic simulation processes, such as long-term climatic effects introducing autocorrelations, multispecies effects introducing ecosystem changes, or data biases caused by inadequate catch reporting. However, it is extremely difficult to identify case-specific effects because of the few observations on any specific case*."

The possible effect of autocorrelation introduced by climatic effects (noted as a possible contributing factor to the under-estimation of uncertainty) has been addressed further by work done under the auspices of SGPRISM (ICES CM 2001/C:02). Working with North Sea Cod, SGPRISM found that in addition to variation in recruitment, variation in weight-at-age is also in important contributory factor to the uncertainty in stock projections. Recruitment in North Sea cod is thought to be related to temperature (O'Brien *et al*., 2000). The studies by SGPRISM did not demonstrate any improvement in the performance of the projections when a temperature term was included in the stock-recruitment model. However, it is possible that the observed changes in weight-at-age may be mediated by temperature, hence subsequent work has focussed on direct modelling of changes in weight-at-age and maturity (Needle *et al.,* 2000, 2001, 2002).

6.1.2 Recommendations for immediate adoption by stock assessment working groups

Relatively few improvements can be made to the present procedures in the immediate future: improving medium-term projections depends largely on additional analyses of data and reformulation of software, neither of which can be done particularly quickly. However, there is one specific area in which relatively rapid improvements could be made. In many cases, the choice of the stock-recruitment model to be used is arbitrary or driven by personal preference, when there is no statistical support for making a particular choice. Because of the requirement for stable medium-term analyses between years, it would be preferable to maintain a consistent choice of model from year-to-year, and furthermore to attempt to ensure that the model used is not overly sensitive to the addition of new data. To this end, **it is proposed that a series of candidate stock-recruitment models are fitted to historically-estimated stock-recruitment pairs, and that a final model is chosen based on consideration of statistical fit, parsimony, biological appropriateness, and robustness (including sensitivity to the addition of new data).**

6.1.3 Suggestions for future adoption

The Methods Working Group is of the opinion that medium-term analyses would be improved if they were based on unbiased or bias-corrected historical assessments. It is proposed that no provisions should be implemented in the medium-term prediction software to correct for the retrospective bias. The correction should be made in the estimation of the state of the stock and the medium-term analyses should ideally start with unbiased input values.

Future implementations of the medium-term software should take into account estimates of the variability of the input parameters, either as time-series or as random noise. Evaluation of historical data may indicate trends in biological

parameters such as in weight-at-age and maturity-at-age, and medium-term software should be adjusted to take account for these trends.

Needle (WP1) presented one example of how changes to medium-term projection methodology might be implemented. In this version, stock-recruitment residuals are modelled as autoregressive moving-average (ARMA) time-series, while cohort-based weights-at-age ogives are expressed as functions of hatch-date biomass (HDB) and subsequently modelled as vector autoregressive moving-average (VARMA). These approaches were generally accepted by the Working Group as being appropriate and potentially beneficial, although VARMA modelling of weights would require considerable further testing and may not be generally applicable. An attempt was made to model maturity-at-age as an age-specific linear function of weight-at-age, but this was less successful. The methodological changes discussed by Needle (WP1) are part of a series which has been developed under the remit of SGPRISM and the UK CFRD WG[1], and which is documented by Needle *et al*. (2000, 2001, in press). It is intended that functional software incorporating these and further developments will be available for ICES' certification by early 2003, and a work programme to this end is given in Section 7.3.3.

The Working Group recommends that the following studies be pursued inter-sessionally, both in an *ad hoc* manner and formally under the auspices of WGMG, SGPRISM (subsequently WGRP), and CFRD, and that additional elements be included as necessary:

- ARMA time-series modelling of recruitment is considered, *a priori*, to represent an improvement on the WGMTERM/ICP method of bootstrapping recruitment residuals. Of particular value would be the avoidance of unrealistic projection realisations in which several large year-classes are generated sequentially. However, to date ARMA models have only been tested on one stock (North Sea cod), and it is not clear that the improvement achieved through their use is universal. Hind-cast testing of competing projection methodologies must be carried out on a wide variety of stocks (and, potentially, simulated data) to evaluate whether a switch to time-series recruitment modelling is worthwhile.

- The choice of drivers for weights-at-age projections should be closely investigated for each specific stock, inter-sessionally if desired. Hatch-date biomass was used by Needle (WP1) as a source of density-dependent effects on subsequent year-class growth, but it is far from certain that this is the most suitable choice. Alternative candidates might include cohort numbers at recruiting or pre-recruiting ages, or hatch-date spawner numbers. Factors such as the abundance of predator or prey species during the lifetime of the cohort, or the fishing mortality it may have experienced, are also likely to be influential on mean weights-at-age, but may be more difficult to model successfully in projections.

- The values used for maturity in projections should also be analysed, but here even more care needs to be taken and there needs to be clear guidelines on the approach to take. For example, maturity-at-age is usually available only from survey data, whereas weights-at-age are taken principally from landings data and the two are therefore not completely comparable. The following Term of Reference is therefore proposed for the next meeting of the Methods Working Group: that **guidelines on the modelling of weights, maturity, and condition factors for both historical stock assessment and medium-term projections should be formulated.**

- **The ICES Fisheries Assessment Scientist should be encouraged to complete the study** (Sparholt WA5) **of the quality of ACFM advice for all stocks currently assessed using an analytical assessment.** This would serve two purposes. Firstly, to identify those stocks with the largest errors in the forecasts and secondly, allow ACFM (through its stock assessment working groups) to periodically up-date its perception of the quality of advice. If large errors are found in the forecasts then the root causes should be investigated; e.g. initial stock numbers-at-age, recruitment, weights-at-age, maturity-at-age, natural mortality, tuning data series, etc. It is suggested that this exercise should be not repeated every year.

- Current Working Group practice does not generally stipulate that quality control procedures should be carried out on medium-term projections, but there is an undoubted need for this. Substantial annual changes in starting population numbers-at-age and assumed stock-recruitment formulations lead to medium-term projections that vary widely from year to year. Plotting the projection from this year's assessment alongside that from last year's (and indeed, several years prior to that) would serve to highlight such variation, and focus Working Group attention on determining the reasons for it. Advice grounded in medium-term projections should become more stable and credible as a result. It should be noted that such comparisons would all have to be based on the same projected fishing mortality, so that

[1] Coordinator of Fisheries Research and Development (CFRD) Working Group on the Application of Recruitment Models in Stock Assessment.

there would probably be a need to re-run previous projections using the current *status quo F*. **The following term of reference is therefore proposed for the next meeting of the Methods Working Group: that quality control of medium-term projections should be investigated and implemented**.

– Developments in the medium-term methodology are resulting in improved algorithms for making projections of future outcomes. However, the reasons for the failure of the current methods, as highlighted by the EU concerted action (PL98-4231), have not been fully explored. It is recommended that further investigations into the underlying causes of the failure of the projection methods be carried out to determine potential areas of research.

6.2 Short-term forecasts

An ICES short-term forecast is a forward extension in time of the results of an historical stock assessment, examining what might be expected to occur under different assumptions of exploitation in the three years following the final historical year. Assumptions also have to made about recruitment, mean weight-at-age and maturity in these years. The prognoses give a prediction of the expected landings in the current and TAC year and resultant SSBs.

Short-term predictions are presented in the ICES' advice at the request of its clients, who use it to select a particular TAC. Management tools such as TACs are intended to control fishing mortality. In order to be able to control fishing mortality by means of a TAC, taking the TAC should result in a level of fishing mortality which does not exceed the required value.

The quality of short-term predictions largely depends on the quality of the assessment and assumed inputs. Short-term prognoses which use a biased assessment as a starting point will also be biased. Many stocks which are assessed by ICES are subject to high levels of exploitation, and expected yields largely depend on the success of incoming year-classes. In these cases short-term predictions are particularly sensitive to estimates of recent recruitment and factors associated with these year-classes such as weight-at-age and maturity.

The short-term forecast requires an assumption about the level of fishing mortality in the mid-year (the current year for projections), in order that population numbers can be projected forward to the start of the TAC year. This assumption is expressed as an F-multiplier, which will often have a value of 1, i.e. the assumption of *status quo* fishing mortality.

In a number of ICES forecasts, considerable deviations from the predicted catch in the mid-year have occurred. These reflect the need to make an assumption about F, and thus catch, in the current year which may be inconsistent with the actual catch or TAC during this year. This is referred to as the mid-year problem. This is very difficult to explain to managers and industry and causes criticism of the ICES advice. The mid-year problem can appear in the short term prognoses in two different forms: 1) when an F constraint is used, it appears as an anomaly in the expected catches and 2) when a catch constraint is used it appears in an anomaly in the expected fishing mortality. The mid-year problem undermines the credibility of, and support for, the ICES advice and therefore deserves considerable attention.

A number of factors can contribute to the mid-year problem, either individually or in combination. These factors include:

1) Bias in assessment.

2) Predictions for heavily exploited stocks depend on recruitment, so poor estimation of recruitment may be an important factor.

3) Comparisons of assumed weight-at-age and measured weight show that inappropriate assumptions about weight-at-age have caused considerable problems for some stocks

Comparison of the weight-at-age assumed in the short-term prognoses and the observed weight-at-age in the relevant year for a number of North Sea stocks show that the traditional procedure to estimate weight-at-age (average of the last 3 years) may contribute to the bias in the prediction in the short-term (and also in the medium-term) (Darby WA2, Darby WA3; see Appendices B and C, respectively). In periods where weight-at-age declines, the values used in the prediction are systematically over-estimated and vice versa.

The Working Group recommends the following, regarding the generation of short-term forecasts:

- Model assumptions for short-term forecasts should be considered more carefully than is currently the case. For example, Darby (WA2, WA3) has shown that observed weights-at-age for many North Sea stocks are strongly influenced by cohort-based year-class effects, with environmentally-driven year-effects being perhaps less important. Consistent declining or increasing trends in weights-at-age are also known for several stocks. With this in mind, any such influences in an extant cohort should be taken into account when generating weights-at-age for short-term forecasts. These comments are equally pertinent to fishing mortality-at-age, which is equally affected by time trends (both in terms of overall F and the selection pattern).

- It would be desirable for the forecasting software to be able to use the output from any available assessment package as starting inputs, and not be tied to one assessment method. It is thus assumed here that assessment and forecasting functions will be performed by separate programs. Given this, any bias correction that is thought to be necessary to estimated starting numbers-at-age for the forecasts should be undertaken within the assessment process rather than the forecast process, as retrospective bias is an assessment problem and is best dealt with in that context.

- The presentation of results from short-term forecasts should be made more probabilistic, in order to reflect assessment uncertainty with more clarity than is currently the case. An example of this would be the re-expression of forecasts in terms of the probability of fishing mortality F being within a range F_1 to F_2 for a given catch. This is more pertinent than producing a catch range for a given F, principally because terminal F is not well-enough defined to allow it to be specified with any precision.

- Working Groups should be encouraged to produce detailed catch forecast tables. These were generated automatically by the IFAP system, but with its recent demise the provision of detailed tables has become inconsistent.

- Managers often comment on the differences between short-term forecasts and the early part of medium-term projections. These result from differences in methodology (short-term forecasts are deterministic, medium-term projections are stochastic) and inputs (short-term forecasts use information on incoming year classes from surveys, medium-term projections use fitted stock-recruitment models). Looking further ahead, it would be beneficial to merge the functions of short-term forecast and medium-term projection software.

7 SOFTWARE TOOLS FOR STOCK ASSESSMENT PURPOSES

7.1 Certification of software

The ICES Fisheries Assessment Package (IFAP) has provided the major assessment tools used for the majority of the ICES stock assessments. However, the Study Group on Future Requirements for Fisheries Assessment Data and Software (SGFADS; ICES 1998/ACFM:9) reviewed existing ICES software and concluded that the speed of development of assessment methodology has been such that it has always outstripped the speed with which new techniques can be incorporated into IFAP. In order to provide a more flexible set of assessment tools, SGFADS proposed moving to a PC-based system. This would consist of standard programs, together with defined file formats for exchange of information between these programs. To ensure a degree of quality control and efficiency, programs would not be incorporated into the standard set unless they conformed to defined minimum standards of programming practice and documentation.

The Workshop on Standard Assessment Tools for Working Groups (ICES, 1999) proposed programming guidelines and an acceptance protocol for assessment software – testing of new methods is not part of this protocol! For new sophisticated analytical methods, it will be expected that ACFM will refer the task of scientific endorsement to the Working Group on Methods on Fish Stock Assessments [WGMG] which possesses the appropriate expertise. This group will undertake detailed testing of the method and recommend the context in which the method should be appropriately applied. However, the accompanying software may not meet appropriate documentation and coding standards, and the required ease of operation. In such cases the software cannot become part of the *standard* endorsed package of software.

ACFM will be responsible for endorsing assessment software.

WGMG endorsed the following proposal for the acceptance of a program as part of the standard ICES' tools. This proposal was initially presented at the May 2000 meeting of ACFM (BPS1) but WGMG have highlighted one addition and one deletion to the proposal in **bold**.

In order for a program to be accepted as part of the standard it would need to go through the following process.

The Secretariat must be provided with:

- *Documentation of the analytical method which gives a complete description of the approach.* ~~**This can be in the form of a peer-reviewed paper in the prime literature.**~~
- *Documentation of the program which gives sufficient information on how to install and run the program, and how to interpret the output.*
- *Documentation of the input and output files.*
- *The program source code.*
- *Example data sets to check that the program is running correctly***; together with test results for these example data sets.**
- *The Secretariat will check that the program installs and runs correctly on the ICES system; i.e. that the software gives the results claimed when applied to one or more test data. Such test data sets must accompany the software.*
- *If the software installs and runs under the ICES system then the ACFM chair will nominate two reviewers, preferably drawn from ACFM membership to check the software. The reviewers need to be satisfied that the program developer has undertaken adequate testing of the program and expect documentation from the developer outlining the testing which has been undertaken. Their written reviews together with the software and documentation will be forwarded to ACFM through the ACFM home page. ACFM will be invited to comment on the review and to conclude if the software meets sufficient quality standard and can be accepted.*

The above process cannot guarantee that every program is free of errors and it is expected that the program source code will be available at the Secretariat. This will enable working groups to check any problems against the source code.

The ICES Secretariat will maintain a library of endorsed *standard* software. These programs, together with documentation and source code, will be publicly available as part of the transparency requirements for the stock assessment work. Commercial code under copyright will not be in the public domain.

Endorsed programs will be the preferred tools to be used by stock assessment working groups. Other tools will only be used if the standard program does not provide the necessary method and working groups will need a strong justification for using such tools. As a minimum, any non-standard method should be used in addition to one of the standard tools so that results can be compared.

After inclusion in the *standard* package it is likely that programs will need some support, particularly where errors need to be fixed. It is the responsibility of the developer to fix errors. Where an error comes to light and is corrected, this will need to be recorded in the program documentation at the Secretariat and a new version of the program identified in the standard library. However, this will not require a renewed certification process.

Where a substantial revision or update is introduced to a program it will be necessary for the new revision to undergo the same acceptance protocol as the original program. However, depending on the nature of the change, ACFM may identify a simpler endorsement procedure than that originally carried out.

In addition**, it is envisaged that there will also be standard programs to pick-up output from these programs and produce standard tables and plots for inclusion in Working Group and ACFM reports. However, as these would be presentational rather than analysis tools WGMG considered it appropriate that such programs be developed by the ICES Secretariat.**

7.2 Comments to ICES on software development and maintenance

A notable feature about the software used by ICES stock assessment working groups is that ICES has little to do with its development and maintenance, but instead relies on national laboratories to carry out these tasks. This is problematic, in that ICES has no direct influence on the tools through which much of its advice is provided. The situation is unlikely to change, but certain small alterations in the way ICES manages assessment software could make a great difference to the quality and transparency of assessments. In this vein, **WGMG recommends that a formal feed-back mechanism be created, through which those scientists with primary responsibility for the maintenance of a particular item of software could be made aware of any problems** individual stock assessment working groups have had in using it, or suggestions for future developments.

In addition, **further improvements to the traceability of assessments could be achieved if assessment runs were correctly labelled and logged**, as used to be the case automatically with IFAP. This would enable runs to be repeated if necessary, and could reduce the time taken by new assessment scientists to become familiar with the accepted methodology.

7.3 Software development of stock assessment tools

WGMG endorsed the further development of TSA into a usable FORTRAN 90 subroutine by FRS Aberdeen, **the likelihood-based development to XSA** by CEFAS Lowestoft **and the development of a new methodological tool for medium-term projections (MedAn)** by FRS Aberdeen. Each of these developments are discussed in the subsequent sections of this Section 7.3.

7.3.1 Time series analysis (TSA)

The TSA (Time Series Analysis) stock assessment methodology of Gudmundsson (1994) has been discussed by several ICES Methods Working Groups, where its performance has been shown to compare well with other stock assessment methods. Recently, a new version of TSA has been implemented. This implementation is based on Gudmundsson's work, but also allows for the modelling of landings-at-age and discards-at-age separately. It has been used to assess VIa cod, haddock, and whiting, and North Sea whiting. WS3 (reproduced as Appendix D of this report) gives a technical description of the new implementation, illustrates the method, and discusses some of its strengths and weaknesses.

In summary, the main strengths are:

- genuinely models the time-series structure of a fishery, allowing; e.g. numbers-at-age, fishing mortalities-at-age, and discards-at-age to evolve forward in time,
- gives the precision of estimates of numbers-at-age, fishing mortalities-at-age, and stock trends. This helps to avoid the over-interpretation of small recent changes in stock trends, and
- can be tailored to an individual fishery, depending on the auxiliary information available.

and the main weaknesses are:

- Very hard to code. This will make it difficult to provide easy-to-use software that is sufficiently flexible to realise the full potential of TSA.
- Favours *status quo*. If the data are not of sufficient quality, sudden changes in the fishery, such as a stock collapse, might not be picked up straight away.
- Requires linear approximations to non-linear equations and assumes normal error structure (although constant variance is not assumed). In practice, this does not cause difficulties with model fitting, but might pose problems with predictions.

It is hoped to make a fully documented FORTRAN 90 subroutine available within the next year that will fit a standard class of TSA models. This would include:

- fitting to catch-at-age data;
- fitting to landings-at-age data and discards-at-age data;
- incorporation of multiple surveys;
- flexible stock-recruitment module;
- flexible survey selectivity module;
- standard ICES output; and
- standard errors or profile likelihood intervals on parameter estimates.

This subroutine will then be tested prior to ICES' certification.

7.3.2 Extended Survivors Analysis (XSA)

The XSA program is currently undergoing development in order to evaluate the potential for improvements to the estimates and variances derived from its use. Many of these developments result from the findings of recent EU-funded studies (Concerted Action FAIR project PL98-4231 and CFP Study Project 98/175 EMAS). A list of the modifications currently being evaluated is given below:

1. Bootstrap algorithms for deriving standard errors of the parameter estimates, covariance matrices, methods of bias correction etc.
2. Comparison of the application of a non-linear search algorithm to find terminal population estimates with the current iterative methodology.
3. An analysis of the sensitivity and bias in the estimated values to the current assumption of a log-normal distribution for catchability.
4. Independent application of the shrinkage constraint across years and ages.
5. Variable q model by fleet.

The development work will continue during the coming year and progress will be reported at the next meeting of this Working Group. Acceptable developments to the program will then be tested prior to ICES' certification.

7.3.3 Medium-term analyses (MedAn)

A new methodology (MedAn) has recently been proposed (under the specific aegis of the UK CFRD Working Group on the Application of Recruitment Models in Stock Assessment) which takes aspects of the approach described in Needle (WP1) and develops them further. It is intended to replace the WGMTERM software for medium-term projections. The development plans for this approach to medium-term projection software are:

- Recruitment residuals could be modelled using ARMA time-series as before, although further testing of the applicability of this will be required. A further development would be that, for each projection, only a subsample of the historical stock-recruitment pairs will be used. Several different model formulations would be fitted to these data, and the best fitting one used to generate projections. Thus, uncertainty about the form of the recruitment model would be modelled, as well as variation about it.

- It is proposed that weights at the recruiting age would still be projected as residuals to a function of some causal driver (the identity of which would be stock-specific), but subsequent ages may be better modelled in terms of increments down cohorts. This is more logical from a biological-process point of view, and also allows for year effects as well as year-class effects. In addition, it would avoid the problem of data from incomplete cohorts, as these could be used directly. There still remains the thorny issue of whether to smooth weights down cohorts before modelling increments.

- Maturity could be modelled in one of two ways, either by a single ogive with weights as the independent variable (thus ignoring age), or by empirical non-parametric kernel distributions.

- For many stocks, spawning-stock biomass is known to be an unsuitable proxy for egg production. The new method may (if possible and appropriate) base stock-recruitment modelling on potential egg production rather than spawning-stock biomass, and will thus be better able to mimic the deleterious effect of a compressed age structure.

- The illustrative examples presented in this paper all assume fixed starting values for population numbers-at-age, fishing mortalities, and so on. The proposed new model should include the facility to accept a different realisation of starting values for each projection iteration, as would be produced by a bootstrapped assessment model. This would lead to a better representation in the projections of the variance-covariance structure of the starting values. If this is available, then the need for sub-sampling of stock-recruitment data would be removed.

- A key justification for work on medium-term projections is to allow managers to determine the likely responses of the stock in question to specified management actions. To this end, the methodology will allow for future imposed changes in fishing effort, gear selectivity, and catch constraints.

Extensive hindcast model testing would be carried out on all these aspects. Beta-test versions should be circulated prior to formal evaluation. The method should be implemented to be as general as possible, whilst allowing for additional model structure if data are available. It is envisaged that functional software incorporating these features, and others as need dictates, will be available during the first quarter of 2003 in time for ICES' certification and subsequent use of the software by ICES stock assessment working groups.

8 RECOMMENDATIONS AND FURTHER WORK

WGMG had started to address its terms of reference at this first meeting since February 1995, and the group felt it had been both a useful and stimulating forum for discussion. However, much work still remained to be undertaken and the

group recommends that a second meeting should be held – the suggested terms of reference for which are given in Section 8.2.

The group has made a number of suggestions and recommendations on issues of data quality, modelling and stock assessment practice throughout this report and these have been highlighted in the text. However, it was felt that due to the importance of some of the points raised that these should be collated together and presented in the next Section 8.1.

The group had focused on the urgent issue of the retrospective problem in stock assessments but it could be anticipated, in advance of the meeting, that the problems of ICES' assessments would not be fixed at short notice. It has, however, become clearer as to the likely causes of the problems and a way to proceed in the development of a solution has been proposed.

8.1 Suggestions and recommendations

Section 4. Data quality

Sub-section 4.2.1. Survey series used for 'tuning'

Assessment procedures could model surveys separately without the need to calculate correct conversion factors when changes in vessel or design take place.

Sub-section 4.2.2. Commercial CPUE data

This Group recommends that effort data be corrected for changes in efficiency by specific analyses prior to setting up the tuning data for assessment. This may involve *a priori* standardisation of effort with, for example, tonnage or engine power, but such an approach may not be appropriate in all cases and should be used only when justified on the basis of analyses.

Sub-section 4.4. Conclusions

A message to assessment working groups is that they should favour fewer data of good quality (as evaluated independently of the assessment model) instead of large quantities of data of unknown properties.

Retrospective patterns should not be taken as the only diagnostic of problems in assessments. Consideration should also be given to all other assessment diagnostics.

Sub-section 4.5. A final word

The definition of fleets for tuning purposes should be improved, and stricter criteria should be used to select the catch and effort data retained for each fleet.

Section 5. Numerical and statistical aspects of fisheries models

Sub-section 5.1. Introduction

WGMG recognises the need to:

- understand the mechanisms which create inconsistencies in the perception of the development of the stock
- investigate the sensitivity to various causes of the retrospective bias of different model formulations, and the extent to which such causes can be accounted for in model structures
- develop diagnostics, and to understand the extent to which problems can be revealed by the diagnostic tools.

Sub-section 5.3.1. Residual patterns in catchability

It is recommended that modelling data sets in order to detect departures from model assumptions should take place prior to the fitting of an assessment model and using data that are independent of the assessment information.

It is recommended that analysis of survey series residuals should be given a high priority during the fitting of assessment models, even if a retrospective pattern is not apparent in the time series of assessment estimates. This analysis should take place over the whole of the available time series, not only for the most recent data. Where surveys show transitions in catchability careful consideration should be given to the underlying cause and the quality of the catch data and any commercial tuning series.

Sub-section 5.3.2. Local influence diagnostics

The local influence method could be used to find perturbations of VPA inputs that remove or reduce the retrospective problem. The method may be useful for identifying a smaller subset of the inputs that are more likely causes of the retrospective problem.

The working group recommends that influence diagnostics should be developed for routine use within stock assessments, addressing both data and modelling issues. It is further recommended that such methods be applied to specific case studies to examine their potential for analysing retrospective problems.

Sub-section 5.3.4. Bounding the scale of the possible causes of the retrospective pattern

Stock assessment working groups should be encouraged to provide bounds for the changes to the assessment data that would be required to remove retrospective patterns, if present.

Sub-section 5.4. Conclusions

WGMG considers that there is a need for extensive studies with data sets with known properties, and has outlined a framework for making such studies with simulated data.

Section 6. Population forecasting

Sub-section 6.1. Medium-term projection

The extreme percentiles (5^{th} and 95^{th} percentiles) of predicted SSB and catch cannot be considered to be reliable.

Sub-section 6.1.2. Recommendations for immediate adoption by stock assessment working groups

It is proposed that a series of candidate stock-recruitment models are fitted to historically-estimated stock-recruitment pairs, and that a final model is chosen based on consideration of statistical fit, parsimony, biological appropriateness, and robustness (including sensitivity to the addition of new data).

Sub-section 6.1.3. Suggestions for future adoption

The Working Group recommends that the following studies be pursued inter-sessionally:

- ARMA time-series modelling of recruitment
- The choice of drivers for weights-at-age projections should be investigated
- The values used for maturity in projections should be analysed

Guidelines on the modelling of weights, maturity, and condition factors for both historical stock assessment and medium-term projections should be formulated.

The ICES Fisheries Assessment Scientist should be encouraged to complete a study of the quality of ACFM advice for all stocks currently assessed using an analytical assessment.

The following term of reference is proposed for the next meeting of the Methods Working Group: that quality control of medium-term projections should be investigated and implemented.

Sub-section 6.2. Short-term forecasts

The Working Group recommends the following, regarding the generation of short-term forecasts:

- Model assumptions for short-term forecasts should be considered more carefully than is currently the case
- Detailed catch forecast tables should be produced that are similar to those previously generated by the IFAP system
- The presentation of results from short-term forecasts should be made more probabilistic
- Looking further ahead, it would be beneficial to merge the functions of short-term forecast and medium-term projection software

Section 7. Software tools for stock assessment purposes

Sub-section 7.1. Certification of software

WGMG endorsed the following proposal for the acceptance of a program as part of the standard ICES' tools. This proposal was initially presented at the May 2000 meeting of ACFM but WGMG have highlighted one addition and one deletion to the proposal in **bold**.

In order for a program to be accepted as part of the standard it would need to go through the following process.

The Secretariat must be provided with:

- *Documentation of the analytical method which gives a complete description of the approach.* ~~**This can be in the form of a peer-reviewed paper in the prime literature.**~~
- *Documentation of the program which gives sufficient information on how to install and run the program, and how to interpret the output.*
- *Documentation of the input and output files.*
- *The program source code.*
- *Example data sets to check that the program is running correctly***; together with test results for these example data sets.**
- *The Secretariat will check that the program installs and runs correctly on the ICES system; i.e. that the software gives the results claimed when applied to one or more test data. Such test data sets must accompany the software.*
- *If the software installs and runs under the ICES system then the ACFM chair will nominate two reviewers, preferably drawn from ACFM membership to check the software. The reviewers need to be satisfied that the program developer has undertaken adequate testing of the program and expect documentation from the developer outlining the testing which has been undertaken. Their written reviews together with the software and documentation will be forwarded to ACFM through the ACFM home page. ACFM will be invited to comment on the review and to conclude if the software meets sufficient quality standard and can be accepted.*

It is envisaged that there will also be standard programs to pick-up output from these programs and produce standard tables and plots for inclusion in Working Group and ACFM reports. However, as these would be presentational rather than analysis tools WGMG considered it appropriate that such programs be developed by the ICES Secretariat.

Sub-section 7.2. Comments to ICES on software development and maintenance

WGMG recommends that a formal feed-back mechanism be created, through which those scientists with primary responsibility for the maintenance of a particular item of software could be made aware of any problems.

Further improvements to the traceability of assessments could be achieved if assessment runs were correctly labelled and logged.

Sub-section 7.3. Software development of stock assessment tools

WGMG endorsed the further development of TSA into a usable FORTRAN 90 subroutine by FRS Aberdeen, the likelihood-based development to XSA by CEFAS Lowestoft and the development of a new methodological tool for medium-term projections (MedAn) by FRS Aberdeen.

8.2 Future terms of reference

The Working Group on Methods on Fish Stock Assessments [WGMG] meet for 8 days during January 2003 (Chair: C. O'Brien, UK) at ICES Headquarters, Copenhagen, Denmark to:

a) develop influence diagnostics for routine use within stock assessments, addressing both data and modelling issues;

b) investigate and test the sensitivities of catch-at-age stock assessment methods to known data problems with particular reference to the retrospective problem;

c) develop and investigate techniques (e.g. Benford's Law) that detect inconsistencies in the data sources currently used by ICES' stock assessments;

d) investigate and implement quality control procedures for medium-term projections;

e) evaluate the reports from the recently funded EU studies of the performance of multi-annual management strategies for flatfish, roundfish and pelagics to identify generic tool components that WGMG can develop;

f) review the software developments in TSA, XSA, MedAn and other assessment methods that are presented to ICES;

g) discuss the choice of model structure (VPA/CAGEAN, age/age and length/length based models, single-/multi-species) taking into account stock dynamics, biology and data availability; and

h) address any ad-hoc requests from ACFM.

WGMG should report for the attention of the Resource Management Committee, the Living Resources Committee and ACFM.

9 WORKING DOCUMENTS AND BACKGROUND MATERIAL PRESENTED TO THE WORKING GROUP

9.1 Working papers and documents (W)

Applications (A)

WA1: Reeves, S. and Hovgård, H. On the retrospective problem in the assessment of Eastern Baltic cod.

WA2: Darby, C. Over-estimation bias in the North Sea cod short-term forecasts.

WA3: Darby, C. Estimation bias in the North Sea short-term forecasts.

WA4: van Beek, F. and Pastoors, M. Bias in stock assessments: consequences for short-term and medium-term prognoses from a practical point of view.

WA5: Sparholt, H.. Quality of ACFM advice: How good have forecasts been since 1988?

WA6: Duarte, R., Murta, A., Azevedo, M. and Cardador, F. Violating the constant catchability assumption: can XSA cope with it?

Data quality (D)

WD1: Mesnil, B. "Biased" estimates of stock size: mostly a problem of method or of data?

WD2: Mahévas, S. Quantification of fishing power explained by differences with technical characteristics: the bottom-trawlers of south-Brittany targeting monkfish from 1983 to 1998.

Stock projections (P)

WP1: Needle, C.L. ARMA and VARMA models in medium-term projections.

Recruitment (R)

WR1: O'Brien, C.M. and Maxwell, D.L. Stock-recruitment modelling based upon a segmented regression approach – work in progress

Methods of stock assessment (S)

WS1: Gavaris, S. Some comments on Terms of Reference a).

WS2: Darby, C. Preliminary results of a simulation study into XSA estimation bias.

WS3: Fryer, R. TSA: is it the way?

WS4: Skagen, D.W. Some possible causes of 'retrospective bias' in an ICA-like assessment model.

Uncertainty (U)

WU1: Darby, C. Is a perfect retrospective pattern correct?

WU2 : Darby, C. A comparison of retrospective bias in stock assessment estimates derived from XSA and ICA analyses.

WU3: Vázquez, A. and Cerviño, S. Covariance among results from a sequential population analysis.

9.1 Background material (B)

Applications (A)

BA1: van Beek, F. (2001). North East Arctic cod. Working document presented to ACFM May 2001.

BA2: van Beek, F. (2000). A note on the working group performance of short term predictions for cod, plaice and sole. WD-WGNSSK2000-2.

Data quality (D)

BD1: O'Brien, C.M., Darby, C.D., Maxwell, D.L., Rackham, B.D., Degel, H., Flatman, S., Pastoors, M.A., Simmonds, E.J. and Vinther, M. (2001). The precision of international market sampling for North Sea plaice (*Pleuronectes platessa* L.) and its influence on stock assessment. ICES CM 2001/P:13.

BD2 : O'Brien, C.M., Darby, C.D., Rackham, B.D., Maxwell, D.L., Degel, H., Flatman, S., Mathewson, M., Pastoors, M.A., Simmonds, E.J. and Vinther, M. (2001). The precision of international market sampling for North Sea cod (*Gadus morhua* L.) and its influence on stock assessment. ICES CM 2001/P:14.

Environment (E)

BE1: Needle, C.L., O'Brien, C.M., Darby, C.D. and Smith, M.T. (2000). The use of recruitment time-series structure and environmental information in medium-term stock projections. ICES CM 2000/V:05.

BE2: O'Brien, C.M. (2001). Cod and climate variability. Paper presented at Marine Conservation Society *Annual Conference for Recreational Sea Anglers*, University of Cardiff, 12th May 2001.

Fishing mortality (F)

BF1: van Beek, F.A. and Pastoors, M.A. (1999). Evaluating ICES catch forecasts: the relationship between implied and realised fishing mortality. ICES CM 1999/R:04.

BF2: Cook, R. (2000). A rough guide to population change in exploited fish stocks. *Ecology Letters*, **3**:394-398.

Management strategies (M)

BM1: Multi-annual TACs (MATAC): An analysis of the possibilities of limiting annual fluctuations in TACs for flatfish. Executive Summary of FISH/2000/02.

Computer programs and software (PS)

BPS1: ICES (2000). Certification of software used for assessment purposes. Working document presented to ACFM May 2000.

BPS2: ICES (2000). Certification of software used for assessment purposes. Extract from minutes of ACFM May 2000.

Recruitment (R)

BR1: O'Brien, C.M. (1999). An approach to stock-recruitment modelling based upon GLMs, HGLMs and DLMs. ICES CM 1999/T:01.

BR2: Bravington, M.V., O'Brien, C.M. and Stokes, T.K. (1999). Sustainable recruitment: the bottom line. ICES CM 1999/P:01.

Methods of stock assessment (S)

BS1: Cadigan, N.G. and Farrell, P.J.. Generalized local influence with applications to fish stock cohort analysis.

BS2 : Cadigan, N.. Estimation and inference for a simple "SPA-like" model.

Uncertainty (U)

BU1: Patterson, K.R., Cook, R.M., Darby, C.D., Gavaris, S., Kell, L.T., Lewy, P. Mesnil, B., O'Brien, C.M., Punt, A.E., Restrepo, V.R., Skagen, D.W. and Stefánsson, G. (2000). Final Report of EU Concerted Action FAIR PL98-4231: Evaluation and comparison of methods for estimating uncertainty in harvesting fish from natural populations. Research Report 4 of a meeting held at Institute of Marine Science, Reykjavik, Iceland, 28-30 August 2000.

BU2: Gavaris, S., Patterson, K.R., Darby, C.D., Lewy, P., Mesnil, B., Punt, A.E., Cook, R.M., Kell, L.T., O'Brien, C.M., Restrepo, V.R., Skagen, D.W. and Stefánsson, G. (2000). Comparison of uncertainty estimates in the short term using real data. ICES CM 2000/V:03.

BU3 : Restrepo, V.R., Patterson, K.R., Darby, C.D., Gavaris, S., Kell, L.T., Lewy, P., Mesnil, B., Punt, A.E., Cook, R.M., O'Brien, C.M., Skagen, D.W. and Stefánsson, G. (2000). Do different methods provide accurate probability statements in the short term? ICES CM 2000/V:08.

10 REFERENCES

Aglen, A. (1994). Sources of error in acoustic estimation of fish abundance. In A. Fernø and S. Olsen (eds.): *Marine Fish Behaviour Related to Capture and Abundance Estimation*. Fishing News Books, Oxford. Pp.107-133.

Cadigan N.G. and Myers, R.A. (2001). A comparison of gamma and Log normal maximum likelihood estimators in a sequential population analysis. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**: 560:567.

Cook, R.D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society, series B*, **48**: 133-169.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Gavaris S. (1988). An adaptive framework for the estimation of population size. CAFSAC (Canadian Atlantic Fisheries Science Advisory Committee) Research Document 88/99.

Godø, O.R., Walsh, S.J. and Engås, A. (1999). Investigating density-dependent catchability in bottom-trawl surveys. *ICES Journal of Marine Science*, **56**: 292-298.

Gudmundsson, G. (1994). Time series analysis of catch-at-age observations. *Applied Statistics*, **43**:117-126.

Harley, S.J., Myers, R.A. and Dunn, A. (2001). Is catch-per-unit-effort proportional to abundance? *Canadian Journal of Fisheries and Aquatic Sciences*, **58**:1760-1772.

Hilborn, R. and Walters, C.J. (1992). *Quantitative Fisheries Stock Assessment. Choice, Dynamics and Uncertainty*. Chapman & Hall, New York.

Hinrichsen, R.A. (2001). The importance of influence diagnostics: examples from Snake River chinook salmon spawner-recruit models. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**: 551-559.

Huse, I. and Ona, E. (1996). Tilt angle distribution and swimming speed of overwintering Norwegian spring spawning herring. *ICES Journal of Marine Science*, **53**: 863-873.

Hutton, T., Casey, J. and O'Brien, C.M. (2001). Stock assessment of cod in the North Sea (IV), including Skagerrak (IIIa) and the Eastern Channel. Annex 1 to the Report of the Working Group on Fishery Systems, 12-15 June 2001. ICES CM 2001/D:06, pp.41-47.

ICES (1984). Report of the Working Group on Methods of Fish Stock Assessment, ICES Headquarters, Copenhagen, Denmark, 20-26 May 1983. *Cooperative Research Report,* **129**, pt II.

ICES (1991). Report of the Working Group on Methods of Fish Stock Assessments, St. John's, Newfoundland, 20-27 June 1991. ICES CM 1991/Assess:25.

ICES (1992). Report of the Workshop on the Analysis of Trawl Survey data, Woods Hole, USA, 4-9 June 1992. ICES CM 1992/D:6.

ICES (1993a). Reports of the Workshop on Methods of Fish Stock Assessment, Reykjavik, 6-12 July 1988, and of the Working Group on Methods of Fish Stock Assessment, Nantes, 10-17 November 1989. *Cooperative Research Report,* **191**, pt II & III.

ICES (1993b). Report of the Working Group on Methods of Fish Stock Assessment, ICES Headquarters, Copenhagen, Denmark, 3-10 February 1993. ICES CM 1993/Assess:12.

ICES (1995). Report of the Working Group on Methods of Fish Stock Assessment, ICES Headquarters, Copenhagen, Denmark, 6-14 February 1995. ICES CM 1995/Assess:11 Ref.:D.

ICES (1997). Report of the Comprehensive Fishery Evaluation Working Group, ICES Headquarters, Copenhagen, Denmark, 25 June – 4 July 1997. ICES CM 1997/Assess:15.

ICES (1999a). Report of the Working Group on the Assessment of Mackerel, Horse Mackerel, Sardine and Anchovy, ICES Headquarters, 28 September – 7 October 1998. ICES CM 1999/ACFM:6.

ICES (1999b). Report of the Workshop on Standard Assessment Tools for Working Groups, Aberdeen, United Kingdom, 3-5 March 1999. ICES CM 1999/ACFM:25.

ICES (2000a). Report of the Study Group on Incorporation of Process Information into Stock-Recruitment Models, Lowestoft, UK, 23-26 November 1999. ICES CM 2000/C:01.

ICES (2000b). Report of the Study Group on Market Sampling Methodology, Marine Laboratory, Aberdeen, Scotland, 24-25 January 2000. ICES CM 2000/D:01.

ICES (2001a). Report of the Workshop on International Analysis of Market Sampling and the Evaluation of Raising Procedures and Data-Storage (Software), CEFAS, Lowestoft, 28-30 November 2000. ICES CM 2001/D:02.

ICES (2001b). Report of the Study Group on the Incorporation of Process Information into Stock-Recruitment Models, Lowestoft, UK, 23-26 January 2001. ICES CM 2001/C:02 Ref.:D.

ICES (2001c). Report of the Working Group on Fishery Systems, ICES Headquarters, Copenhagen, Denmark, 12-15 June 2001. ICES CM 2001/D:06 Ref.:ACFM.

ICES (2001d). Report of the Baltic Fisheries Assessment Working Group, Gdynia, Poland, 18-27 April 2001. ICES CM 2001/ACFM:18.

ICES (2002). Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak, Hamburg, Germany, 19-28 June 2001. ICES CM 2002/ACFM:01.

Mohn, R. (1999). The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. *ICES Journal of Marine Science*, **56**: 473-488.

Morgan, M.J. and Brodie, W.D. (2001). An exploration of virtual population analyses for Divisions 3LON American plaice. NAFO SCR Doc. 01/04.

Needle, C. L., O'Brien, C. M., Darby, C. D. and Smith, M. T. (2000) The use of recruitment time-series structure and environmental information in medium-term stock projections. ICES CM 2000/V:05.

Needle, C. L., O'Brien, C. M. and Darby, C. D. (2001) Ogive characterisations in medium-term stock projections. ICES CM 2001/V:19.

Needle, C. L., O'Brien, C. M., Darby, C. D. and Smith, M. T. (2002). Incorporating time-series structure in medium-term stock projections. *Scientia Marina* (in press).

O'Brien, C.M., Fox, C.J., Planque, B. and Casey, J. (2000). Climate variability and North Sea cod. *Nature*, **404**: 142.

Patterson, K.R., Cook, R.M., Darby, C.D., Gavaris, S., Kell, L.T., Lewy, P. Mesnil, B., O'Brien, C.M., Punt, A.E., Restrepo, V.R., Skagen, D.W. and Stefánsson, G. (2000). Final Report of EU Concerted Action FAIR PL98-4231: Evaluation and comparison of methods for estimating uncertainty in harvesting fish from natural populations. Research Report 4 of a meeting held at Institute of Marine Science, Reykjavik, Iceland, 28-30 August 2000.Skagen, D.W., Stefánsson, G. and Smith, M. (2000). Validating three methods for making probability statements in fisheries forecasts. ICES CM 2000/V:06.

Reeves, S.A. (2001). The implications of age-reading errors for stock assessment and management advice: a case-study based on Eastern Baltic Cod. ICES CM 2001/P:18.

Restrepo, V.R., Patterson, K.R., Darby, C.D., Gavaris, S., Kell, L.T., Lewy, P., Mesnil, B., Punt, A.E., Cook, R.M., O'Brien, C.M., Skagen, D.W. and Stefánsson, G. (2000). Do different methods provide accurate probability statements in the short term? ICES CM 2000/V:08.

Rivard, D. (1989). Overview of the systematic, structural, and sampling errors in cohort analysis. *Amer. Fish. Soc. Symp*., **6**: 49-65.

Shelton, P.A. and Lilly, G.R. (1998). Interpreting the collapse of the northern cod stock from survey and catch data. *Canadian Journal of Fisheries and Aquatic Sciences,* **57**: 2230-2239.

Shepherd, J.G. (1999). Extended survivors analysis: an improved method for the analysis of catch-at-age data and abundance indices. *ICES Journal of Marine Science*, **56**: 584-591.

Simmonds, J. and Rivoirard, J. (2000). Vessel and day/night effects in the estimation of herring abundance and distribution from the IBTS surveys in the North Sea. ICES CM 2000/K:32.

Sinclair, A., Gascon, D., O'Boyle, R., Rivard, D. and Gavaris, S. (1990). Consistency of some northwest Atlantic groundfish stock assessments. NAFO SCR Doc. 90/96.

Sinclair, A., Gascon, D., O'Boyle, R., Rivard, D. and Gavaris, S. (1991). Consistency of some northwest Atlantic groundfish stock assessments. NAFO Scientific Council Studies, **16**:59-77.

Stansbury, D.E., Shelton, P.A., Murphy, E.F., Lilly, G.R. and Brattey, J. (1999). An assessment of the cod stock in NAFO Divisions 3NO. NAFO SCR Doc. 99/62.

Stratoudakis, Y., Fryer, R.J., Cook, R.M. and Pierce, G.J. (1999). Fish discarded from Scottish demersal vessels: estimators of total discards and annual estimates for targeted gadoids. *ICES Journal of Marine Science*, **56**: 592-605.

APPENDIX A – WORKING DOCUMENT WA5

Quality of ACFM advice: How good have forecasts been since 1988?

by
Henrik Sparholt

ABSTRACT:

The forecasted SSB surviving the "TAC year" given the realised catch in the "TAC year", are compared with the realised SSB as estimated from the latest assessments for six fish stocks, for which ICES provides annual advice on catch options. All six stocks are among the best known stocks in terms of biology and stock dynamics and they are well monitored each year, some of them several times a year. Thus, the quality of the advice on these stocks is supposed to be among the best in ICES. Generally, the forecasts made by ICES have been very imprecise. They have been wrong by a factor of two or more in 20 (3 too small by a factor of 2 and 17 too large) out of 67 assessments for the six stocks analysed here. In three assessments the SSB was overestimated by a factor of about 4. A clear bias towards overestimating SSB is apparent in 5 of the stocks. The precise reasons for the poor quality are generally not known. According to presented estimates of CV the "magic formula" used by ACFM to get $\mathbf{B}_{pa}$ and $\mathbf{F}_{pa}$ from $\mathbf{B}_{lim}$ and $\mathbf{F}_{lim}$ should be changed from the present values of 1.4-1.6 to a general value of at least 2.4 or to (generally high) stock specific values.

Introduction

On the assumption that a converged VPA is a precise description of the past history of the dynamic of a given fish stock, the ICES advice can be evaluated retrospectively.

The ICES advice is based on short-term forecasts, which from estimates of surviving population size by age of the last data year (as well as other biological parameters like weight at age, maturity at age etc.), and assumption about the fishery in the "current year" (year of assessment), forecasts catch and SSB given fishing mortality for the next year ("TAC year"), the so-called "catch option table". Given some relatively well defined rules (ACFM decision diagram, ACFM Meeting May and October 2001) and defined PA reference points, the advice about TAC is based on keeping F below a certain level and keeping SSB surviving the "TAC year", above a certain level. Precision in the forecasts is therefore crucial for the quality of the ICES advice.

The present paper considers the precision in the short-term forecasts made in the period 1988-2000, for six of the major fish stocks assessment annually by ICES: North Sea cod, North Sea plaice, North Sea sole, North Sea herring, central Baltic (Sub-divisions 25-32) cod, and Northeast Arctic cod. These stocks are among the best known stocks in terms of biology and stock dynamics and they are well monitored each year, some of them several times a year. Thus, the quality of the advice on these stocks is supposed to be among the best in ICES.

Several methods have in the past been applied to evaluate the precision in the ICES advice. Brander (1987) compared the catch predicted with the actual catch. He concluded that the "current year" forecast is better than the "TAC year" forecast. Cook et al. (1991) made sensitivity analysis on the procedures for forecasting yield and SSB for North Sea cod and found that yield is most sensitive to estimates of recruitment, while SSB is most sensitive to mortality rates. Both authors found relatively small CVs - in the order of 10-20% for forecasted yield and SSB. Mohn (1999) considered retrospective patterns as has been observed in a number of assessments in both ICES and in USA and Canada. He used simulated data to investigate the influence of errors in assumptions about catchability, natural mortality, discards, and partial recruitment. He found that for the stock he considered, cod in eastern Scotian Shelf, retrospective problems could lead to catch level advice that would be twice or more the intended level. In an analysis made be van Beek and Patoors (1999) it was found that the expected fishing mortality associated with the realised catches did not show any relationship to the realised F. Pattersson *et al*. (2000) validated three methods XSA/WGMTERM, ICA/ICP and a stochastic projection methods that was first applied to North Sea herring. By comparing the frequencies of expected and actual outcomes, they concluded that some methods can be used to make reliable probability statements about the relative biomass, but that for most other quantities and the accuracy of the probability estimates is very poor.

The method used in the present paper is largely the same as the method used by Brander (1987), but is focusing on precision of the SSB forecasted to be surviving the "TAC year", taking into account the catch actually taken in the

"TAC year". The reason for focusing on SSB is that SSB has become more important for the advice after the precautionary approach has been adopted.

A few examples of what the advice actually should have been in some selected years are also given and compared with the advice that was actually given at the time.

Material

"True" historical stock sizes are obtained from ACFM 2001 assessments. Forecasts are taken from ACFM reports from 1988-2000. For calculating what the advice actually should have been in some selected years, stock data are taken from ICES assessment wgs 2001/2002.

Method

The performance parameter: SSB surviving the "TAC year", is taken from the short-term forecasts, e.g. for short-term forecasts made in 1999 with advice for TAC in 2000, SSB(2001) is the parameter to consider.

For each assessment year the forecast tables from the ACFM reports are considered. A new option has been calculated, which corresponds to the catch actual taken in the forecast year.

Example:

The table below is from the ACFM 1999 report, with a row added, namely one with on option corresponding to what the catch for 2000 turned out to be. The performance parameter is the SSB for 2001.

Catch forecast for 2000:
Basis: TAC, Landings (99) = 480, F(99) = 0.73, SSB(2000) = 275.

F(2000)	Basis	Catch (2000)	Landings (2000)	SSB (2001)	Medium term (2003) effect of fishing at given level
0.00			0	560	<5% probability of SSB<$\mathbf{B}_{pa}$
0.13	$B_{2001}=\mathbf{B}_{pa}$ ($0.14\ F_{98}$)		110	500	<5% probability of SSB<$\mathbf{B}_{pa}$
0.22	$\mathbf{F}_{max}$ ($0.24\ F_{98}$)		184	460	<5% probability of SSB<$\mathbf{B}_{pa}$
0.32	$B_{2003}>\mathbf{B}_{pa}$ with high prob. ($0.35\ F_{98}$)		260	420	<5% probability of SSB<$\mathbf{B}_{pa}$
0.42	$\mathbf{F}_{pa}$ ($0.46\ F_{98}$)		328	386	11% probability of SSB<$\mathbf{B}_{pa}$
0.46	$\mathbf{F}_{med}$ ($0.51\ F_{98}$)		355	372	19% probability of SSB<$\mathbf{B}_{pa}$
	Actual catch in 2000		**414**	**343 - the performance parameter**	
0.91	F_{98}		610	256	78% probability of SSB<$\mathbf{B}_{pa}$

Weights in '000 t.

Results

For all six stocks it is clear that both the variation and the bias are large, with the exception of sole for which there is no bias (Table 1).

The variation is so large that for none of the stocks are the forecasted SSB significantly correlated with the "true" SSB.

There seems to be significant autocorrelation in the deviations from the "truth".

For North Sea cod the forecasted SSB was in 5 out of 12 year more than twice the "true" value and up to almost 4 times the "true" value in the assessment made in 1998, where the forecasted SSB(2000) was 205 000t while it has turned out that SSB was only 54 000t. All forecasted SSB were higher than the "true" SSB; on average about twice as high.

For North Sea plaice the forecasted SSB was in 2 out of 12 year more than twice the "true" value. All forecasted SSB were higher than the "true" SSB except the assessment made in 1988, which was 5% lower than the "truth". On average the forecasted SSB were 50% above the "truth".

For North Sea sole the forecasted SSB was never more than 49% above the "true" value. In the beginning of the period the forecast underestimated the actual SSB by a factor of two. On average the assessment were only 1% lower than the "true" average value.

For North Sea herring the forecasted SSB was in 3 out of 12 year more than twice the "true" value and up to over 3 times the "true" value in the assessment made in 1992, where the forecasted SSB(1994) was 1.409 million t while it has turned out that SSB was only 0.445 million t. Since the assessment in 1991 all forecasted SSB were higher than the "true" SSB. On average for the whole time series they were 72% higher.

For Central Baltic cod the forecasted SSB was in 4 out of 12 year more than twice the "true" value and up to more than 4 times the "true" value in the assessment made in 1996, where the forecasted SSB(1998) was 416 000t while it has turned out that SSB was only 102 000t. On average the forecasted SSB were about twice as high as the "true" SSB.

For Northeast Arctic cod the forecasted SSB was in 3 out of 12 year more than twice the "true" value and up to almost 4 times the "true" value in the assessment made in 1996, where the forecasted SSB(1998) was 1.550 million t, while it turned out that SSB was only 0.389 million t. Since the 1990 and 1991 assessment, which forecasted SSB to be only about half the "true" value the forecasts have been above the true value; on average for the whole time series 59% higher.

For each stock the forecasted SSB is plotted against the "true" SSB (Figure 1). It is clear that there is very little correlation between the two measures for each of the stocks. The tendency to overestimate SSB in the forecasts is also clear for all stocks except sole. When all stocks are plotted on the same graph (Figure 2) giving a larger dynamic range of data, the correlation becomes significant ($R^2 = 0.61$).

For Northeast Arctic cod a status quo forecast has been compared to a TAC one for the period 1996-1999 (Table 1f). A status quo forecast is marginally better than a TAC constraint forecast.

Also for Northeast Arctic cod the "correct" advice has been calculated based on the "true" stock size, and true values for biological parameters according to AFWG (2001). This was done for two years, the assessment years 1991 and 1996. In 1991 the advice should have been a catch of less than 475 000t in 1992, while the actual advice was 250 000t, and in 1996 the advice should have been a catch of less than 390 000t in 1997, while the actual advice was 993 000t.

Discussion

It must be remembered that what is called the "truth" in the present paper is not the real truth, which is unknown. First of all because a VPA is a very simple model of the truth and secondly because misreporting, discards and biological input to the model are uncertain. Thus, some extra and unknown uncertainty must be added to the already great variability observed in the forecasted SSB.

The reasons for the large errors in the forecasted SSB are likely to be found in model errors, as they are too large to be ordinary sampling errors only.

For instance the large SSB(1998) forecasted for the NEA cod in 1996, has been claimed to be due to unusual high catchabilities on some of the surveys, which are used for tuning the VPA for this stocks (Tore Jakobsen, pers. comm.). Cod in this area are semi-pelagic and if most cod in one particular year stay close to the bottom and becomes easy to catch the survey will give a too optimistic picture of the stock situation. As estimation of annual fluctuations in catchability is extremely difficult if not impossible, this is a very serious problem for the assessment. A likely consequence is that the survey cannot be used at all in the assessment.

For cod in the North Sea the large deviations found in the most recent years can to a large extent be ascribed to calculations errors in one of the important CPUE series used for tuning. This error has now been found and corrected.

For the herring in the North Sea a sudden shift in the level of stock number estimates in the acoustic tuning data series in the end of the 1980s and beginning of the 1990s can explain the major deviation in the forecasts made in the beginning of the 1990s. The overestimation in the later years are still unexplained, but are clearly not due to random noise as they are too large and too consistent in direction. For this stock we have three different tuning data series: the

IBTS (bottom trawl survey) age 2-5, the MLAI (herring larvae index) and the hydro-acoustic survey, and quite different signals in these series is at least a part of the explanation (see Figure 3).

For plaice and sole no obvious explanation has been put forward.

For Baltic cod significant misreporting have been suspected as the course of the error. Errors in age determination has been considered as a possible explanation, but according to Stuart (2001), who analysed this carefully, this seems not to be the reason.

In general, there are some speculations as given above, but no consensus about the reasons for the large errors in the past forecasts. It might be a fruitful exercise to try to find out the reasons.

In general the deviation of the assessment from the "truth" is very high especially if bias is not corrected for (cod North Sea 103%, plaice North Sea 48%, sole North Sea 28%, herring North Sea 75%, Baltic cod 128%, NEA cod 175%), and a CV of 20-30% as used when defining $\mathbf{B}_{pa}$ points from $\mathbf{B}_{lim}$ by ACFM, is clearly an underestimate of the uncertainty. Based on a regression analysis (Figure 2) CV is on average 54%. This means that the "magic formula" of 1.4 and1.64, which corresponds to CVs of 20% and 30%, should be revised to a general value of at least 2.4 [=exp(1.645*0.54), the 5% one-sided confidence limit] or to stock specific values, which generally should be much larger than 1.4-1.6.

The errors in the forecasts are larger than estimated by Brander (1987), Cook *et al*. (1991), and maybe also those calculated by Pattersson *et al*. (2000) although this study is difficult to compare with. The study by van Beek and Pastoors (1999) showed that there was no correlation between predicted and realised fishing mortality. The error discovered by Mohn (1999) for cod in eastern Scotian Shelf is of similar magnitude as the errors observed here. The Newfoundland cod assessments in the years when the stock collapsed was off by a factor of about 2 (F(1990) was assessed in 1991 to 0.51 and in 1992 to 0.96, Bishop and Shelton 1997). This will normally translate into an even larger error in the forecasted SSB for the "TAC year"+1, due to extra uncertainty introduced with the additional data needed on recruitment, weight at age, maturity at age etc. Thus, this assessment was probably as bad as the worst assessments presented here.

The following issues seems to be important for further consideration:

1. Investigate what precisely caused the largest of the errors in the forecasts (initial stock number, recruitment, weight at age, maturity at age, natural mortality, tuning data series (commercial vs. surveys etc.), like has been done for the Newfoundland cod stock.

2. Has the quality of the forecasts deteriorated in the recent decade compared to what Brander (1987) and Cook *et al*. (1991) found for the previous decade? And if yes, is that due to:

    a. depleted stocks which are more dependant on recruitment,
    b. more variable CPUE data (fleets are more dynamic now than in the past),
    c. more variable discarding and misreporting,
    d. lack of scientists with intimate knowledge of the fishery,
    e. more stressed working conditions for assessment work.
    f. working groups using too much time on medium-term projections etc. instead of on getting the basic short-term forecast correct.

3. Is the concept of using the surviving SSB more uncertain that other management procedures?

4. Is it better just to use survey data alone and forget about commercial catch data and VPAs?

5. Could climatic changes play a role?

6. Do we need to revise the $\mathbf{B}_{pa}$ and $\mathbf{F}_{pa}$ values, which have been defined based on the "magic formula", due to the fact that CV is not 0.20-0.30 as we thought they were, but more commonly several times larger?

Conclusion

Generally, the forecasts made by ICES have been very imprecise, when measured as SSB surviving the "TAC year". They have been wrong by a factor of two or more in 20 (3 too small by a factor of 2 and 17 too large by a factor of 2)

out of 67 cases for the six stocks analysed here. In three cases the SSB was overestimated by a factor of about 4. A clear bias towards a tendency to overestimate SSB, is apparent in 5 of the 6 stocks. The precise reasons for the poor quality are generally not known. The ”magic formula” used by ACFM to get $\mathbf{B}_{pa}$ and $\mathbf{F}_{pa}$ from $\mathbf{B}_{lim}$ and $\mathbf{F}_{lim}$ should be changed from the present values of 1.4-1.6 to a general value of at least 2.4, or to stock specific values, which generally should be much larger than 1.4-1.6.

References

van Beek, F. A. and Pastoors, M. A. 1999. Evaluating ICES catch forecasts: the relationship between implied and realised fishing mortality. ICES CM 1999/R:04.

Bishop, C.A., and Shelton, P.A. 1997. A narrative of NAFO 2J3KL cod assessments from extension of jurisdiction to moratorium. Can. Tech. Rep. Fish. Aquat. Sci. No. 2199.

Brander, K. 1987. How well do working groups predict catches?- J.Cons.int.Explor. Mer, 43:245-252.

Cook, R. M., Kunzlik, P. A., and Fryer, R. J. 1991. On the quality of the North Sea cod stock forecasts. –ICES J. mar. Sci., 48:1-13.

Mohn, R. 1999. The retrospective problem in sequential population analysis: An investigation using cod fishery and simulated data. – ICES Journal of Marine Science, 56:473-488.

Reeves, S.A. (2001). The implications of age-reading errors for stock assessment and management advice: a case-study based on Eastern Baltic Cod. ICES CM 2001/P:18.

Table 1a. North Sea cod. ACFM 2001 estimate of "the truth" from 1997 corrected with SSB from the Skagerrak and east Channel stocks, as the assessment before 1997 was based on a stock definition including only the North Sea.

				SSB surviving the "TAC year", i.e. SSB ("TAC year + 1) given the actual catch taken in the "TAC year" ('000 t)				
"Current year"	"TAC year" +1	Actual basis used by ACFM	Catch in "TAC year" '000t	S.q. basis	TAC basis	ACFM 2001 estimate "the truth"	Difference between S.q. basis and "the truth"	Difference between TAC basis and "the truth"
1988	1990	SQ	115	113		64	78%	
1989	1991	SQ	104	95		60	59%	
1990	1992	SQ	88	93		59	57%	
1991	1993	SQ	97	74		57	30%	
1992	1994	SQ	104	58		57	1%	
1993	1995	SQ	94	91		62	47%	
1994	1996	SQ	120	104		68	52%	
1995	1997	SQ	106	151		72	109%	
1996	1998	SQ	124	152		72	113%	
1997	1999	SQ	146	203		62	229%	
1998	2000	SQ	96	205		54	282%	
1999	2001	SQ	71	152		55	176%	
						Average	103%	

Table 1b. North Sea plaice.

				SSB surviving the "TAC year", i.e. SSB ("TAC year + 1) given the actual catch taken in the "TAC year" ('000 t)				
"Current year"	"TAC year" +1	Actual basis used by ACFM	Catch in "TAC year" '000t	S.q. basis	TAC basis	ACFM 2001 estimate "the truth"	Difference between S.q. basis and "the truth"	Difference between TAC basis and "the truth"
1988	1990	SQ	170	353		372	-5%	
1989	1991	SQ	156	376		314	20%	
1990	1992	SQ	148	347		278	25%	
1991	1993	SQ	125	392		244	61%	
1992	1994	SQ	117	420		204	106%	
1993	1995	SQ	110	365		181	102%	
1994	1996	SQ	98	253		159	59%	
1995	1997	SQ	82	234		137	71%	
1996	1998	SQ	83	258		199	30%	
1997	1999	SQ	72	288		200	44%	
1998	2000	SQ	81	365		251	46%	
1999	2001	SQ	83	317		289	10%	
						Average	47%	

Table 1c. North Sea sole.

				SSB surviving the "TAC year", i.e. SSB ("TAC year + 1) given the actual catch taken in the "TAC year" ('000 t)				
"Current year"	"TAC year" +1	Actual basis used by ACFM	Catch in "TAC year" '000t	S.q. basis	TAC basis	ACFM 2001 estimate "the truth"	Difference between S.q. basis and "the truth"	Difference between TAC basis and "the truth"
1988	1990	SQ	22	SHOT forecast		91	NA	
1989	1991	SQ	35	37		78	-53%	
1990	1992	SQ	34	43		78	-45%	
1991	1993	SQ	29	33		56	-41%	
1992	1994	SQ	31	67		75	-11%	
1993	1995	SQ	33	52		60	-13%	
1994	1996	SQ	30	50		38	32%	
1995	1997	SQ	23	45		30	49%	
1996	1998	SQ	15	34		23	48%	
1997	1999	SQ	21	55		49	12%	
1998	2000	SQ	23	49		48	3%	
1999	2001	SQ	23	40		39	3%	
						Average	-1%	

Table 1d. North Sea herring.

				SSB surviving the "TAC year", i.e. SSB ("TAC year + 1) given the actual catch taken in the "TAC year" ('000 t)				
"Current year"	"TAC year" +1	Actual basis used by ACFM	Catch in "TAC year" '000t	S.q. basis	TAC basis	ACFM 2001 estimate "the truth"	Difference between S.q. basis and "the truth"	Difference between TAC basis and "the truth"
1988	1990	SQ	788	1271		1225	4%	
1989	1991	SQ	645	997		1116	-11%	
1990	1992	SQ	658	863		915	-6%	
1991	1993	SQ	717	1180		684	73%	
1992	1994	SQ	671	1409		445	217%	
1993	1995	SQ	568	1174		473	148%	
1994	1996	SQ	639	985		467	111%	
1995	1997	SQ	306	856		434	97%	
1996	1998	SQ	273	673		529	27%	
1997	1999	SQ	380	1010		702	44%	
1998	2000	SQ	372	1481		815	82%	
1999	2001	SQ	372	1347		772	75%	
						Average	72%	

Table 1e. Central Baltic (Sub-divisions 25-32) cod .

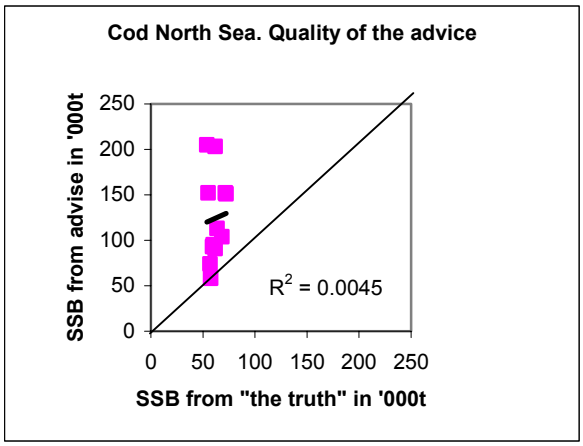
				SSB surviving the "TAC year", i.e. SSB ("TAC year + 1) given the actual catch taken in the "TAC year" ('000 t)				
"Current year"	"TAC year" +1	Actual basis used by ACFM	Catch in "TAC year" '000t	S.q. basis	TAC basis	ACFM 2001 estimate "the truth"	Difference between S.q. basis and "the truth"	Difference between TAC basis and "the truth"
1988	1990	SQ	179	398	0	215	85%	
1989	1991	SQ	154	255	0	151	68%	
1990	1992	SQ	123	267	0	96	179%	
1991	1993	SQ	55	163	0	118	39%	
1992	1994	SQ	45	81	0	197	-59%	
1993	1995	SQ	93	130	0	242	-46%	
1994	1996	SQ	108	NA	0	162	NA	
1995	1997	SQ	122	NA	0	130	NA	
1996	1998	SQ	89	416	0	102	307%	
1997	1999	SQ	67	267	0	75	258%	
1998	2000	SQ	73	233	0	87	168%	
1999	2001	SQ	66	162	0	95	71%	
						Average	107%	

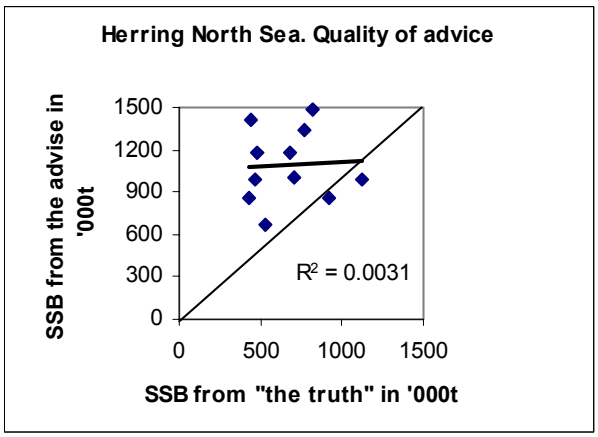
Table 1f. NEA cod.

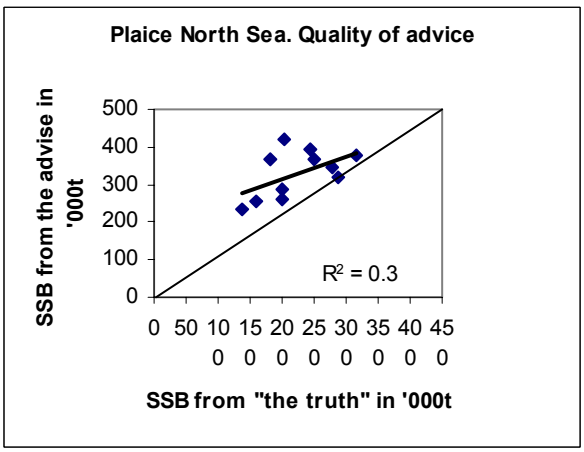
				SSB surviving the "TAC year", i.e. SSB ("TAC year + 1) given the actual catch taken in the "TAC year" ('000 t)				
"Current year"	"TAC year" +1	Actual basis used by ACFM	Catch in "TAC year" '000t	S.q. basis	TAC basis	ACFM 2001 estimate "the truth"	Difference between S.q. basis and "the truth"	Difference between TAC basis and "the truth"
1990	1992	TAC	319		320	873		-63%
1991	1993	TAC	513		398	735		-46%
1992	1994	TAC	582		630	603		4%
1993	1995	TAC	771		671	501		34%
1994	1996	TAC	740		724	571		27%
1995	1997	TAC	732		704	565		25%
1996	1998	TAC	762	1280	1550	389	229%	298%
1997	1999	TAC	593	830	716	259	220%	176%
1998	2000	TAC	485	446	495	223	100%	122%
1999	2001	TAC	414	328	343	300	9%	14%
						Average		59%
						Average (1996-1999)	140%	153%

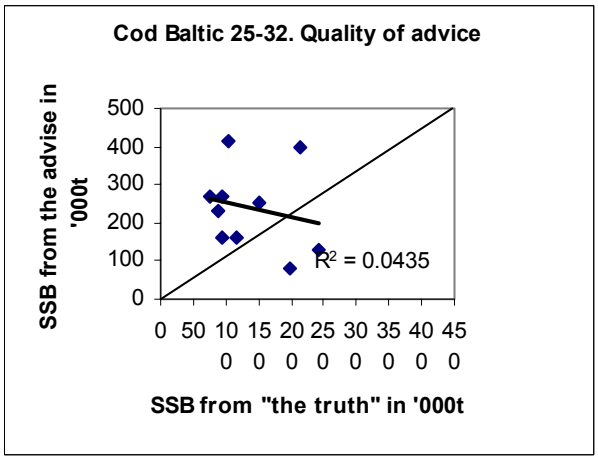
Table 2. (only partly filled out) NEA cod. The advice given by ACFM in the assessment year, what the advice would have been if the new advice principles had been applied in the assessment year and what the advice actually should have been (according to the new advice principles) if the stock had been correctly assessment (in this context meaning that the most recent VPA gives the truth).

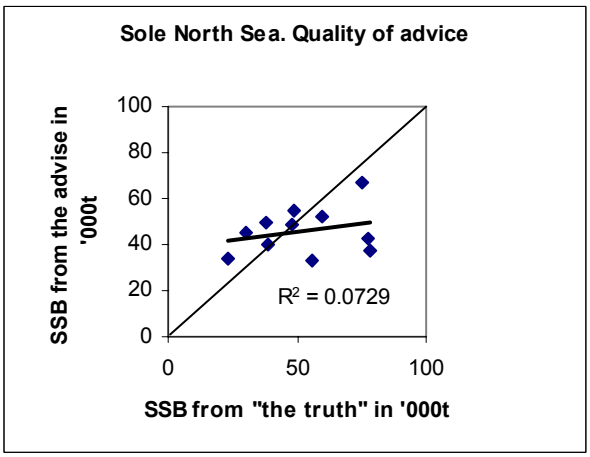
Assessment year	**A**. Actual advice in forecast year In '000t	**B**. Advice according to the present principle and with the present $\mathbf{B}_{pa}$ and $\mathbf{F}_{pa}$, but with the perception of stock size as per 1991 In '000t	**C**. As B, but given the knowledge we have now of the stock size, mean weight at age etc. at that time In '000t
1990			
1991	250	384	475
1992			
1993			
1994			
1995			
1996	Well below 993 (managers agreed on 850)	919	390
1997			
1998			
1999			
2000			
2001			











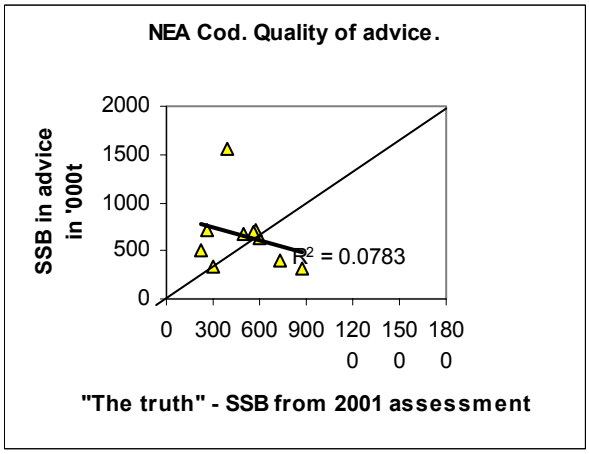


Figure 1. Forecasted SSB in "TAC year"+1 plotted against "true SSB".

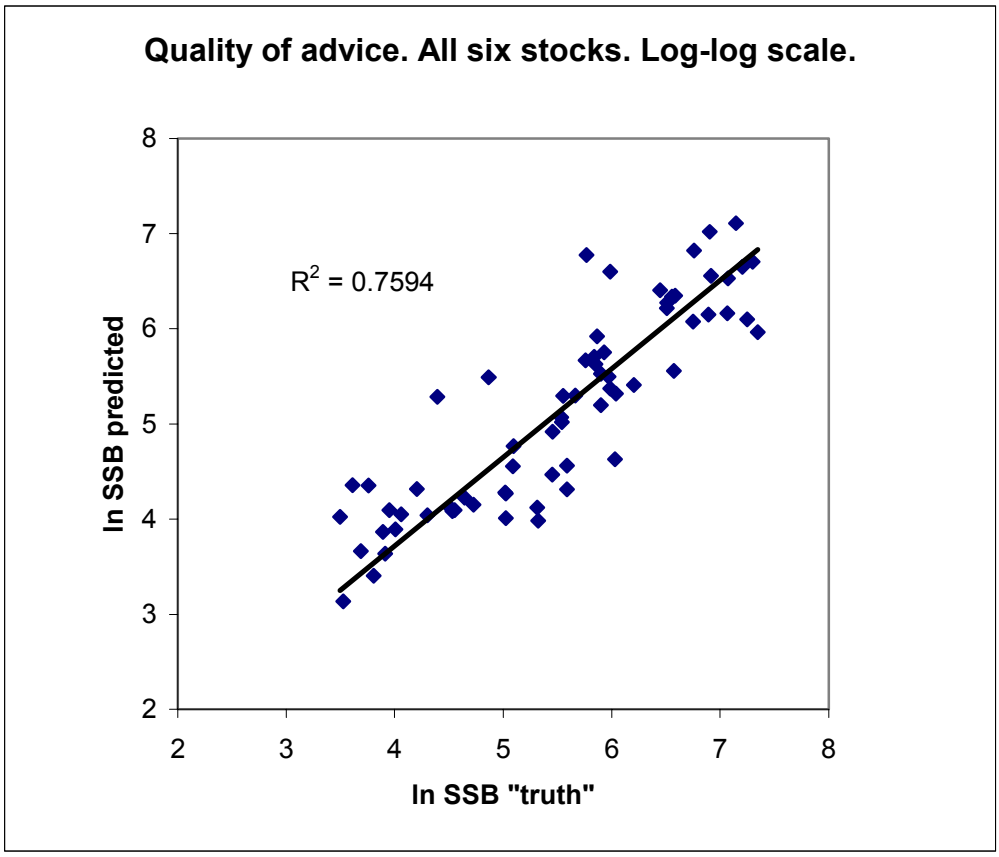


Figure 2. Forecasted SSB in the "TAC year"+1 plotted against the "true SSB". All stocks on the same plot.

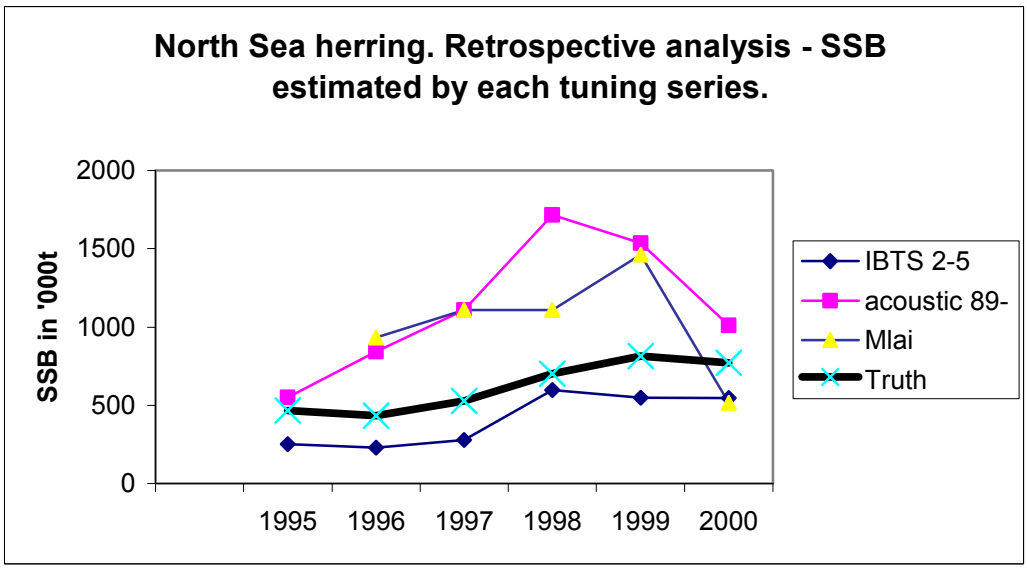


Figure 3. North Sea herring. Retrospective analysis of the signals in the tuning data. The "truth" is the ACFM assessment from May 2001. For each of the tuning series separate ICA runs have been made, which include each tuning series at a time. For instance the IBTS 2-5 point for 1995 is an ICA run where only the IBTS data were used in tuning and only data from 1995 and backwards were included. From the plot it can be seen that the IBTS follows the trends in SSB quite well, but is too low. The two other tuning series are too high.

APPENDIX B – WORKING DOCUMENT WA2

Over-estimation bias in the North Sea cod short-term forecasts

by
Chris Darby

ABSTRACT

During the past five years the total reported catch landed from the fishery for North Sea cod has not reached either the landings corresponding to the advice from the ACFM, or the final TAC agreed by the Council of Ministers. This paper discusses the historic over-estimation bias in the assessment and forecasts for landings. It highlights one assumption made within the current ICES method for forecasting future levels of landings, constant weight at age, which may have generated a substantial over-estimation bias within the forecasts for total landings' weight. The bias may have resulted in a continuation of high fishing mortality rates at a time when management was endeavouring to achieve a reduction in the level of exploitation.

Introduction

During the past five years the total reported catch landed by the fishery for North Sea and Skagerrak cod stocks (ICES areas IIIa, IV) have not reached either the landings corresponding to the advice from the ICES Advisory Committee on Fisheries Management (ACFM), or the final Total Allowable Catch (TAC) agreed by the Council of Ministers. Figure 1 illustrates the deficit for ICES area IV. Advice for reduced catch levels, intended to reduce the level at which the stocks are exploited, has not achieved that objective. Fishing mortality estimated for the combined North Sea and Skagerrak area has remained high and relatively stable at around 1.0.

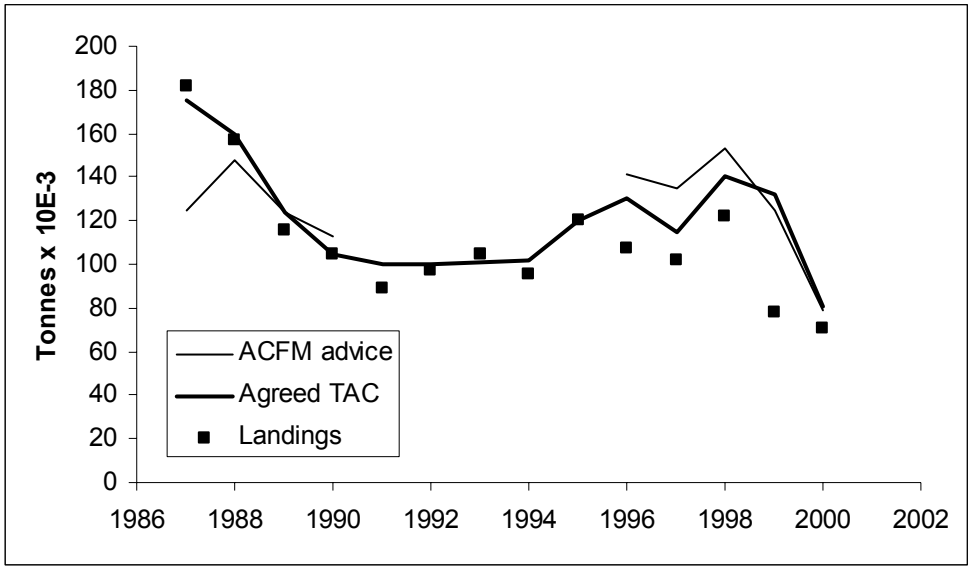


Figure 1. Time series of the ACFM advised landings, agreed TAC and recorded landings for the North Sea (Sub-area IV) cod stock. Data from ACFM 2000

Van Beek & Pastoors (1999) and Van Beek (2000) have described discrepancies between Working Group forecasts and the eventual outcome. The study by Van Beek (2000) has been summarised in the most recent Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak (North Sea WG) report (ICES CM 2002/ACFM:2). For most of the period 1986 – 1997 the agreed North Sea cod TAC was set at a level lower than the forecast landings under an assumption of status quo fishing mortality. Landings as used by ACFM, were close to the agreed TAC. It would have been expected that such a management regime should have resulted in a reduction in the fishing mortality imposed by the fishery. However, current estimates of the level of exploitation show that fishing mortality has not been reduced. Even more worrying is the trend during the past five years, in which the total landed catch has not reached either the landings corresponding to the advice from the ICES ACFM, or the final agreed TAC by a substantial margin. If the landings information is "reliable", at a time when the stock is at its lowest historical level, it implies that the TAC has not been restrictive.

It will be assumed that the shortfall in the landings results from uncertainties or bias in the Working Group assessments and forecasts and not as a result of excessive under-reporting, discarding or effort controls. That is, there is no incentive to under-report or to transfer effort to other species, if the TAC is not reached.

The ICES quality control sheets and historic North Sea WG reports illustrate that there has been an inconsistency/bias in the estimates of North Sea cod population assessment parameters and predictions estimated by consecutive Working Groups (Figure 2). At each new assessment meeting, the most recent estimates of fishing mortality have been revised upwards and spawning stock biomass reduced. Stock forecasts, made under status quo assumptions, have consistently over-estimated biomass levels and landings.

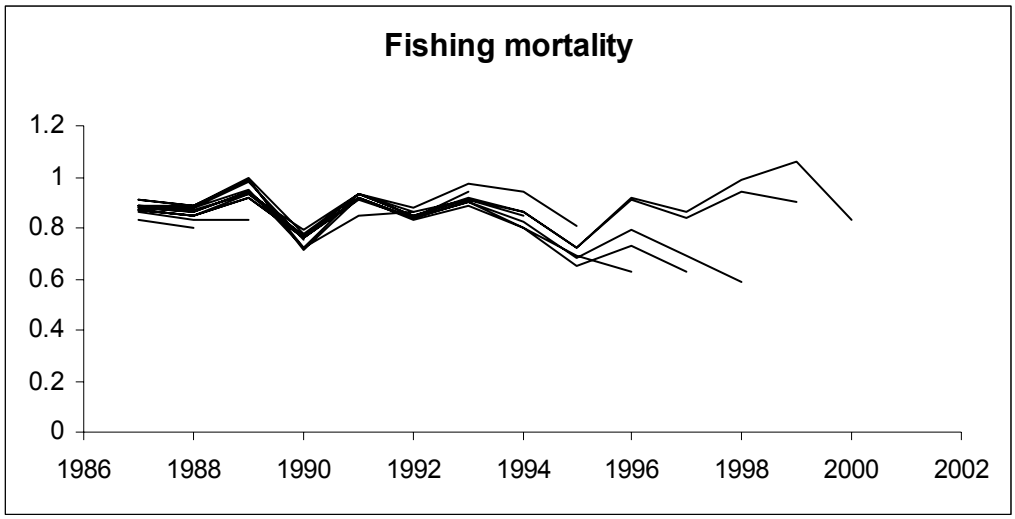
Biases in forecasts of stock abundance and landings can result from inappropriate assumptions within the assessment model used to estimate the current state of the stock, form within the model used to forecast future populations and landings, or both.

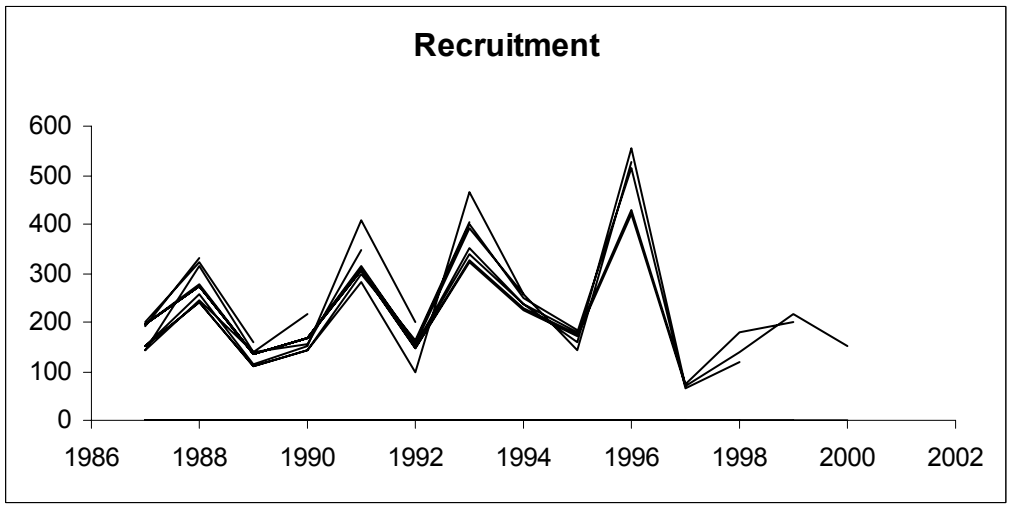
Bias in the assessment model

Recent Study Group (Bannister *et.al*., 2000) and North Sea WG (ICES CM 2001/ACFM:07, ICES CM 2002/ACFM:2) analyses have established that the commercial catch per unit effort data used to fit the Extended Survivors Analysis (XSA, Shepherd 1999) model to the North Sea cod was biased in the most recent years. At the 2000 Working Group the majority of the data sets were omitted from the fitted XSA model resulting in a substantial improvement in the consistency of the retrospective analysis results, although some bias remained. At the 2001 North Sea WG (ICES CM 2002/ACFM:2) all commercial information was excluded and the improved consistency of the XSA retrospective pattern was maintained. The change to the XSA data structure explains the difference in trends between the time series of estimates from the final two and the preceding assessments, illustrated in Figure 2.

The North Sea WG concluded that "The [current] configuration of XSA has little effect on the estimate of SSB … The tuning configuration can however, significantly affect the terminal exploitation pattern which can influence the mean fishing mortality and therefore the catch forecast. The Working Group is of the opinion that the greatest uncertainty is associated with the catch predictions rather than the assessment of the current state of the stock."

It should be noted that implicit within the search for an unbiased retrospective pattern is the assumption that the North Sea WG believes that the landings that they are using within the assessment are not substantially biased by under-reporting and discarding. If such biases were present in the data then, given unbiased CPUE tuning series, an over-estimation retrospective pattern would be expected from the model. The time series of final year values from consecutive assessments would indicate the trend in the "true" stock trajectory, rather than the time series trend of the final assessment. Chasing the Holy Grail of a perfect retrospective pattern can, in this situation, lead to a consistent estimate of a biased quantity.





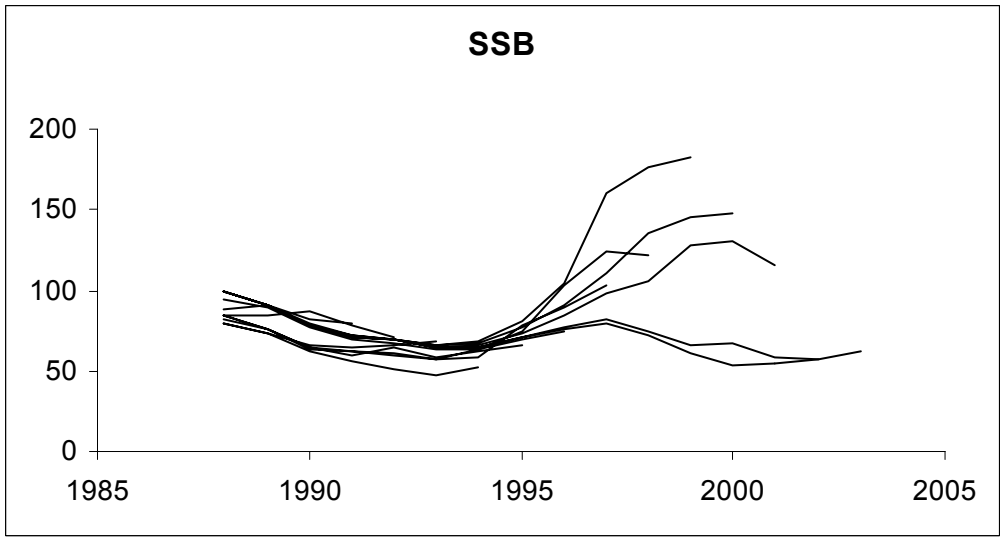


Figure 2. Consecutive Working Group estimates of fishing mortality, recruitment and spawning stock biomass as recorded in the ICES Quality control sheets. The final two values in each SSB series are stock forecasts. The significant adjustment in the final two years occurred after exclusion of the commercial fleet CPUE data series from the XSA tuning.

Bias in the stock forecast model

Work carried out at the ICES Study Group on Incorporation of Process Information into Stock Recruitment Models (SGPRISM, ICES CM 2001/C:02) indicated that the assumption of constant catch weight at age in stock forecasts for North sea cod can induce bias within medium term projections.

Systematic variation in the weight at age in the North Sea cod has been analysed by Cook *et. al*. (1999). The authors noted that, although inter-annual variability in the North sea cod was not large (c.v. 5-10%), there had been an increase in weight at age during 1990 – 1995 that could be modelled with a year class effect. The likely cause being a decline in the stock biomass and a release from density dependent competition.

Figure 3 plots the updated time series of relative weight at age fitted with a three year moving average smoother (y-1, y , y+1) and with each series offset in order to allow visualisation of the year class effects noted by Cook *et. al*. The increase in weight at age can be traced from age 1 in 1988/89 through to age 6 in 1998. Following the increase there has been a decrease in weight at age that can also be followed along the cohorts and although not formally analysed appears to be a year class effect.

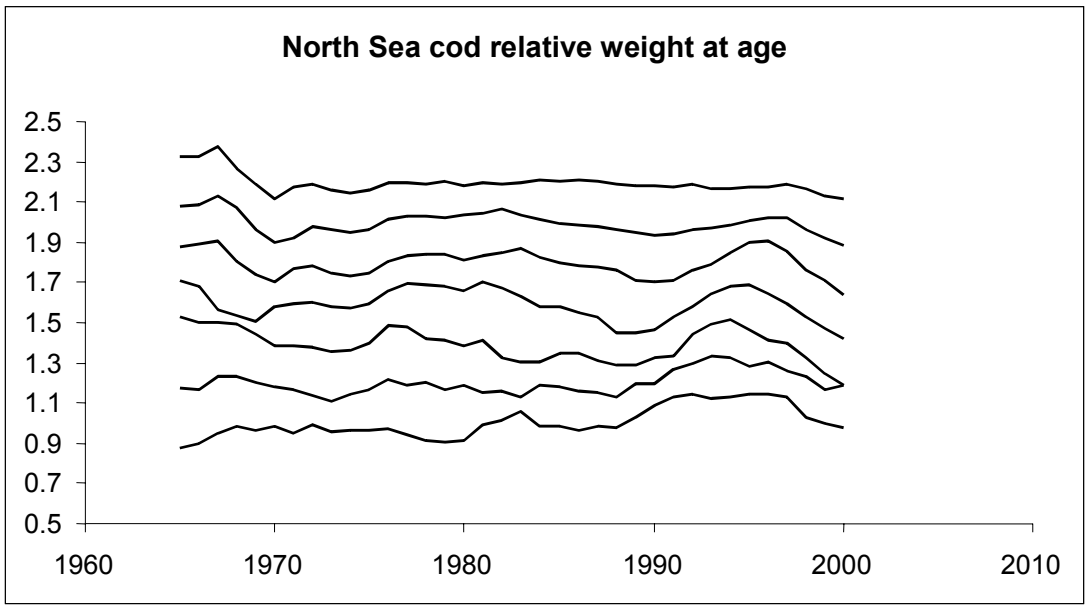


Figure 3. The time series of locally smoothed relative catch weights at age 1 - 7 for the North Sea cod illustrating the year class effect in the increase in weight during 1989 – 1993 and the subsequent decrease.

Given that there have been substantial systematic changes in the weight at age of the North Sea cod in recent years, an assumption of constancy could result in biased short-term projections. In order to illustrate this, Figure 4 shows the time series of cod weight at age for ages 1 – 5 together with the average unweighted arithmetic mean weight, from earlier years, that would be used for a catch prediction in that year. During the early years of the time series the three-year running mean provided weights at age that were consistent with the values recorded subsequently. However, in recent years cyclical changes in the weights at age have induced a lag in the response of the three year average and there is a systematic tendency to over or underestimate weight at age in the year of the forecast.

A minimal approach has been used to examine the bias induced through the use of a lagged mean weight, within the standard ICES short term forecast algorithm used for determining future landings levels for the North Sea cod stock. Arithmetic mean catch weight at age for a three year period ending in year y was used with the WG estimates of catch numbers at age from year y+2, to calculate a forecast total landings weight conditional on known catch numbers. The estimates of total landings were then compared to the landings recorded subsequently by the North Sea Working Group (ICES CM 2001/ACFM:02); all data were taken from that report. Conditioning on the recorded catch numbers at age removes any confounding bias in the forecast generated by the XSA model. The estimated bias in the total catch forecast resulting from the use of the mean weights at all ages is illustrated in Figure 5; a positive value indicates over-estimation of the total landings.

If the standard ICES approach to forecasting the TAC had been adopted during 1967 – 1989 there would have been a tendency to overestimate landings in 13 of the 23 years. The over-estimation bias would have been in the region of 5%. Under-estimation of the landings would have occurred in only three of the 23 years. During the years 1989 – 1993 there was a systematic increase in the weight at age covering the majority of the ages in the stock. The delayed response of the mean weight at age could have resulted in under-estimation of the total landings weight by approximately 10%. More recently weight at age in the landings has shown a decreasing trend and a simple historic arithmetic mean has over-estimated total landings weight with an increasing trend within the range 5 – 25%.

Figure 6 plots the percentage deficit between the landings and the agreed TAC along with the estimated bias introduced into the landings forecast for that year by the use of a simple mean weight. There is a significant correlation between the two series, even with the omission of the influential 1999 deficit. The TAC for that year was the last to be forecast from the results of an XSA that included the commercial fleet data.

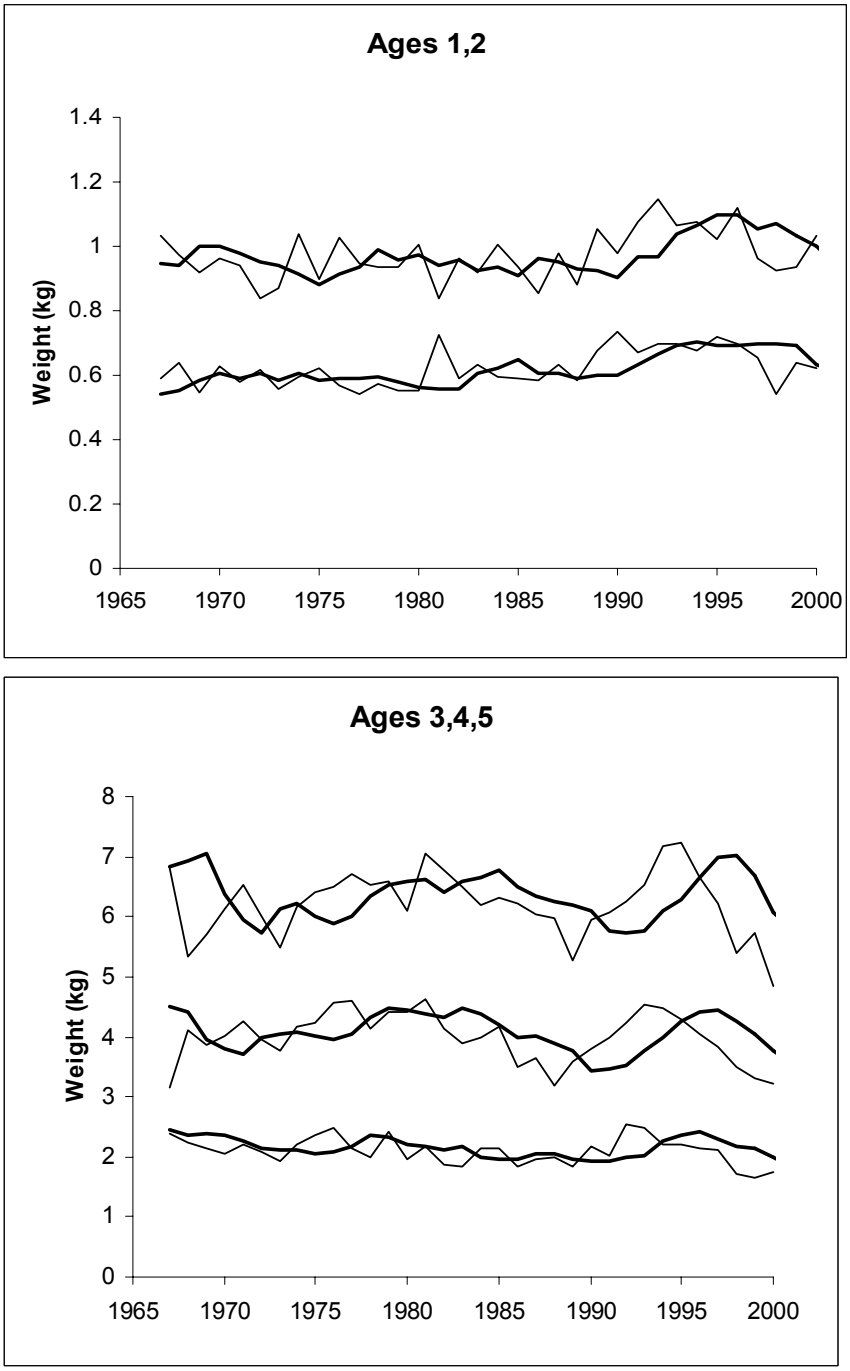


Figure 4a,b. The time series of weights at age of the North Sea cod as recorded at the 2001 ICES North Sea Working Group (thin line) and the unweighted arithmetic mean weight at age (bold line) that would be used for the short-term forecast management option table giving the TAC advice in the same year.

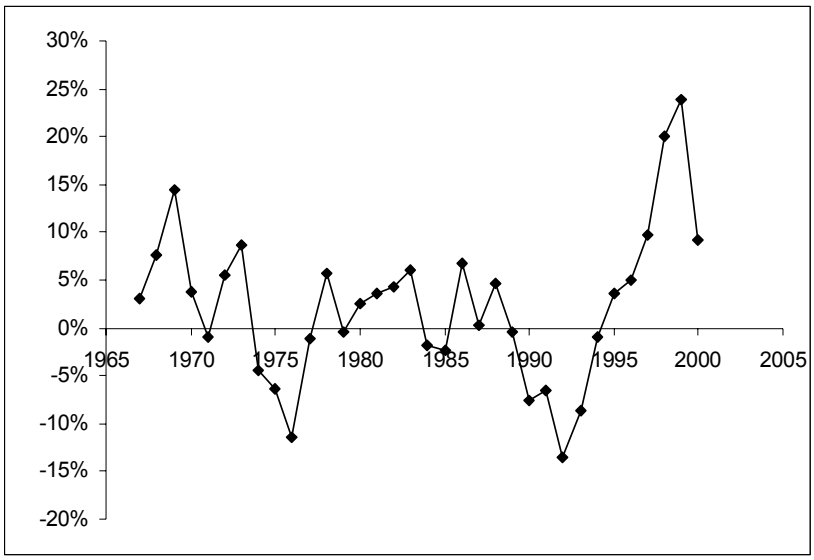


Figure 5. The percentage bias in the estimated landings for year y resulting from the use of an unweighted arithmetic mean weight at age calculated using the years y-4 to y-2, as currently applied within the ICES short-term forecast algorithm.

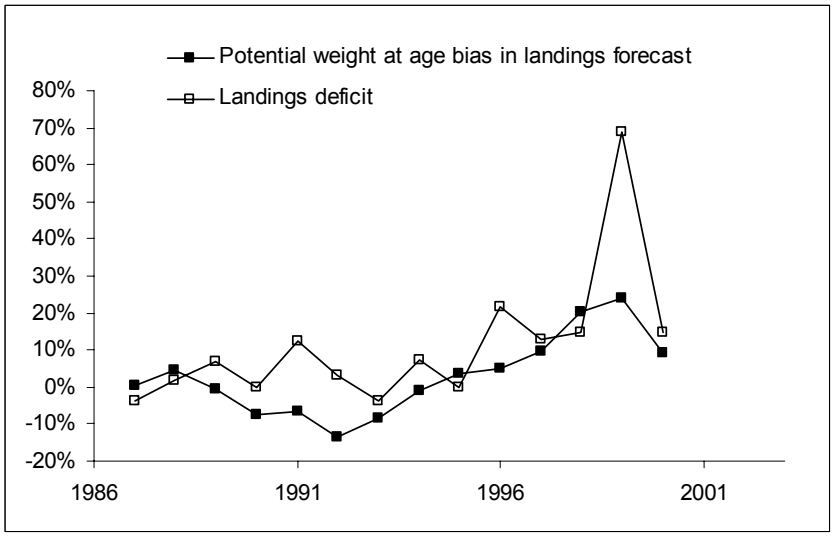


Figure 6. Time series of the percentage bias in landings resulting from the use of an unweighted arithmetic mean weight and the percentage deficit between the landings as recorded by ACFM and the agreed TAC in the same year. Positive values indicate over-estimation of landings in the short-term forecast and a higher TAC than recorded landings.

Discussion

The use of an arithmetic mean of historic catch weights at age in catch forecasts for 2 years into the future can induce bias into the ACFM management advice table for the North Sea cod.

For the period of 1990 - 1993 an under-estimation bias in the predicted landings would have negated some of the over-estimation of population abundance by the XSA model which included commercial catch per unit effort data. During the past five years landings could have been over-estimated within the range 5 – 25%, a bias that is consistent with the deficit between the total catch landed by the North Sea cod fishery and the final agreed TAC. If the recent landings information is "reliable" the magnitude of the bias is such that it could have removed the pressure of TAC regulation from the fishery.

The use of a biased estimate of weight at age in recent years could have resulted in overestimates of potential landings, and would certainly have amplified any retrospective over-estimation bias within the assessment structure. Even if the retrospective over-estimation bias within the XSA model has been reduced by the removal of the commercial fleet CPUE data there is still potential for over-estimation of future landings at a time when the stock is at its historically lowest levels.

Removal of the bias from the North Sea cod short-term forecast will require modelling of the temporal changes in weight at age. If available, commercial catch weight at age recorded in the year of the Working Group could be used to reduce the uncertainty in the forecast as shown by Figure 7. Cook *and al*. (1999) showed that a year class effect model can be fitted to the commercial weight at age data. Such a model could be used to predict weight changes two years into the future. The authors showed that similar effects could be detected in the weights at age from the International Bottom Trawl Survey (IBTS) time series. The results from the first quarter surveys would be available at the Working Group meeting and could be used to provide a more up to date indication of variation.

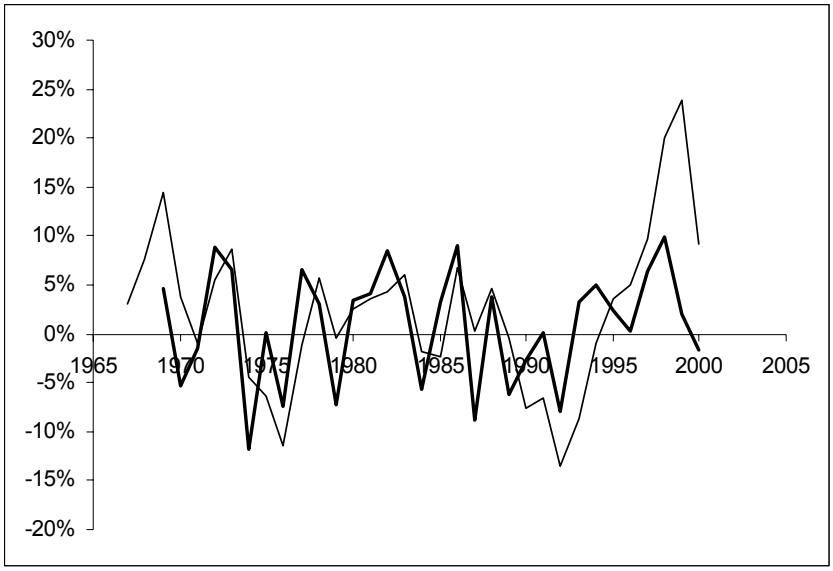


Figure 7. The bias induced into the short-term forecast of North Sea cod total catch weight through the use of a three year unweighted average weight for each age (thin line) and the weight at age recorded in the year of the assessment Working group (bold line).

REFERENCES

Bannister, C., Casey, J., Cook, R., Darby, C., Horwood, J. O'Brien, C., Reeves, S., Scott, R. North Sea cod meeting: Lowestoft 22-23 August 2000. WD-1 in ICES CM 2001/ACFM:07

Cook, R.M., Kunzlik, P.A., Hislop, J.R.G., Poulding, D. (1999). Models of growth and maturity for North Sea Cod. Journal of the Northwest Atlantic Fishery Science 25:91-99.

ICES 2001. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak, October 2000. ICES CM 2001/ACFM:7

ICES. 2001. Report of the Study Group on the Incorporation of Process Information into Stock-Recruitment Models. ICES CM 2001/C:02.

ICES 2002. Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak, June 2001. ICES CM 2002/ACFM:2

Van Beek, F.A., Pastoors, M.A. (1999). Evaluating ICES catch forecasts: the relationships between implied and realized fishing mortality. ICES C.M 1999 / R:4.

Van Beek, F. A. (2000). A note on the Working Group performance of short term predictions (Working Document). WD-2. WGNSSK 2000.

APPENDIX C – WORKING DOCUMENT WA3

Estimation bias in the North Sea short-term forecasts

by
Chris Darby

Introduction

The estimation bias resulting from the use of a simple model for future weight at age in the catch in forecasts for total allowable catch is examined for the North Sea stocks.

Method

A minimal approach has been used to examine the bias induced through the use of a lagged mean weight, within the standard ICES short term forecast algorithm used for determining future landings levels for the North Sea stocks. Arithmetic mean catch weight at age for a three year period ending in year y was used with the WG estimates of catch numbers at age from year y+2, to calculate a forecast total landings weight conditional on known catch numbers. The estimates of total landings were then compared to the landings recorded subsequently by the North Sea Working Group (ICES CM 2001/ACFM:02); all data were taken from that report. Conditioning on the recorded catch numbers at age removes any confounding bias in the forecast generated by the XSA model.

Results

The time series of relative weights at age for each of the stocks examined are illustrated in Figures 1,3,5,7,9. In each of the figures, year class effects and systematic trends in time are apparent in relative weight at age. The effects on the subsequent catch forecasts are illustrated in Figures 2,4,6,8,10; a positive value indicates over-estimation of the total landings.

Discussion

The assumption made within the current ICES method, of constant weight at age can generate a substantial estimation bias within the forecasts for total landings' weight. In recent years, for most of the stocks, the bias may have resulted in a continuation of high fishing mortality rates at a time when management was endeavouring to achieve a reduction in the level of exploitation. There appear to be systematic changes in the weight at age that could be modelled using time series and/or cohort effect models.

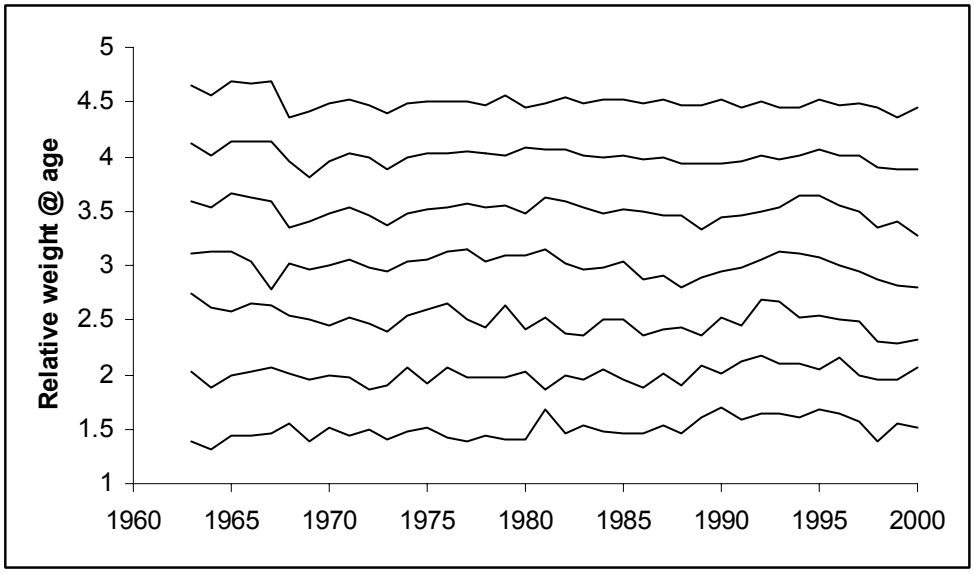


Figure 1. The time series of relative catch weights at age 1 - 7 for the North Sea cod illustrating the year class effect in the increase in weight during 1989 – 1993 and the subsequent decrease.

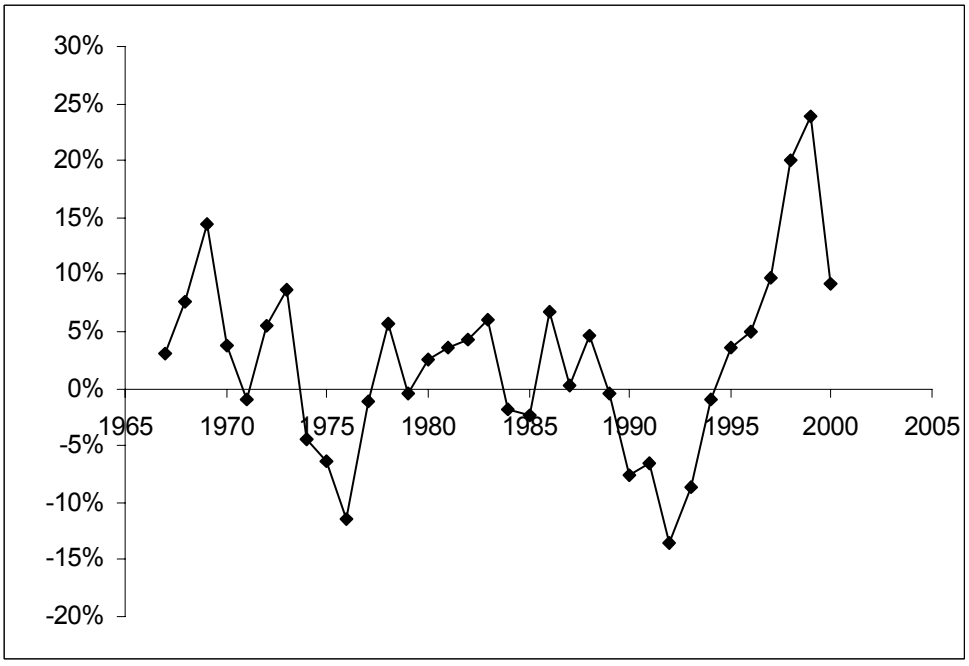


Figure 2. The percentage bias in the estimated landings for North Sea cod in year y resulting from the use of an unweighted arithmetic mean weight at age calculated using the years y-4 to y-2, as currently applied within the ICES short-term forecast algorithm.

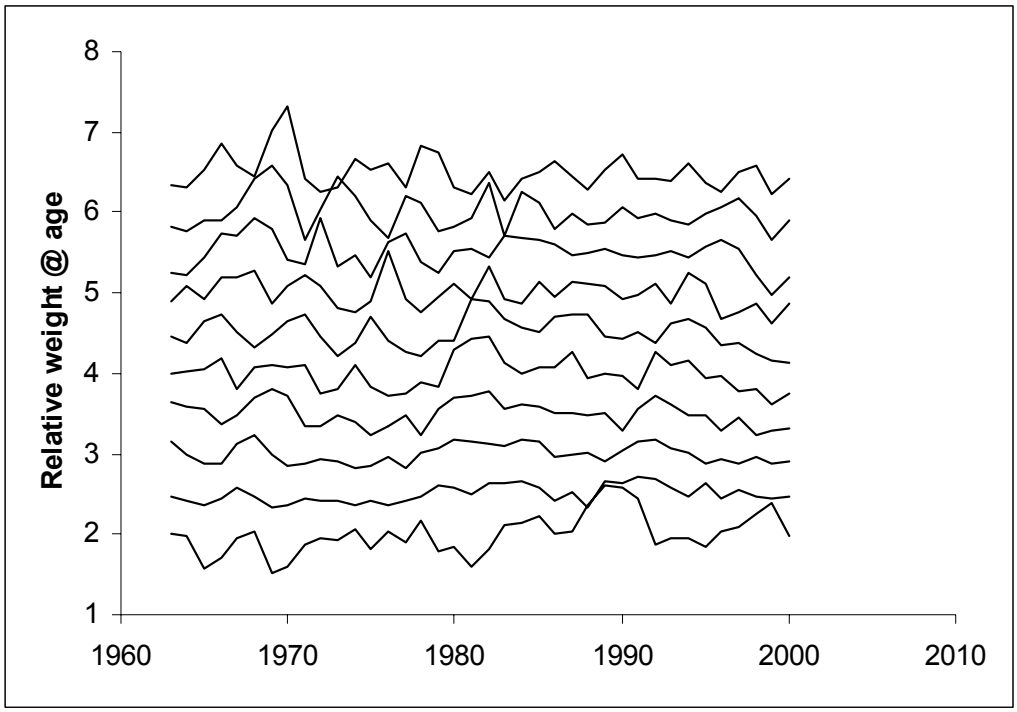


Figure 3. The time series of relative catch weights at age 0 - 9 for the North Sea haddock.

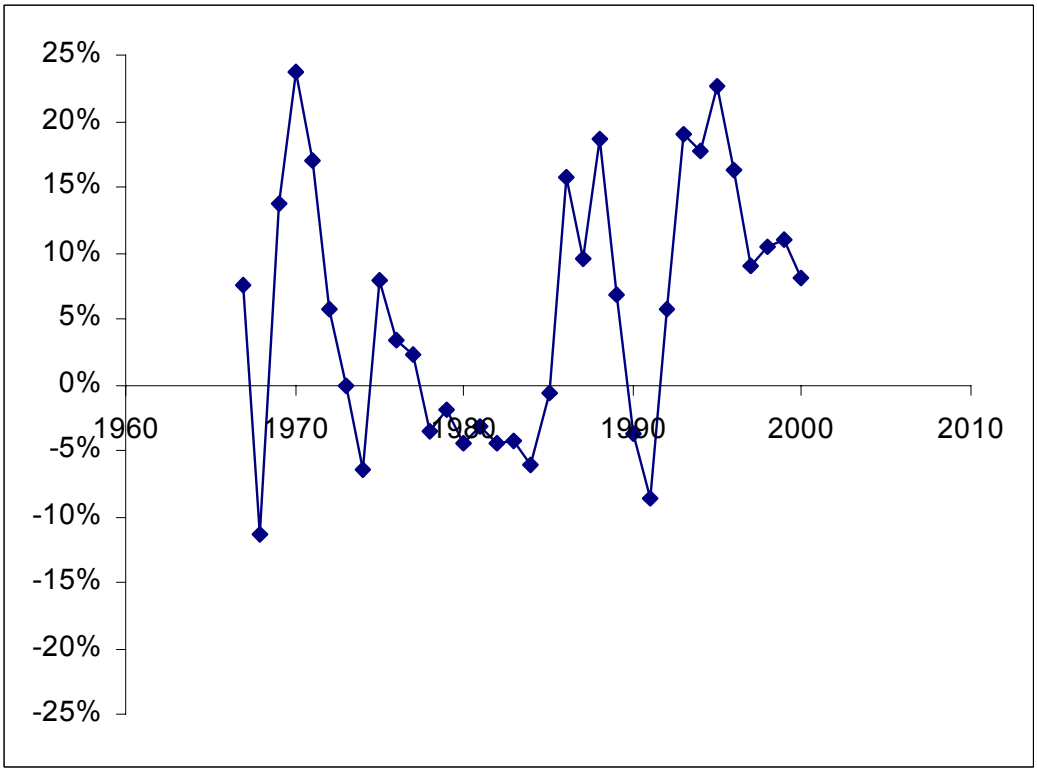


Figure 4. The percentage bias in the estimated landings for North Sea haddock in year y resulting from the use of an unweighted arithmetic mean weight at age calculated using the years y-4 to y-2, as currently applied within the ICES short-term forecast algorithm.

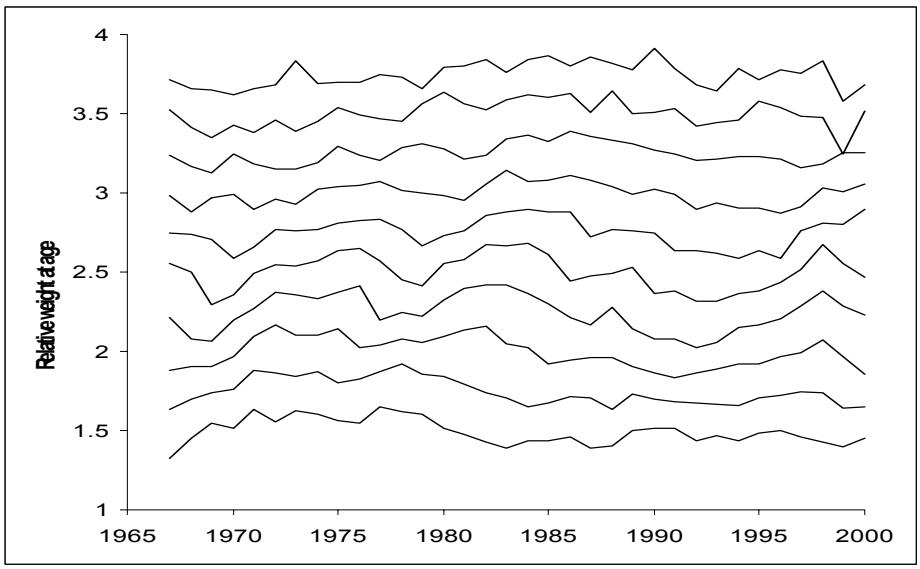


Figure 5. The time series of relative catch weights at age 1 - 10 for the North Sea plaice.

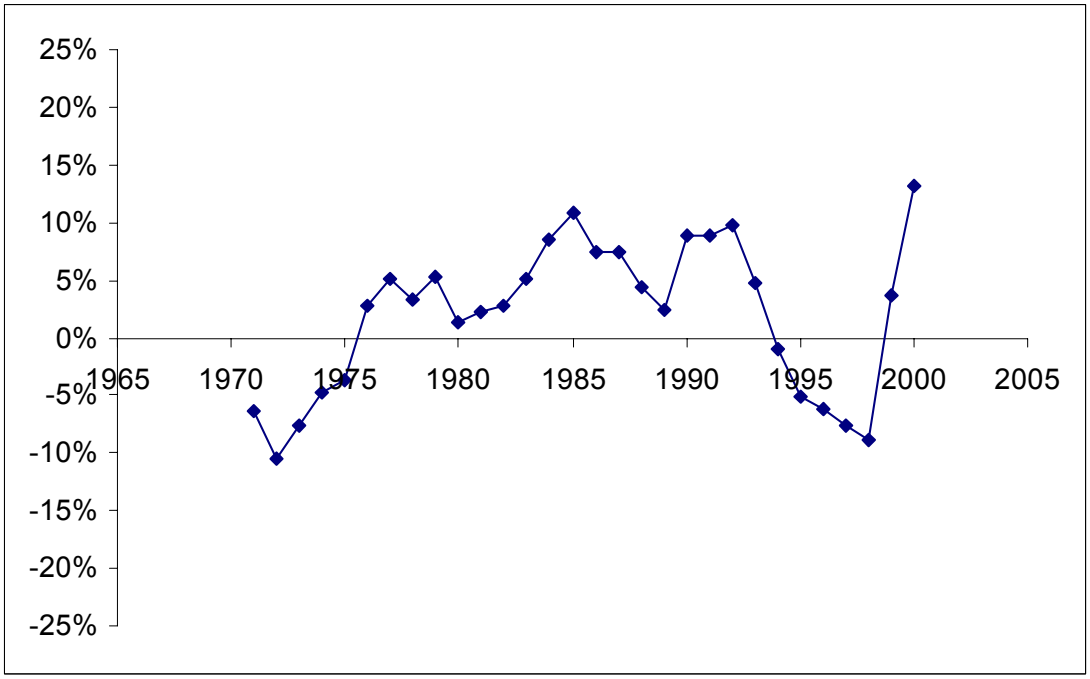


Figure 6. The percentage bias in the estimated landings for North Sea plaice in year y resulting from the use of an unweighted arithmetic mean weight at age calculated using the years y-4 to y-2, as currently applied within the ICES short-term forecast algorithm.

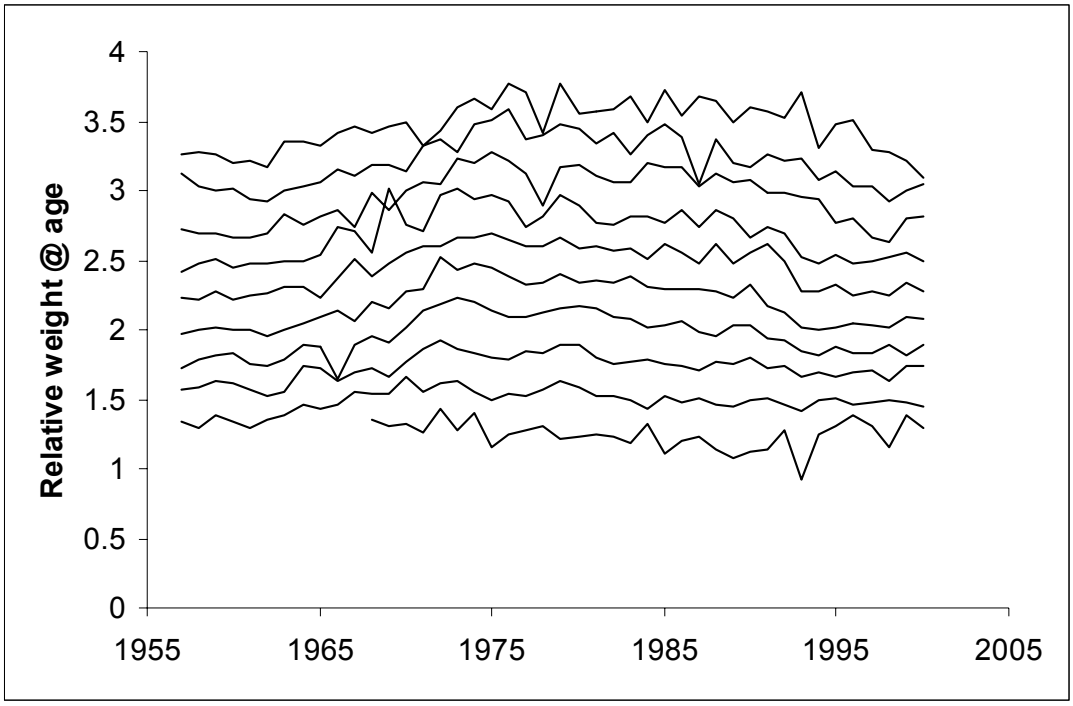


Figure 7. The time series of relative catch weights at age 1 - 10 for the North Sea sole.

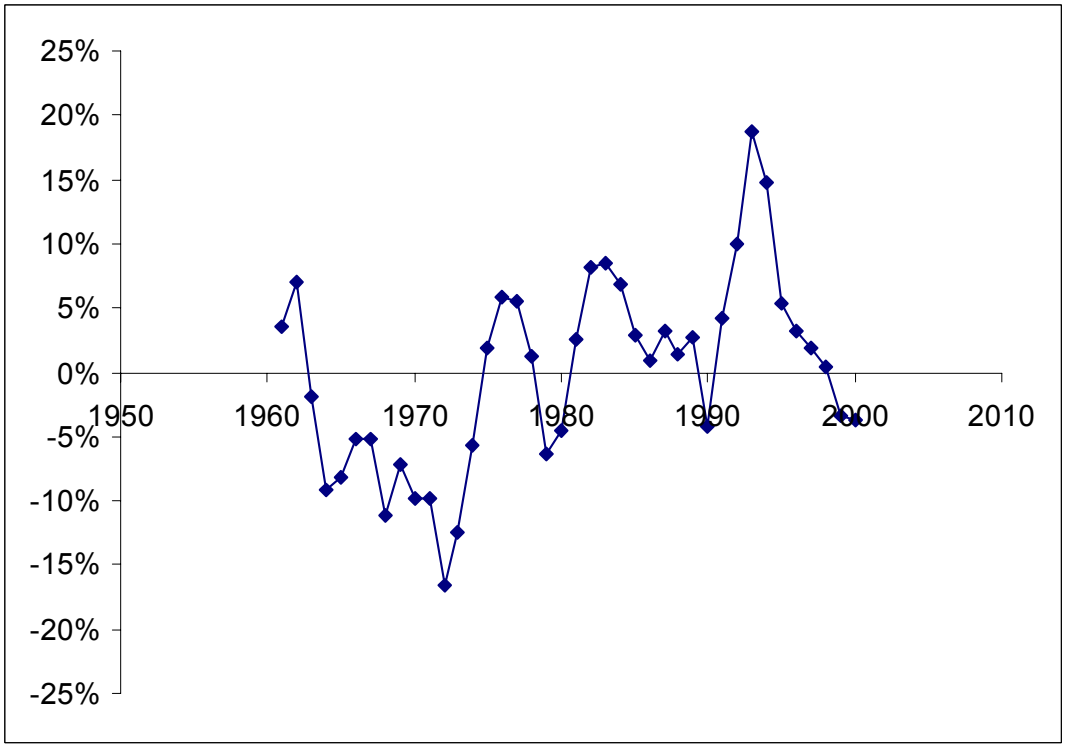


Figure 8. The percentage bias in the estimated landings for North Sea sole in year y resulting from the use of an unweighted arithmetic mean weight at age calculated using the years y-4 to y-2, as currently applied within the ICES short-term forecast algorithm.

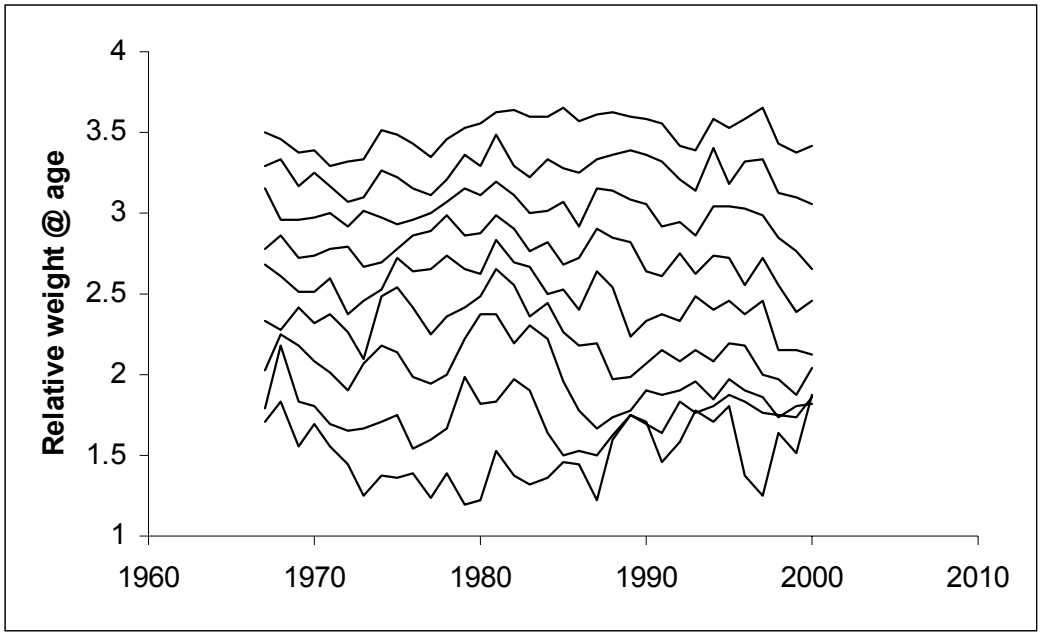


Figure 9. The time series of relative catch weights at ages 2 - 10 for the North Sea saithe.

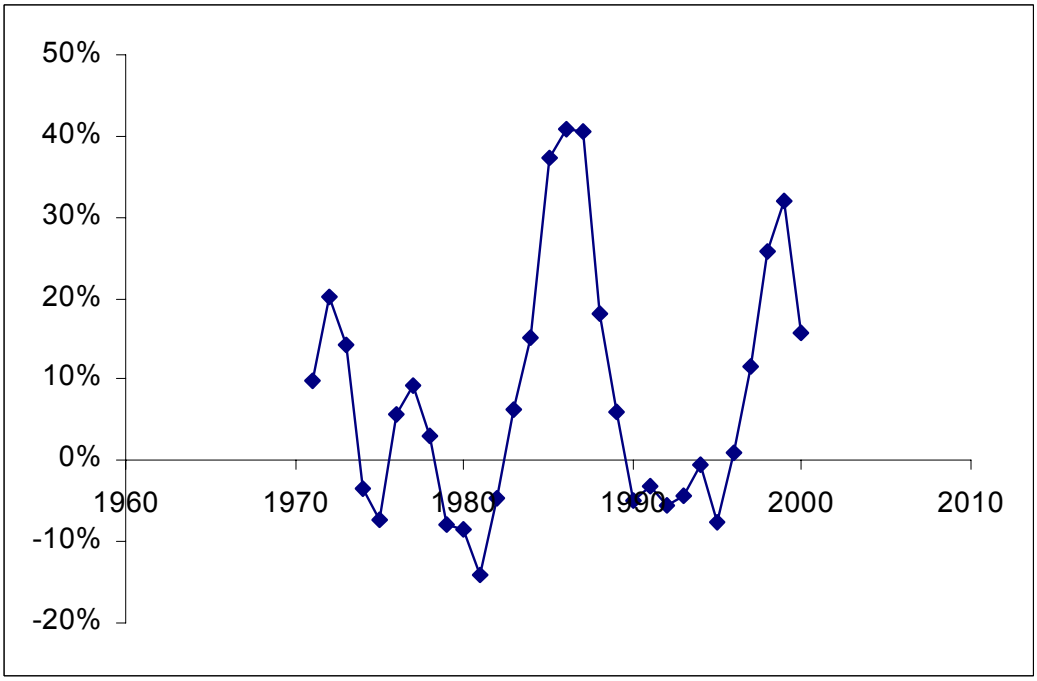


Figure 10. The percentage bias in the estimated landings for North Sea saithe in year y resulting from the use of an unweighted arithmetic mean weight at age calculated using the years y-4 to y-2, as currently applied within the ICES short-term forecast algorithm.

APPENDIX D – WORKING DOCUMENT WS3

TSA: is it the way?

by
Rob Fryer

TSA, or 'Time Series Analysis', provides an attractive framework for modelling commercial catch-at-age data. Despite its name, TSA is not a 'traditional' time series model involving e.g. autoregressive or moving average terms. Rather, TSA represents a fish stock / fishery in *state space* form. The *state* of the fishery in year *y* is described by the *state vector*, which contains all the information we need to know about numbers-at-age and fishing moralities-at-age in year *y*. The state vector evolves forward over time as determined by the *state equations*. For example, the state equations describe how the numbers-at-age in year *y*+1 depend on the numbers-at-age and fishing moralities-at-age in year *y*. The state vector is unobservable and inference about it is made using observations, typically catches-at-age, that are related to the state vector through *observation equations*. The Kalman filter is the algorithm used to estimate the state variables.

TSA was first developed by Gudmundsson (1994). It has been discussed by several Methods Working Groups, where its performance has been shown to compare well with other stock assessment methods. However, TSA failed to catch on (outside Iceland), presumably due to the lack of available and easy-to-use software. In 1997, needing to assess a cod time series containing several years with survey data but no reliable catch data, I coded a new implementation of TSA. This implementation was later extended to model landings-at-age and discards-at-age separately (Fryer *et al* 1998), and has since been used to assess five North Sea or VIa demersal stocks.

This working document has three objectives:

- to summarise the technical details of TSA
- to illustrate the technique (using VIa whiting)
- to discuss the strengths and weaknesses of TSA and to consider where it is going.

Theory

This section summarises the technical details of (the new implementation of) TSA. In essence the approach (for catch-at-age data at least) is identical to that of Gudmundsson (1994), the few modifications being mainly related to model parameterisation. Some details have been omitted for brevity, but these can be tracked down in Gudmundsson (1994), Harvey (1989), or Jones (1993). I first consider catch-at-age analysis, and then go on to consider the modelling of landings-at-age and discards-at-age separately.

The state vector and the state equations

The state of the fishery in year *y* is described by the state vector **s**(*y*), which contains all the information we need to know about numbers-at-age *N(a, y)* and fishing mortalities-at-age *F(a, y)* in year *y* (*a* = 1…A, *y* = 1…*Y*). The state equations describe how the state vector evolves forward in time. The state vector and the state equations clearly go hand in hand, but the state equations are more familiar territory so I'll begin with these.

The numbers-at-age in year *y*+1 depend on the numbers-at-age and fishing mortalities-at-age in year *y* through the usual equation:

$$N(a+1, y+1) = \exp\big(-Z(a,y)\big)\, N(a,y),$$

(with the familiar adjustments for a plus group).

Recruits in year *y*+1 are given by:

$$N(1, y+1) = f\left(N(\cdot, y)\right) + \varepsilon_{recruit}(y+1)$$

where *f*(.) is any specified stock-recruit function. The errors $\varepsilon_{recruit}(y+1)$ are assumed to be normally distributed with zero mean and standard deviation $\mathrm{cv}_{recruit} f(N(\cdot,y))$; i.e. recruitment is assumed to be distributed with constant coefficient of variation $\mathrm{cv}_{recruit}$. The parameters of the stock-recruit function and $\mathrm{cv}_{recruit}$ are estimated by maximum likelihood (see later). Note that other recruitment formulations are possible: in particular, recruits could be related to a pre-recruit index (see Gudmundsson, 1994).

Fishing mortalities evolve according to the following model (where NID stands for Normal Independent Deviate):

$$\log F(a,y) = U(a,y) + V(y) + \mathrm{NID}\left(0,\left(H(a)\sigma_F\right)^2\right)$$

$$U(a,y) = U(a,y-1) + \mathrm{NID}\left(0,\sigma_U^2\right) \quad a \le a_m < A$$

$$U(a,y) = U(a_m,y) \quad a > a_m$$

$$\text{with the constraint that } \sum_1^{a_m} U(a,y) = 0$$

$$V(y) = Y(y) + \mathrm{NID}\left(0,\sigma_V^2\right)$$

$$Y(y) = Y(y-1) + \mathrm{NID}\left(0,\sigma_Y^2\right)$$

This looks really complicated. And it is. But the salient features are that:

log fishing mortality is separated into an age component $U(a,y)$ and a year component $V(y)$, both of which can evolve over time,

a_m is an age above which fishing mortality is assumed to be constant (except for local transitory departures),

the variance σ_Y^2 induces persistent changes in fishing mortality (through the year component *V*),

σ_V^2 induces transitory changes in fishing mortality (through the year component *V*),

σ_U^2 induces persistent changes in fishing mortality (through the age component *U*),

σ_F^2 induces transitory changes in fishing mortality around the separable model *U* + *V*,

$H(a)$ allows the variability in fishing mortalities to be age dependent; typically $H(a)$ is initially taken to be unity, but can be adjusted if fishing mortalities for some ages (usually the young ages) are more variable than for others, the constraint on the $U(a,y)$ is necessary for identifiability.

Finally, the state vector consists of the $N(a,y)$, $\log F(a,y)$, $U(a,y)$, $V(y)$ and $Y(y)$.

The observation equations

Catches-at-age depend on the state vector through the usual catch equation:

$$C(a,y) = \frac{F(a,y)}{Z(a,y)}\left(1-\exp\left(-Z(a,y)\right)\right) N(a,y) + \varepsilon_{catch}(a,y)$$

The $\varepsilon_{catch}(a,y)$ are assumed to be NID with zero mean and standard deviation $\sigma_{catch}\, B_{catch}(a)\, q_{catch}(a,y)$ and represent measurement error in estimating the catch. The $B_{catch}(a)$ are initially taken to be unity, but can be adjusted later if the measurement errors associated with some ages (typically the older ages) are larger than for others. The $q_{catch}(a,y)$ are pre-determined from the catch data, as described by Gudmundsson (1994); if necessary, they can be inflated to decrease the influence of outliers.

The Kalman recursion

The Kalman filter is the algorithm used to estimate the state vector and the model parameters. It is an iterative procedure and works as follows. Suppose we have an estimate of the state vector in year *y* based on all the information available up to and including year *y*. Denote this estimate $\mathbf{s}(y|y)$ and let $\mathbf{P}(y|y)$ be the variance of $\mathbf{s}(y|y)$. The Kalman filter then moves forward to year *y*+1 by:

- using the state equations to predict the state vector in year *y*+1, denoted $\mathbf{s}(y+1|y)$, and its associated variance $\mathbf{P}(y+1|y)$,
- using the catch equations to predict the catches in year *y*+1, denoted $\mathbf{c}(y+1|y)$,
- calculating the *innovation* $\mathbf{I}(y+1)$, the difference between the observed catches $\mathbf{c}(y+1)$ and their predicted values $\mathbf{c}(y+1|y)$, with variance $\mathbf{V}(y+1)$,
- combining the innovation $\mathbf{I}(y+1)$ and its variance $\mathbf{V}(y+1)$ with the one-step ahead prediction of the state vector $\mathbf{s}(y+1|y)$ and its variance $\mathbf{P}(y+1|y)$ to give a new estimate of the state vector $\mathbf{s}(y+1|y+1)$ and its variance $\mathbf{P}(y+1|y+1)$.

The whole process requires staring values **s**(1|1) and **P**(1|1) (see Gudmundsson, 1994).

The estimates of the state vector in year *y* are based on the data up to and including that year, so only the estimates in the final year are based on all the available data. We therefore obtain final estimates of the state vector, based on all the data, by a further (backwards) recursive procedure known as *smoothing*.

At each stage of the recursion, we can calculate the log-likelihood of the innovation vector. Maximising the sum of these log-likelihoods allows us to estimate the unknown parameters in the model. These are the parameters of the stock-recruit curve and the associated coefficient of variation $\mathrm{cv}_{recruit}$, the four variances associated with the fishing mortality model $\sigma^2_F, \sigma^2_U, \sigma^2_Y,$ and σ^2_Y, and the variance of the catch data σ^2_{catch}. Three fishing mortalities $F(1,1), F(2,1),$ and $F(a_m,1)$ are required to provide sensible starting values of **s**(1) and these must also be estimated. Standard errors of the parameter estimates can also be calculated, but I have not yet implemented this. This is not critical, since it is the variances associated with the state vector that are necessary for making inferences about numbers-at-age, fishing mortalities-at-age, and associated variables such as spawning stock biomass, and these variances just drop out of the Kalman recursion.

Model assessment and adjustment

Model assessment is typically based on standardised catch prediction errors. Although these are not residuals in the true sense, they are useful for identifying outliers or ages where the catch data are more variable. Common adjustments are:

- increasing $q_{catch}(a,y)$ to downweight outliers,
- increasing $B_{catch}(a)$ for older fish, because catch estimates at these ages are based on few individuals,
- increasing $H(a)$ for younger fish, because fishing mortalities are more variable here.

Other adjustments are possible if there are long term trends in the state variables. For example, a long term trend in fishing mortality can be incorporated by including a trend parameter θ_Y in the state equation:

$$Y(y) = Y(y-1) + \theta_Y + \mathrm{NID}\left(0, \sigma^2_Y\right).$$

The trend parameter is estimated by maximum likelihood.

Occasional very large year classes are not well modelled by

$$N(1,y+1) = f(N(\cdot,y)) + \varepsilon_{recruit}(y+1)$$

A pragmatic solution is to allow:

$$N(1,y+1) = \lambda f(N(\cdot,y)) + \varepsilon_{recruit}(y+1)$$

where $\lambda > 1$ is a multiplier based on prior knowledge of the fishery. Recruitment is still assumed to be distributed with constant coefficient of variation; i.e. the error $\varepsilon_{recruit}(y+1)$ is assumed to be normally distributed with zero mean and standard deviation $\text{cv}_{recruit}\lambda f(N(\cdot,y))$. This approach can be thought of as putting an uninformative prior of the size of very large year classes. In practice the choice of λ does not appear to be particularly important.

Survey data

Survey data is incorporated as follows. Let $S(a,y)$ be the survey index of abundance at age *a* in year *y*. These data are assumed to be related to the state vector by the observation equation:

$$S(a,y) = \Phi(a)\,\Omega(y)\,N(a,y)\exp\left(-\tau\,Z(a,y)\right) + \varepsilon_{survey}(a,y)$$

where $\varepsilon_{survey}(a,y)$ are assumed to be NID with zero mean and standard deviation $\sigma_{survey}B_{survey}(a)\,q_{survey}(a,y)$ and τ denotes the time through the year of the survey. The $\Phi(a)$ are age-specific selectivities, assumed to be constant throughout the survey. Various parameterisations of the age-specific selectivities are possible, but all require some parameters to be estimated by maximum likelihood. Catchability $\Omega(y)$ is allowed to evolve over time, and enters the state vector rather like the year component *V*(*y*) in the fishing mortality model:

$$\Omega(y) = \beta(y) + NID\left(0,\sigma_{\Omega}^2\right)$$
$$\beta(y) = \beta(y-1) + NID\left(0,\sigma_{\beta}^2\right)$$

The variances σ_{Ω}^2 and σ_{β}^2 induce transitory and persistent changes in catchability respectively, and are estimated by maximum likelihood.

In practice, any number of surveys can be included, but the penalty is the increase in the number of parameters that have to be estimated by maximum likelihood.

Landings-at-age and discards-at-age

Now suppose that we have separate estimates of landings-at-age $L(a,y)$ and discards-at-age $D(a,y)$ and let $P(a,y)$ be the proportion of age *a* fish discarded in year *y*. The $P(a,y)$ are assumed to evolve as:

$$\text{logit}\ P(a,y) = a_1(y) + a_2(y)\times a + \text{NID}\left(0,\sigma_P^2\right)$$
$$a_1(y) = v_1(y) + \text{NID}\left(0,\sigma_{a1}^2\right)$$
$$v_1(y) = v_1(y-1) + \text{NID}\left(0,\sigma_{v1}^2\right)$$
$$a_2(y) = v_2(y) + \text{NID}\left(0,\sigma_{a2}^2\right)$$
$$v_2(y) = v_2(y-1) + \text{NID}\left(0,\sigma_{v2}^2\right)$$

Here:

- the proportions discarded at age in year *y* vary around a logistic discard curve with intercept $a_1(y)$ and slope $a_2(y)$,

- the discard curves evolve in time; σ^2_{a1} and σ^2_{v1} induce transitory and persistent changes in the intercept $a_1(y)$ respectively; similarly σ^2_{a2} and σ^2_{v2} induce transitory and persistent changes in the slope $a_2(y)$,
- the variables $\text{logit}\, P(a,y), a_1(y), a_2(y), v_1(y), v_2(y)$ enter the state vector, and the variances σ^2_{a1}, σ^2_{a2}, σ^2_{v1} and σ^2_{v2} are estimated by maximum likelihood.

The observation equations become:

$$D(a,y) = P(a,y)\,\frac{F(a,y)}{Z(a,y)}\left(1-\exp\left(-Z(a,y)\right)\right)N(a,y) + \varepsilon_{discards}(a,y)$$

$$L(a,y) = \left(1-P(a,y)\right)\frac{F(a,y)}{Z(a,y)}\left(1-\exp\left(-Z(a,y)\right)\right)N(a,y) + \varepsilon_{landings}(a,y)$$

where $\varepsilon_{discards}(a,y), \varepsilon_{landings}(a,y)$ are assumed to be NID with zero mean and standard deviation $\sigma_{discards}\ B_{discards}(a)\ q_{discards}(a,y)$, $\sigma_{landings}\ B_{landings}(a)\ q_{landings}(a,y)$ respectively.

In practice, virtually all very young fish are discarded and virtually all old fish are landed. It is therefore pragmatic to assume that all fish of age < a_{d1} are discarded, and all fish of age > a_{d2} are landed. The few fish landed at ages < a_{d1} are simply added to the discards data; similarly, any fish discarded at ages > a_{d2} are added to the landings data.

Show-time

To demonstrate TSA in action, consider the 2001 assessment of VIa whiting (ICES 2001b). The assessment reveals many of the typical features of TSA and shows up some problems too.

The data consist of landings-at-age and discards-at-age from 1978-2000 for ages 1 - 7+, and survey indices-at-age from 1985-2001, ages 1 - 6. The survey takes place in March, so $\tau = 0.25$.

The model specification was as follows:

- $a_m = 4$: based on inspection of previous XSA runs.
- $a_{d1} = 1$: many fish landed at age 1.
- $a_{d2} = 5$: usually some age 5 fish discarded, but zero age 6 discards were quite common.
- $B_{landings}(a) = 2$ for ages 6, 7+: extra measurement variability for older ages (fewer landings).
- $B_{discards}(a) = 2$ for age 5: extra measurement variability for the age with fewest discards.
- $B_{survey}(a) = 2$ for age 6: extra measurement variability for the oldest age (smallest catches).
- $H(1) = 2$: more variable fishing mortalities for age 1 fish.
- Downweighted following points by multiplying the relevant q by 3:
  - $D(5, 1993)$: unusually large discard value,
  - $S(4, 1992)$ and $S(5, 1992)$: unusually large values indicated by exploratory prediction error plots.

- Exploratory fits revealed an increase in $a_1(y)$, the intercept of the discard curve, over the assessment period; there was also a decrease in $a_2(y)$ over the assessment period, corresponding to a steeper discard curve; trend parameters were therefore incorporated as follows:

$$\nu_1(y) = \nu_1(y-1) + \theta_{v1} + \text{NID}(0,\sigma^2_{v1}),\quad \nu_2(y) = \nu_2(y-1) + \theta_{v2} + \text{NID}(0,\sigma^2_{v2}).$$

- Ricker stock-recruit model: $f(.) = \eta_1 S \exp(-\eta_2 S)$ where S denotes spawning stock biomass.
- The 1979 year class was unusually large, so the state equation for $N(1, 1980)$ was adjusted by taking $\lambda = 5$. The factor of 5 was chosen by comparing maximum recruitment to median recruitment from 1966-1996 for VIa cod, haddock, and whiting in turn using previous XSA runs.

$\Phi(1)$

- Age-specific selectivities: $\Phi(1)$, $\Phi(2)$, and $\Phi(a_m)$ estimated. Selectivity was assumed to be flat for all ages above a_m (i.e. $\Phi(a) = \Phi(a_m)$ for $a>a_m$). Selectivities between ages 2 and a_m were assumed to vary linearly on a log-scale.

The results are summarised in Table 1 and Figures 1-6

- Parameter estimates are given in Table 1.
- Standardised landings, discards, and survey prediction errors are shown in Figures 1-3.
- Stock trends with approximate pointwise 95% confidence limits are shown in Figure 4.
- The fitted Ricker stock-recruitment curve is shown in Figure 5.
- The proportions discarded at age and the fitted discard curves are shown in Figure 6.

The prediction errors are broadly centred around zero and are generally reasonable (Figures 1-3). However, some patterns are evident, most notably in the survey prediction errors. Table 1 shows that both σ_Y^2 and σ_β^2 are positive, indicating evidence of persistent changes in fishing mortality and survey catchability respectively. Both types of persistent change would induce patterns in the prediction errors. A limitation of the current implementation is that persistent changes in selectivity are assumed to be the same for all ages, whilst the survey prediction errors suggest that the changes might be age-specific: this is an area for further work.

The landings data are fitted much better than the discards data (Figure 4). This is to be expected given the limited data available for estimating discards-at-age, and is consistent with the estimates of measurement error in Table 1. The pointwise 95% confidence intervals are typical of TSA output. In particular, note the wide confidence intervals around mean fishing mortality and the widening confidence intervals around recruitment and spawning stock biomass in the most recent years. This serves as due warning not to interpret too closely small apparent changes in fishing mortality and spawning stock biomass in the terminal year.

There is only weak evidence for the Ricker stock-recruit curve (Figure 5), although the four most recent values are all on the ascending left-hand limb of the curve and this part of the relationship may become more clearly defined over the next few years if spawning stock biomass remains low. Having only limited information about the ascending left-hand limb can be a problem if a stock is about to collapse. This is because TSA favours the status quo, with any changes in the state variables constrained by the size of the relevant variance components. When there is limited information about the left-hand limb, recruitment can be over-estimated and the stock will appear to be healthier than it actually is. After a few years of declining spawning stock biomass, the stock-recruit curve will become better defined, the recruitment estimates will be revised downwards, and the true state of the stock will be revealed. This retrospective pattern was found with VI cod (Fryer 2000), where estimates of recent recruitment and spawning stock biomass were consistently revised downwards as information accrued about the Ricker stock-recruit curve.

There has been a large change in discarding practices over the assessment period with more older fish being discarded over time (Figure 6). The discard curves and the discard prediction errors (Figure 2) suggest that discarding at ages 2 and 3 has been over-predicted in recent years, although the stock trends (Figure 4) suggest this is not a serious problem at present. One cause of this over-prediction might be the presence of the trend parameters in the state equations describing the evolution of the discard curves. Whilst these trend parameters are necessary to account for the long-term changes in discarding over the assessment period, their presence implies that there will be further changes in the future and of course this will not necessarily be so.

Figure 7 shows retrospective plots for mean fishing mortality and spawning stock biomass. There is a tendency to underestimate fishing mortality and overestimate spawning stock biomass in the terminal year, although when judged relative to the 95% confidence bands, there is a particular problem only from 1994 to 1996. The retrospective estimates have improved from 1997 onwards. The retrospective pattern could be due to the suspected misreporting of whiting between 1992 and 1995. However, it could also be due to the tendency of TSA to maintain the status quo. In the retrospective runs, the estimate of σ_β^2 (persistent change in survey catchability) only became positive in 1997. Persistent changes in survey catchability probably began in the early 1990s but not until 1997 was there sufficient evidence in the data to identify these changes as more than just measurement error. For more details, see pages 130-131 of ICES (2001a).

Appraisal

Advantages

- Genuinely models the time-series nature of a fishery.
- Generally tracks the catch-at-age (or landings- and discards-at-age) data well.
- Gives precision of estimates of numbers-at-age and fishing mortalities-at-age: avoids over-interpretation of small recent changes in stock trends.
- Allows fishing mortalities-at-age to evolve in a constrained way: halfway house between separable assumption and fully unconstrained.
- Partitions the variability in the data into interpretable components.
- Predicts ahead (and gives precision of predictions).
- Can omit catch and / or survey data in some years if the data are suspect.
- Allows survey catchability to evolve.
- Can add in auxiliary information (i.e. independent estimates of measurement error, knowledge about 'tie-ups', multiple surveys).
- Can model landings-at-age and discards-at-age separately.
- Allows discard curves to evolve.
- Can omit discard data in some years if these have not been collected.

Disadvantages

- Requires normally distributed errors (but constant variance is not a requirement). This is not a particular problem in model fitting, but does have big limitations when it comes to predicting.
- Requires linear approximation of non-linear equations.
- There is some arbitrariness in the starting values.
- The likelihood can be very flat, so it can be difficult to estimate the model parameters.
- Maximum likelihood estimation can take a long time when there is lots of auxiliary data (and hence lots of parameters).
- Initial coding is hard. It is easy to modify the code to incorporate new model formulations (e.g. a different stock-recruit function or to account for a tie-up) if you did the initial coding.
- Favours status quo so struggles to pick up a collapsing stock.
- Has retrospective patterns (perhaps only until there is sufficient evidence in the data to confirm persistent changes in some of the state variables).
- Ad-hoc approach to modelling large year-classes.

Extensions to the current implementation

- Needs to be tidied up for general use: requires error trapping and documentation, but might not be sufficiently flexible to meet all needs.
- Needs to produce standard 'ICES output'.
- Needs to provide flexible stock-recruit and survey selectivity modules.
- Requires standard errors or profile likelihood regions for the model parameters.

The big picture

I believe that state-space formulations will be at the heart of stock-assessment models in years to come. However, I also believe that the current implementation of TSA is a dinosaur waiting for the meteor to strike. Its evolutionary weaknesses are twofold. First, the investment in coding is huge, and no matter how clever one is, it will be impossible to provide easy-to-use software that has the flexibility to deal with all the case studies that ICES will provide. Second, the restriction to normally distributed errors will limit the usefulness of short- and medium-term predictions made with TSA.

In a series of recent papers - unfortunately I don't have the references to hand - Millar & Meyer have shown how MCMC techniques can be used to fit similar state-space formulations to that used by TSA, but in a Bayesian context. This, I suggest, is the way forward. MCMC has three big advantages. It allows for different error distributions, it easily deals with non-linear equations, and it is simple to code. It will not be all win, however. The choice of priors is non-trivial, some skill is required to find model parameterisations with good chain-convergence, and there will be lots of time to look after your garden whilst the computer chugs away. On second thoughts, maybe it is all win after all.

References

Fryer RJ, 2000. Preliminary assessments of VIa cod. Working document to the Working Group on the Assessment of Northern Shelf Demersal Stocks 2000.

Fryer RJ, Needle C, Reeves SA, 1998. Kalman filter assessments of cod, haddock, and whiting in VIa. Working document for the Working Group on the Assessment of Northern Shelf Demersal Stocks 1998.

Gudmundsson G, 1994. Time series analysis of catch-at-age observations. Applied Statistics 43: 117-126.

Harvey AC, 1989. Forecasting, structural time series models and the Kalman filter. Cambridge University Press.

ICES, 2001a. Report of the Working Group on the Assessment of Northern Shelf Demersal Stocks. ICES CM 2001/ACFM:01.

ICES, 2001b. Report of the Working Group on the Assessment of Northern Shelf Demersal Stocks 2001.

Jones RH, 1993. Longitudinal data with serial correlation: a state-space approach. Chapman & Hall.

Table 1 Parameter estimates

parameter		estimate
Initial fishing mortality	F (1, 1978)	0.29
	F (2, 1978)	0.57
	F (4, 1978)	0.86
Survey selectivities	$\Phi(1)$	1.03
	$\Phi(2)$	0.66
	$\Phi(4)$	0.34
Standard deviations		
fishing mortalities	σ_F	0.11
	σ_U	0.00
	σ_V	0.00
	σ_Y	0.21
survey catchabilities	σ_Ω	0.00
	σ_β	1.72
measurement	$\sigma_{landings}$	0.12
	$\sigma_{discards}$	0.69
	σ_{survey}	0.52
discard curves	σ_P	0.33
	$\sigma_{\alpha 1}$	0.13
	σ_{v1}	0.00
	$\sigma_{\alpha 2}$	0.17
	σ_{v2}	0.00
Trend parameters	θ_{v1}	0.101
	θ_{v2}	-0.037
Recruitment	η_1	8.77
	η_2	0.035
	$cv_{recruits}$	0.29

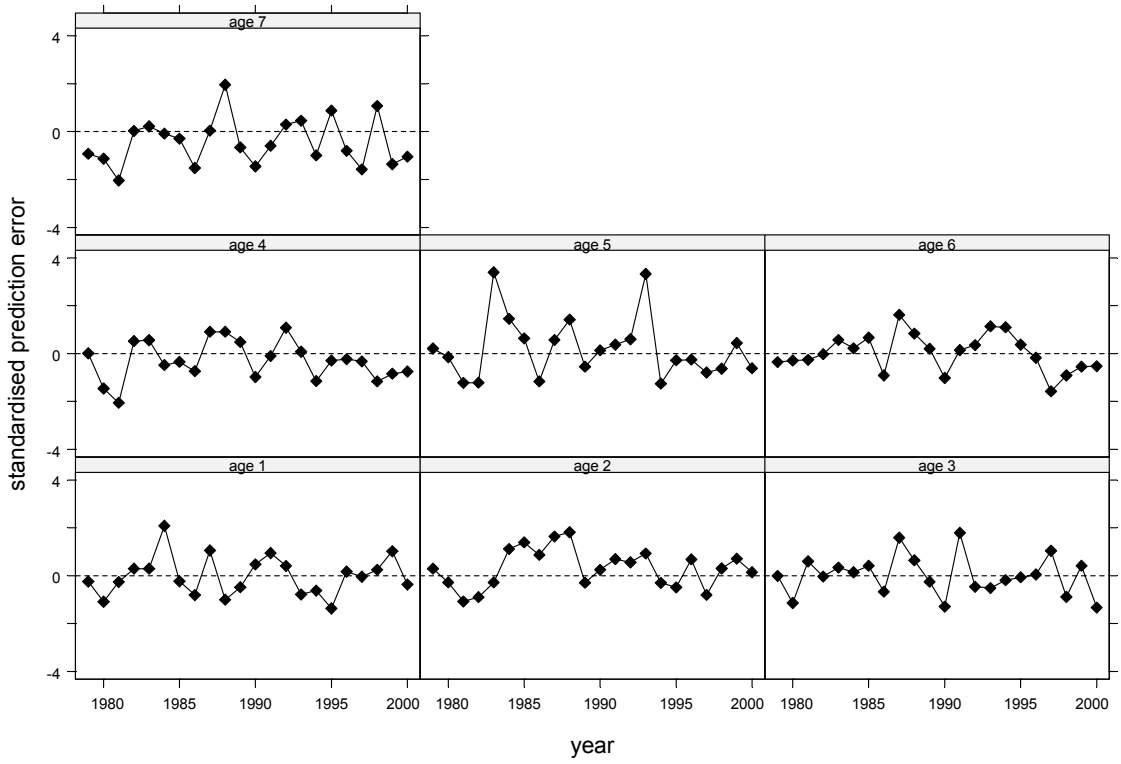


Figure 1 Standardised landings prediction errors.

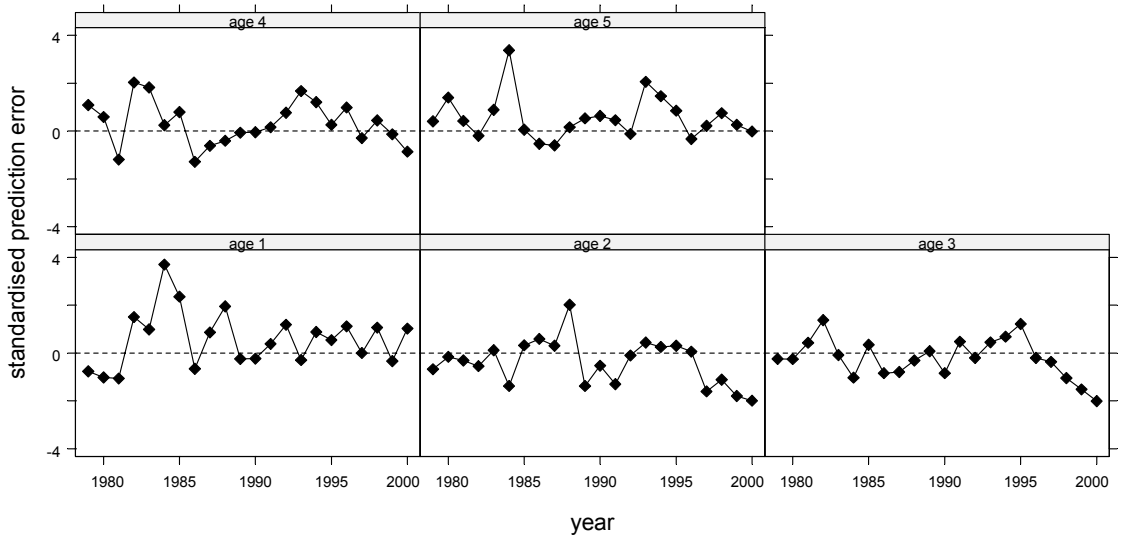


Figure 2 Standardised discards prediction errors.

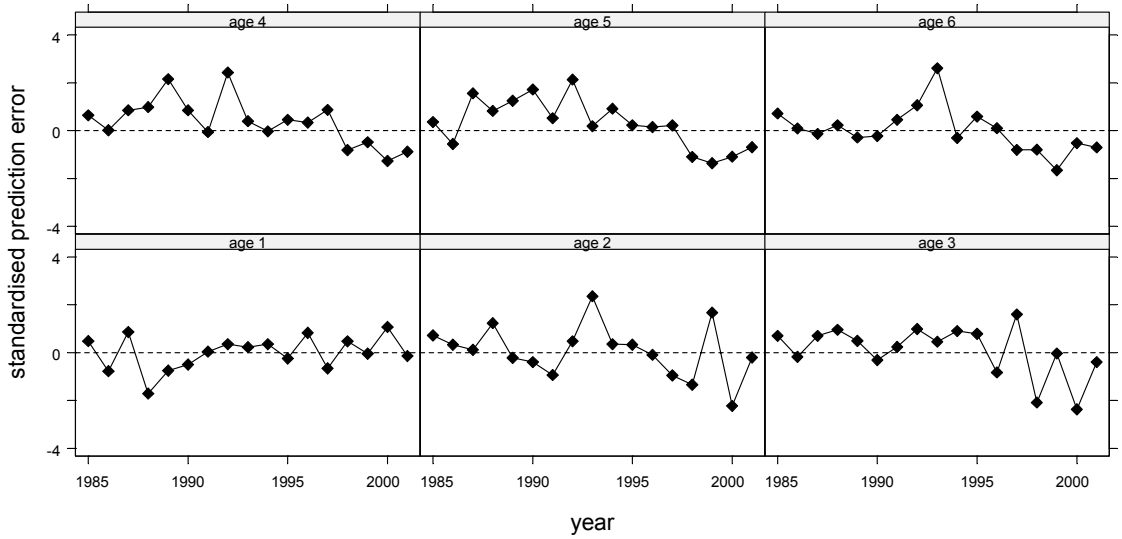


Figure 3 Standardised survey prediction errors.

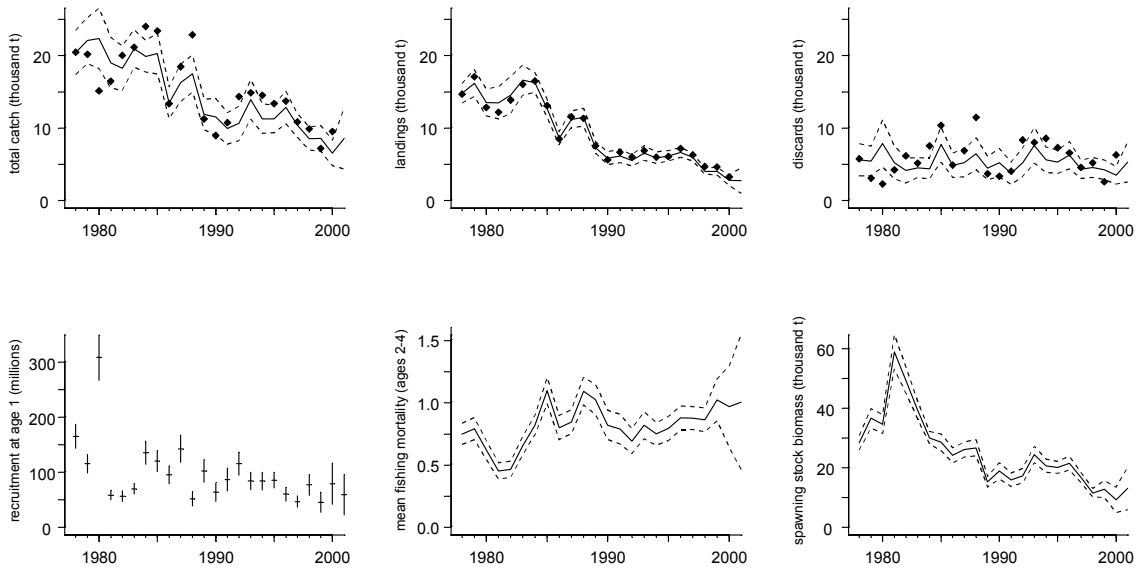


Figure 4 Stock summary

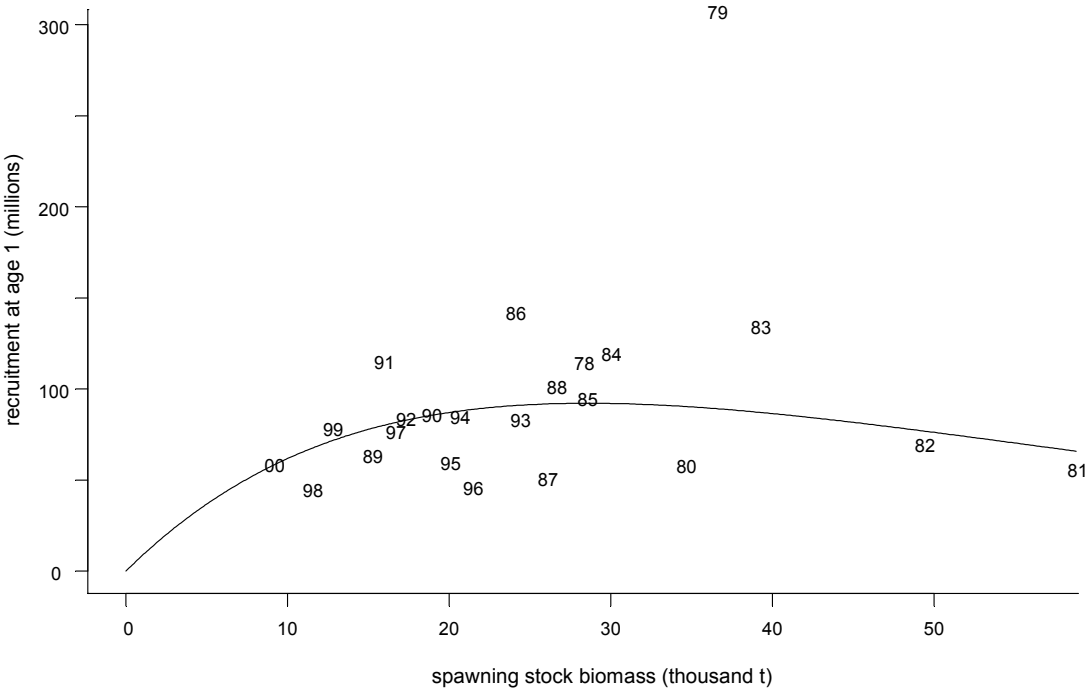


Figure 5 Estimated Ricker stock-recruit curve. The plotting symbols denote the year class.

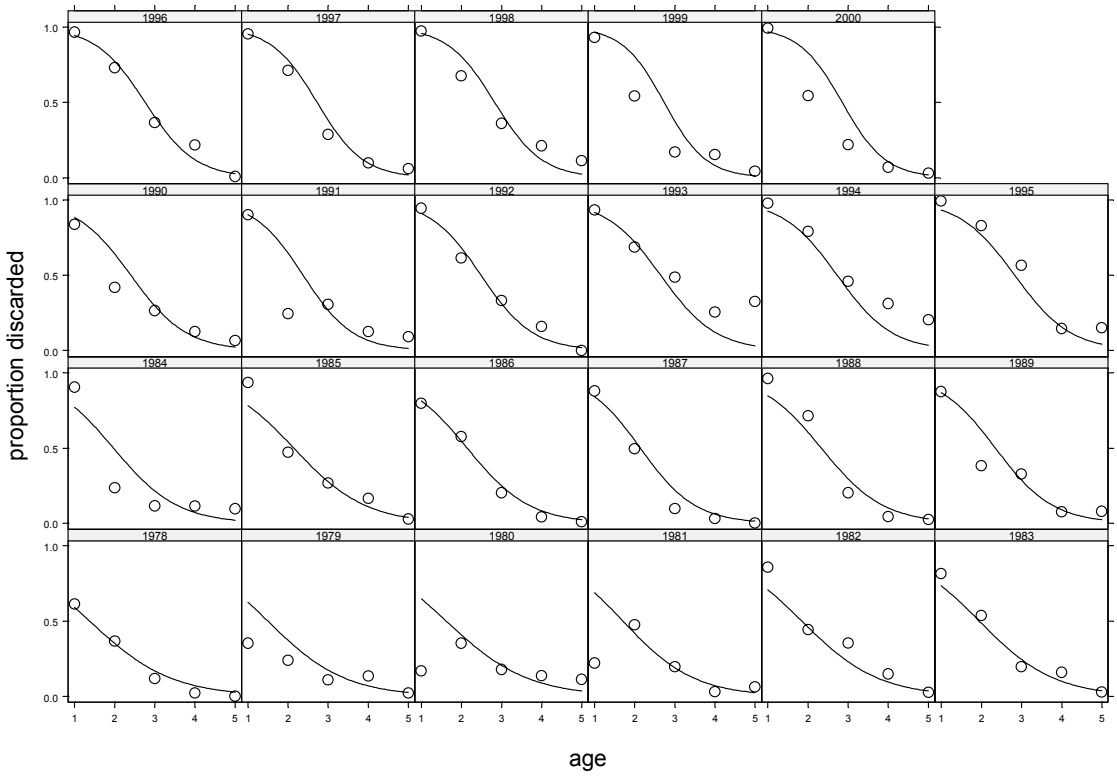
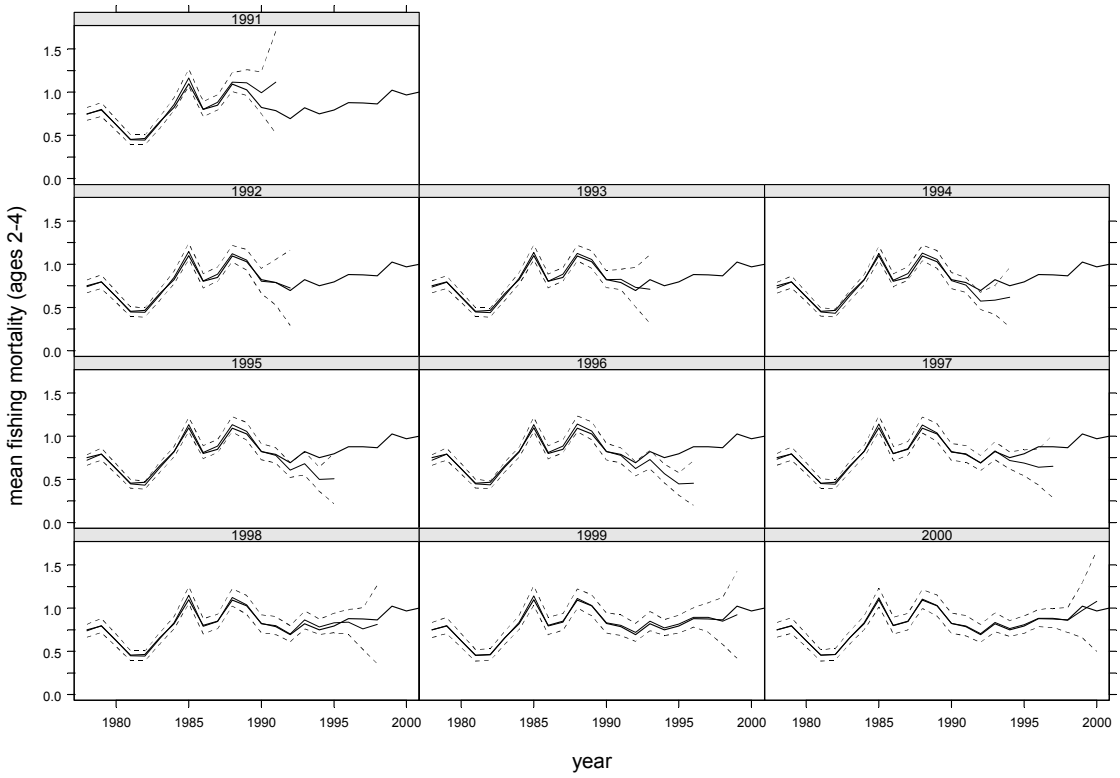


Figure 6 Proportion discarded at age, and fitted discard curve, by year.



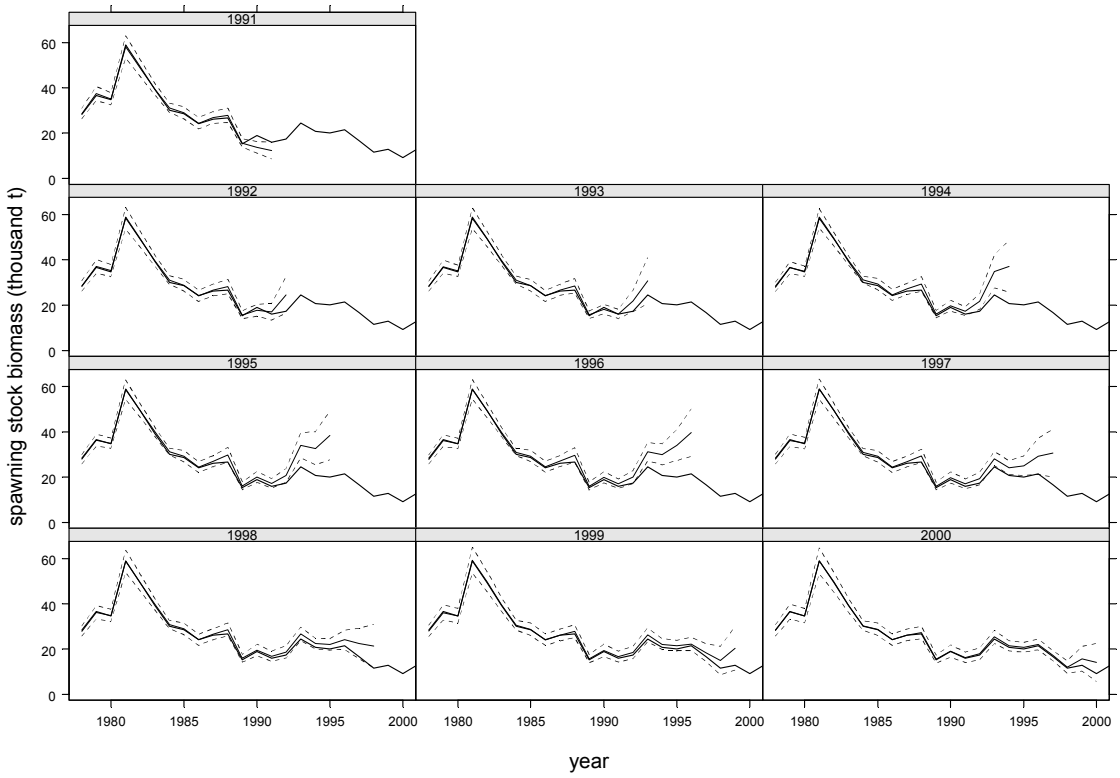


Figure 7 Retrospective plots