Resource Management Committee

ICES CM 2003/D:03
Ref. ACFM, G

REPORT OF THE

Working Group on Methods on Fish Stock Assessments

ICES, Headquarters
29 January–5 February 2003

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

Carl O'Brien (Chair)	UK (England & Wales)
Noel Cadigan	Canada
Chris Darby	UK (England & Wales)
Yuri Efimov	Russia
Kristin Guldbrandsen Frøysa	Norway
Daniel Howell	Norway
Sigurdur Thor Jónsson	Iceland
Knut Korsbrekke	Norway
Yuri Kovalev	Russia
Sarah Kraak	Netherlands
Benoit Mesnil	France
Coby Needle	UK (Scotland)
Martin Pastoors	Netherlands
Dankert Skagen	Norway
Stuart Reeves	Denmark
Marina Santurtún	Spain
Victor Tretyak	Russia
Dmitri Vasilyev	Russia

1.2 Terms of reference

The **Working Group on Methods on Fish Stock Assessments** [WGMG] (Chair: C. O'Brien, UK) will meet at ICES Headquarters from 29 January – 5 February 2003 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 approaches, methods and software tools for the investigation of management strategies;

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

g) discuss the choice of model structure (age-based, length-based, age-length) taking into account stock dynamics, biology and data availability; and

h) review and further develop the specification of software to generate stock assessment data, taking into account spatial, temporal and multispecies characteristics of fisheries.

WGMG will report by 15 March 2003 for the attention of the Resource Management Committee, the Living Resources Committee and ACFM.

1.3 Scientific justification for this meeting

Prior to the last meeting of WGMG in 2001 (ICES 2002a), ICES had lacked an active forum for developing new methods and investigating properties of fish stock assessment methods. ACFM had discussed this problem and concluded that there still remained a strong need for regular meetings of the Working Group on Methods on Fish Stock Assessments. Primarily, the Group was to work with the estimation and projection procedures in a statistical context. The Working Group would concentrate on methodological procedures and would not evaluate any case studies in any

detail. The Methods Working Group would also, as part of its remit, serve as the ICES' focal point for the discussion of new methods.

WGMG had started to address issues of data quality, modelling and stock assessment practice at its last meeting. The Group had focussed 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 within one meeting. The likely causes of the retrospective problem have become clearer and a way to proceed in the development of a solution had been proposed (ICES 2002a). Much work still remained to be undertaken and the group suggested that this second meeting should be held.

In addition to the agreed ToRs for each meeting of WGMG it was recognized that there is a requirement for the Group to be flexible enough within its remit to deal with ad-hoc requests from the ICES Advisory Committee on Fishery Management (ACFM) that are of a methodological nature. One such request was received prior to the present meeting and is detailed in the next Section 1.4. An answer to the special request is presented in the later Section 8.5 of this report.

1.4 Special request to ICES

Extract from a letter dated 21 November 2002 to ICES from the Royal Ministry of Fisheries, Norway on behalf of the EC, the Faroe Islands, Greenland, Iceland, Norway and the Russian Federation:

During the coastal state meeting on blue whiting in Oslo November 7 -8 2002 the assessment and management advice given by ICES was presented and discussed. Further, assessment and prediction based on an alternative model (ISVPA) developed by Russian scientists was presented and compared to the official output from the model applied by ICES (AMCI). The two models give significant different estimates of the stock size. It is known that both these models are used by ICES in 'The Northern Pelagic and Blue Whiting Working Group" along with other available tools.

1. The parties noted that similar discrepancies exist for the assessment of the Norwegian Spring Spawning (Atlanto-Scandian) herring stock and that a request to ICES to evaluate the two assessment models with respect to Norwegian Spring Spawning herring has been put forward by the Russian Federation on behalf of the coastal states. The Parties request ICES to extend these evaluations to also include assessment of blue whiting.

Within ICES and ACFM, the Norwegian Spring Spawning herring is assessed using both ISVPA and SeaStar; whilst blue whiting is assessed using AMCI and ISVPA (ICES 2002b). The originators of the computer programs: Dankert Skagen (AMCI), Sigurd Tjelmeland (SeaStar) and Dmitri Vasilyev (ISVPA); were invited to attend this meeting of WGMG in order for an evaluation of the models to be undertaken. Unfortunately, Sigurd Tjelmeland was unable to attend this meeting of WGMG but will prepare an evaluation of SeaStar for the next meeting of the Northern Pelagic and Blue Whiting Fisheries Working Group [WGNPBW] to be held shortly after this meeting of WGMG in 2003. Therefore WGMG will only address the special request to ICES with respect to blue whiting, and not Norwegian Spring Spawning herring, at this meeting.

WGMG agreed to address the *methodological issues* behind this request; namely, to consider ways in which to deal with different assessment model formulations which are apparently equally valid but nonetheless lead to different perceptions of stock status. However, given the absence of both documentation and software for SeaStar, WGMG could only address the evaluation of the two programs AMCI and ISVPA at this meeting. Their evaluation and comparison is presented in Section 8 of this report; with respect to both simulated data generated at this meeting and blue whiting stock assessment data based upon the last ICES assessment (ICES 2002b).

1.5 Structure of the report

The terms of reference (ToRs) are addressed within the seven main sections of the report. Specifically, ToR a) is addressed within Section 5 of the report, ToR b) is addressed within Sections 5 and 6, ToR c) is addressed within Section 4, ToRs d) and e) are addressed within Section 7, ToR f) is addressed within Section 4 and ToRs g) and h) are addressed in Section 2. The special request to ICES is dealt with in Section 8.

In Section 2, the background to the choice of assessment model structure (e.g. age-based, length-based or age-length based) is briefly discussed and a specification for software to generate simulated data in order to investigate hypotheses is presented. Section 3 details the simulated and real data sets that were used in sensitivity testing during this meeting. Section 4 addresses issues pertaining to software tools for stock assessment purposes and current plans for the further development of stock assessment tools by members of the Working Group; Section 5 addresses the use of local

influence diagnostics to detect mis-specifications in either the data inputs or assumptions to sequential population analysis; and Section 6 presents details of analyses using SURBA and CSA, together with a description of the application of Benford's law to fisheries stock assessment. In Section 7, the issue of applying methods in the provision of fisheries management advice is addressed and Section 8 deals with the use of AMCI and ISVPA for the stock assessment of blue whiting. An extensive selection of graphical outputs and diagnostics have been produced from the assessment models investigated in Section 8 and these have been collated into a separate Section at the end of Section 8 for ease of reference. A compilation of the Working Group's recommendations from the main body of the report is provided in Section 9; together with details of further work needed to be undertaken. The Appendices A and B present brief technical details of catch-survey analysis (CSA) and survey-based assessment (SURBA), respectively. The Working Paper by Cadigan & Farrell (WAB1) is reproduced in Appendix C for completeness.

2 MODEL STRUCTURE AND DATA SIMULATION

2.1 Introduction

ToRs g) and h) for this meeting of WGMG (Section 1.2) required the Group to discuss choice of assessment model structure and to further develop the specification for software to generate simulated stock assessment data. Such software is required to produce data sets with fully known properties. These two issues are related as the use of simulated data is a useful tool with which to investigate choice of model structure. For this reason these two issues are considered together here in this Section 2.

2.2 Background

The term of reference g) makes the distinction between age-based, length-based and age-length models. To some extent these distinctions are rather restricting as they cover only a limited range of the models available for stock assessment; some (such as production models) do not require any of these data-types, whereas others can use all of these data types and more besides. Models may also be divided between those which use catch data to directly reconstruct population levels (e.g. VPA), and models explicitly simulating population dynamics (e.g. Gadget; Stefánsson & Palsson 1998). The first type requires data input in order to run, whilst the second may be completely determined by input parameters. Despite the diversity of available model types, catch-at-age models are by far the most widely used assessment approach within the ICES context. The reasons for this are to a large extent historical and institutional (Skagen & Hauge 2002), and given that many ICES stocks now have relatively long time-series of catch-at-age data, the choice of such methods is natural in many cases. However, it is still useful to question whether such models should be the automatic first choice in all cases. A case in point is the assessments of *Nephrops*, where routine age determination is not possible, so pseudo catch-at-age data are derived by applying a splitting procedure to catch-at-length data, in order that an age-based assessment can be performed. Given the absence of true age-based data in this case, the use of a catch-at-age approach would not seem to be the most natural choice.

Such concerns about the applicability of catch-at-age models also apply in other cases, such as where catch-at-age data exist but are of poor quality, or where only limited data are available such as for a new or developing fishery. It is clearly desirable that the assessment model used is appropriate for the stock it is applied to. This in turn is determined by the nature of the data available and the kind of questions the assessment is intended to answer. Increasing model complexity also means increased data demands. To give an example in relation to age-length models, if a stock shows considerable variation in growth rate to the extent that it influences catchability, then it would seem desirable to include length data and to model these effects as part of the assessment procedure. However, if there are problems in the reliability or coverage of catch or tuning data, then it is possible that such a model would perform less well than a simpler catch-at-age model which disregarded these length-based processes. These trade-offs mean that it is not appropriate to give hard-and-fast rules concerning which sort of assessment model is most appropriate for a given stock. Choice of appropriate model should be governed by both theoretical and practical considerations. The different models should be examined from a mathematical point of view, to see which assumptions are made and how the assumptions influence the resulting model. Particular effort should be directed at clarifying the implicit assumptions and how they influence the results. It should also be made clear what assumptions are made concerning the data structure. The theoretical studies should be combined with practical studies on simulated and real data sets. Examples of such analysis are given in Section 5 and Cadigan & Farrell (WAB1; Appendix C). The artificial data sets used in these studies should represent the characteristics of the stock and the real-world data, and provide a known outcome against which the model estimates can be evaluated.

In order to generate test data which can be used to evaluate the applicability of different classes of model, rather than just catch-at-age models, it would be necessary to generate catch data which retain the full level of complexity of real data; e.g. by length, age, and aggregated catch weights. This would enable that, for example, length and age-based approaches and production models could be applied to the same data. Through the use of such test data, it should then be possible, not only to evaluate the relative merits of different catch-at-age methods in a given set of circumstances, but also of the applicability of other classes of model of both lesser and greater degrees of complexity. Of particular interest might be the possibility to investigate assessment approaches to developing and data-limited fisheries, such as those on some elasmobranch species. In addition, it should be recognised that in some cases it may not be possible to give detailed information on the state of the stock, irrespective of the model used, and tests using simulated data may shed light on when and how such cases may arise.

Given the diversity of available models, choosing the correct tool for a particular job becomes a non-trivial task. A number of factors will influence this decision. Highly complex age- and length- structured models may be required in order to model certain processes. Data availability will impose limits on the feasible complexity of the model used.

Limitations may also be imposed by lack of knowledge of the processes in the modelled system. There are thus conflicting pressures towards a greater or lesser level of complexity.

In order to choose an appropriate model for any given task several things must be known – the purpose of the modelling task; and features of the studied population, data and models. Each of these is discussed in turn next.

The purpose of the modelling task

The model chosen must be appropriate for the problem to be solved. For example, stock assessment to support management decisions may require a different model than that appropriate for exploring the population dynamics of a species.

Features of the studied population

In particular, which are the critical features that need concentrating on in order to answer the question at hand.

Features of the data

The possible deficiencies in the data coverage and reliability.

Features of the models

What are the strengths and differences of the different available models in relation to these issues?

It is to address this last point that a wide range of simulated data sets are required. There currently exist a number of different simulated data sets (e.g. ICES 2002a; Mesnil (WB2); Restrepo *et al*. 2000). These have been generated in different times and places to meet specific needs. These data sets address some of the current issues in stock assessment modelling, such as trends in catchability (Sections 3, 5 and 6). It is not currently possible to access all of these data sets to choose the one that best meets the problem being considered, nor is it clear that suitable data sets exist for all problems which may need to be examined. In general a method of generating custom-designed data would be of great utility, allowing specific data sets to be created to meet the needs of specific problems.

2.3 Data simulator

There is a need for a standardized data simulator for use in testing and developing fisheries models. This software would allow for the rapid creation of simulated data sets suitable for testing the performance of a variety of models in a range of situations. It would also be helpful to have **a standardized data set which can be used for initial testing of models, so that any new model designs or formulations can be run against a single common data set**. It is important to note that it is not enough to merely produce such a tool, it must also be widely distributed, available, and user friendly. This requires both adequate documentation of the tool and also easy access. It is therefore suggested that any **final program and sample data sets, with accompanying documentation, should be placed on a freely accessible website, which could be hosted at ICES Headquarters**. The code for such a system should, if possible, be open source, and able to run on as many computers as possible. This means that proprietary languages, such as SAS, should be avoided.

The ability to create specific data sets with defined characteristics is important. Data sets should be constructed that violate the common assumptions in assessment models, and which replicate real-world situations. Such a suite of data sets can then be used to test the performance of different models, with different assumptions (e.g. separability), under this range of situations. This will indicate which models and assumptions perform well for different kinds of problems, and thus which are appropriate for different real-world situations. For instance, such data sets would be of use in analysing the factors involved in the retrospective problem, as previously discussed in Section 5.1 of the Report of the last WGMG meeting (ICES 2002a).

A proposal was put forward for a standardized data simulator in that report. This proposal was for an age-structured simulation. This is too limited to address many of the issues we are faced with today, and is clearly not sufficient to allow for comparisons between age-based, length-based and age- and length-based models. This framework is extended here to include a greater degree of complexity, and a more concrete structure is proposed. Sub-samples of this complex simulation can then be taken to produce the form of data required by a particular assessment model.

The data simulator must satisfy a number of criteria. It must be: flexible, robust, easy-to-use and implemented. Obviously these criteria are, to a certain extent, incompatible. It is suggested that a two-stage approach should be used. An initial specification defines a simulator that is complex enough to produce useful data, but simple enough to have a chance of being written. In addition, specifications are also produced that give examples of extensions to this basic framework which could be added at a later date. It is to be stressed that these involve a considerable increase in complexity, and should not be undertaken until the basic form is stable and useable. In general we are avoiding any form of feedback within the population simulation at this stage. Thus factors such as predation or dynamic responses in fishing effort to variable year classes are avoided here.

In order to provide maximum flexibility, a principle is adopted that any parameter required by the simulation program can be varied at every time step. In practice it is unlikely that users will wish to directly specify each value by hand, thus a system must be provided to automate this process. This will generate full time-dependent parameter values based on a user choice from a small number of options. The resulting parameter set should then be available for possible manual editing to allow for maximum flexibility.

In order to preserve an understandable and extendable structure the system will be split into three distinct parts, a pre-processor, a simulator, and output filters. These will remain distinct, will data passing between them in the form of ASCII files. There will be no iteration built into the system, data $\mathbf{F}_{low}$ will be in one direction only. This structure will allow different parts of the system to be developed and extended independently of each other, and permit manual editing of the files between each stage if required for specific problems. This structure is summarized in Figure 2.3.1.

Details

Pre-processor

This will take user inputs and convert them to produce parameter values for every time step that the simulator will run. The output of this procedure will be stored as human-readable ASCII file(s). By using a pre-processor in this way the maximum amount of time-dependant flexibility can be maintained, without overwhelming the user by requiring each value to be entered manually.

Simulator

This will read in the time-dependant parameter values generated by the pre-processor and conduct a simulation of the population and catches based on those values. The simulator will output highly disaggregated data on the population and catches through time in an ASCII format. This will represent the *truth* against which models can be tested.

Output filters

The selection filters will take the *true* highly-structured output from the simulation, and perform a series of manipulations on it. Survey data will be generated as a statistical function of the simulated population. Data from the survey and the catch will be aggregated to whatever level is desired for the model test being undertaken. Filters will introduce errors degrading the *true* data into the sample that will be used in testing the assessment models. Random error will allow for multiple samples to be taken from a single *truth*. Systematic errors (e.g. under-reporting or discarding) will also be added at this stage. Finally, the data will be converted into the correct format for use in stock assessment models. Each process should be kept separate. The user can they decide to use as many, or as few, of the filters as required.

It is anticipated that the following levels of detail can be included in the basic simulator:

Age and length structure

The simulator will be fully age- and length- structured, with length classes specified by the user via the pre-processor.

Time step

The length of the time step used in the system will be definable. Yearly, quarterly or monthly output should all be available.

Population groups

The system should be able to handle multiple population groups. A *population group* is defined to be a group of fish that are treated as having uniform characteristics. This may be a species, a stock within a species, or a maturity (life) stage within a stock within a species. There may be multiple species within the system. Each species may be split into multiple stocks, and into an immature and a mature component. Except for maturation between immature and mature individuals there will be no direct interactions between the population groups.

Simple area structure

Each population group should be able to exist on one or more areas. Movement between the areas will be governed by a pre-defined matrix for each time step, and characteristics of a given population group (e.g. growth) will not vary between areas.

Fleets

It should be possible to define multiple fleets, each fishing on one or more areas and catching one or more population groups.

This structure preserves a high degree of complexity. A wide variety of flexibility is possible, allowing for many real-life problems to be replicated in the simulated data set. It will be possible to construct sample stocks with a wide range of species characteristics (e.g. fast or slow growth, long or short life spans, etc). Because all parameters will be generated by the pre-processor and read in for each time step it will be possible to include time-dependant variation in selection, natural mortality, growth, recruitment, and migration (in multi-area simulations). The filters that generate samples for use in the stock assessment models will be able to produce time-dependant variations in survey catchability, discarding and mis-reporting, and biases and errors in the sampling procedure (e.g. possible aging difficulties). It will also be possible to generate different output data sets with conflicting signals, to create a few years of exceptionally poor data or to create individual years with missing data. The time-dependant changes can be user defined.

It may be possible to simplify this structure further while still retaining the ability to produce useful datasets. Removing the possibility of multiple fleets would be possible, but at the cost of preventing simulation of multiple fleet situations. Removing the ability to include multiple population groups would remove the ability to consider mature and immature fish separately, and prevent reconstructions of situations involving by-catch. Removing the capability of having multiple areas would make it impossible to simulate a situation where information is available from only part of the geographical areas covered by a stock, or where quality of information is different in separate areas. Any or all of these simplifications would make all parts of the system easier and faster to write.

2.3.1 Further enhancements

Because the basic goal must be to produce a system that can actually be implemented, a number of more complex features have been left out of this specification. It should not be taken from this that these topics should never be included. Rather, the initial target will be a stable, understandable, and usable system. The topics listed below could then be incorporated into this at a later date. In general these all involve feedback within the simulation itself. This not only increases the difficulty of programming the simulator, it also makes constructing simulated data sets significantly more complex. The feedback loops will make it difficult to predict how variations in input parameters will affect the final simulation, and thus make parameter selection considerably more difficult.

Species interaction

An ecosystem simulation should include cannibalism and predation between two or more species. This is not included **at present, but is an area that should be examined in future.**

Dynamic fisheries

Fishery patterns may vary to account for stock density or large year classes, with the fleets changing selection pattern or geographical location in response to fish-stock dynamics. This is not accounted for here, but would be a worthwhile extension.

2.4 Details

No attempt is made to specify the equations governing the simulation at this stage. Rather, the processes that must be considered are outlined. It is essential that any simulations conducted are fully repeatable. Therefore any random numbers required by the system must be capable of being started using user-defined random number seed(s). Parameter choices and the random number seed(s) should be recorded in such a way that the simulation can be reproduced at a later date. It is also vital that full and clear documentation be made available.

2.4.1 Pre-processor

The pre-processor should present a series of choices to create the parameters needed by the simulator. The user should be able to select constant or time-dependant values for simulation parameters, with one or more functions being provided to define these parameters. The pre-processor will need to provide all of the parameters required by the simulator. It should start with a set of default values corresponding to a single standard benchmark data set, only changes from this default would need to be selected by the user.

2.4.2 Simulator

The simulator would need to output ASCII files containing details of the population and the catches over time. It would also need to produce data on mortality (both natural and fishing) over time. These files would then pass to the output filters. The following processes will need to be modelled in the simulator:

Recruitment

Either as pre-set time-varying values, or via a recruitment function.

Growth

Growth would need to be a function, and growth in both length and weight would be required.

Natural mortality

Natural mortality would need to be an age- and length- based function.

Fishing

Catches must be specified with a combination of a length selection, and some measure of the magnitude of the catch and mortality (for example as actual catch, F or effort). The selection should be at least length-, and possibly also, age-based.

Maturation

As a function of age and length, and possibly weight.

Migration

This will be read from an input file, with different values read for each time step.

2.4.3 Output filters

In order to make the filters powerful and flexible it is proposed that they be written as a series of independent modules. Each would take an ASCII file as input and conduct an operation on that file, and output the results as a file of the same format. This would allow multiple filters to be used in any sequence required by the user. The following types of filter would be required:

Aggregation

The data required by the assessment models may need to be at a coarser scale than the output of the simulation model. Filters to perform this aggregation would be needed.

Survey

If the survey is to be treated as a statistical sample from the 'true' population then filters would need to be written to create this sample. The results could then be passed through error filters if required.

Errors

The filters would need to introduce errors into the output data. User definable random noise would enable multiple samples to be taken from a single *true* population for use in uncertainty estimates. Systematic errors (either a trend or step-wise) would be able to represent factors such as discards and mis-reporting, or changes in catchability. Multiple uses of simple filters would enable complex time-dependant variations to be produced.

Summary measures

Summary measures from the population (e.g. fishing mortality, SSB) will prove useful for assessing model performance. All of the standard output of assessment models should be replicated.

Formatting

A final series of filters capable of formatting simulated data ready for use in standard assessment models would increase the utility of the system.

2.5 Implementation

There are two possible routes to implementation. Either a package can be specifically written for this purpose, or an existing simulation model can be modified to the requirements presented here. Both approaches have advantages. If the simulation is specifically written for this purpose then it is likely to be smaller and easier to understand than software adapted from a different purpose. It will exist independently of any model, and thus not be tied to the changes and development of any specific model. The source code of the simulation should be included in the distribution, increasing both the degree of confidence users will have in the system and their ability to modify it.

On the other hand adapting an existing model may require less effort, and may thus lead to an increased chance of this system actually coming into being. If this option is chosen then the structure suggested here, of separate pre-processor, simulator, and output filters should be retained, with pre-processor and output filters being written around the existing model. Using the pre-processor will reduce the difficulties in setting up the complex data files required by age- and length- structured simulation models. For example Gadget (Stefánsson & Palsson, 1998) can provide most or all of the functionality described here, but is not simple to use. There may be other models that are also suitable for consideration, such as that described in Punt *et al*. (BEH1).

WGMG proposes that funding be sought to support the development of this system as stand-alone software influenced by knowledge gained from existing simulation models. In the shorter term it may be possible to begin by writing simple pre-processor and output filters to use an existing simulation model, such as Gadget, as a population generator. The modular nature of these parts of the system would allow such a beginning to be expanded incrementally as specific problems need to be addressed. It would also enable a number of workers in different sites to collaborate in the development, by separating the project into a number of discrete tasks. Finally it would allow the population generator to be replaced at some future date while retaining the pre-processor and output filters.

It would also be useful to take a number of the currently available simulated data sets, and the corresponding data simulators, and place them on a web or ftp site. Along with the data sets themselves there should be a description of their characteristics and an indication of which data sets are appropriate for which sorts of problem. This repository could then be expanded to include the flexible data simulator system described here when this has been written. A series of data sets designed for specific problems can then be generated and made

available. In order to preserve its utility this web site, and the associated data simulation system, must then be maintained.

To this end, WGMG proposes the following ToR for their next meeting:

> ***To examine software capable of generating simulated data, and agree an initial suite of standard data sets for use in model-testing and evaluation that will be made generally available from the ICES website.***

The Working Group considered that as a first step towards the testing of assessment models, data sets similar to those used at the Reykjavik meeting of this group (ICES 1988) should be generated and made available within the current year for inter-sessional work prior to the next meeting in 2004. The data sets should simulate different types of fisheries exploiting the stocks with *biases* and levels of noise that represent the current perception of stock assessment issues. Examples are the trend in catchability described in Section 3 of this report, changes in selection by the fleet, mis-reporting etc.

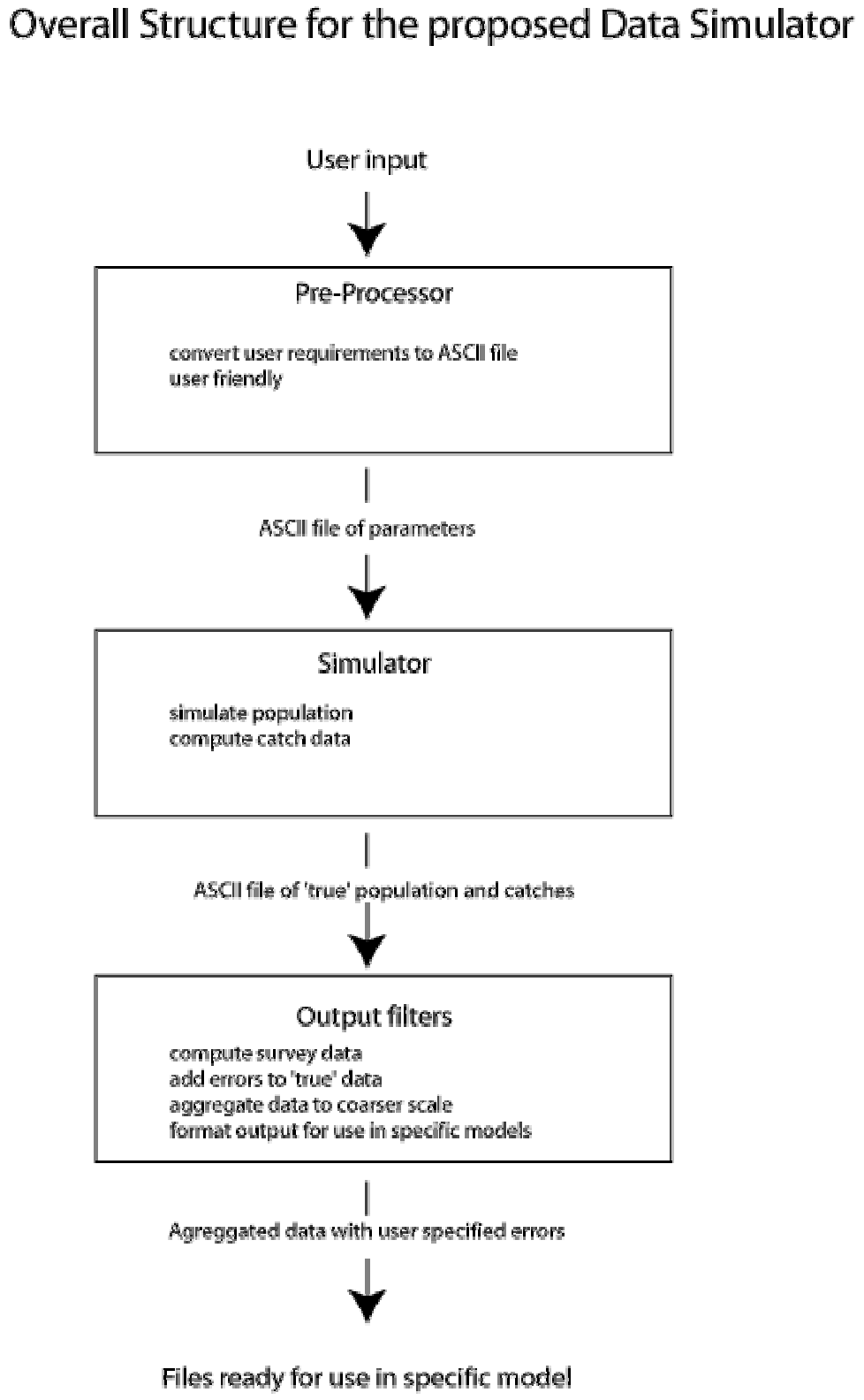


Figure 2.3.1 Overall structure of the proposed data simulator.

3 SPECIFICATION OF DATA SOURCES

In order to address the ToRs a), b) and c) the Group agreed to apply candidate methods of stock assessment to simulated data sets during the meeting whose properties were known. The details of the description of the data sets are presented in Sections 3.1 and 3.2. These simulated data sets were also used in the evaluation of the software AMCI, ISVPA, CSA and SURBA. In addition to the analyses of simulated data sets, the ICES stock assessment data for blue whiting was investigated using ICA, AMCI, ISVPA, CADAPT and XSA. The results are presented and discussed in the later Section 8.

3.1 Simulated data without noise

This set of *clean* simulated data (no measurement error in catch or survey indices) had been intended for testing software rather than for the *evaluation of* methods. It was nevertheless used for that latter purpose because it was immediately available.

The data generation procedure used is an adaptation of that used by Restrepo *et al.* (2000). It considers an age-structured population comprising 15 ages (1-15, no plus-group: contributions of ages 16 and older are ignored). A constant natural mortality of 0.2 is assumed for all ages and years. The population structure in the first year is generated under equilibrium and with a recruitment of 14558 individuals. This population is then simulated forward over 41 years, with nominal fishing mortality maintained at $0.5*\mathbf{F}_{MSY}$ ($\mathbf{F}_{MSY} = 0.166$) during a burn-in period of 17 years, then increased gradually to twice $\mathbf{F}_{MSY}$, maintained there during years 27-33, and subsequently reduced toward half $\mathbf{F}_{MSY}$ ("two-way trip"). The fishery has a specific age-dependent exploitation pattern which is fixed over the period. Recruitment in each year is stochastic about a Beverton-Holt stock-recruitment relationship, with auto-correlation. The specifications are summarised in Table 3.1.1.

In addition, it is further assumed that the population is length-structured. Modal lengths-at-age follow a von Bertalanffy growth schedule. Lengths within age are assumed normally distributed about the mode, with SD= MIN(intermodal distance/2, mode/10). This 'rule' for SD is fairly arbitrary but the choice has no reason to affect the tests conducted here; it allows SD to increase first then decrease with age. The distributions extend ±3*SD about the mode. The relative length composition within each age does not change in time.

The survey is length selective with a logistic selection curve. The selection factor is 0.4 and the ratio of range to L50 is about 0.4 ('cod-like'); the selection range thus increases (flatter curve) with mesh and L50 (=SF*mesh). Three instances of this survey were simulated, supposing that mesh was 20, 50 or 80 mm, with a survey's nominal catchability of 0.001 for the fully selected animals. In a fourth instance (q-trend scenario), it was assumed that the nominal catchability of the 80-mm survey had increased by 4% per annum during the last 10 years.

This scenario emulates a model mis-specification error for the methods that assume constant catchability in the tuning fleets. It was found, however, that this was **insufficient to create a significant retrospective pattern in estimates** obtained with the various methods tried during this meeting (see Section 5). A likely cause is that the simulations above assume relatively low fishing mortality compared to M, and a steady decrease in F over the same period when a q-trend is supposed to occur. Several modifications were tried (creating a retrospective bias proved nearly as difficult as eliminating one in real assessments), leading to a scenario where F was about 0.2-0.5 overall, with a plateau in recent years, and a strong 7% annual trend in the tuning fleet was assumed (this is referred to as the *high F scenario*).

Eventually, data for only the final 20 years are retained (re-coded 1 to 20) and are formatted in line with the specific requirements of the assessment methods (CSA, SURBA, XSA, QLSPA) investigated in the later Sections of this Report.

3.2 Simulated data with noise

In order to examine the behaviour of methods, which assume noise in the catch-at-age and index series, the data generation procedure discussed in Section 3.1 was modified to produce random stochastic noise in the age disaggregated output. Fishery or survey data were modified using a random variable drawn from a specified log-normal distribution with a known coefficient of variation (CV).

The CV of the fishery was 0.2 and of the two indices 0.4. The simulator was used to generate a single realisation of the fishery and population dynamics. As with the *clean* simulated data series, described in Section 3.1, one index series was generated with a 4% trend in catchability at all ages over the last ten years of the time-series.

3.3 Blue whiting combined stock (Subareas I-IX, XII and XIV)

Blue whiting is widely distributed in the eastern North Atlantic. It consists of several populations with genetic *leakage* between them, but it is treated as one stock for the purpose of ICES stock assessment, as it so far has not been possible to define an unambiguous border between populations.

The analytical assessment is based on catch data, acoustic surveys and commercial CPUE series data. Details of the stock assessment data files are to be found in the Report of the Northern Pelagic and Blue Whiting Fisheries Working Group [WGNPBW] (ICES 2002b). These stock assessment data files have been used without modification.

Table 3.1.1 Details of the specification of the age-structured population and fishery simulation. Note that the notation *n*value* used for the specification of maturity-at-age and selectivity-at-age signifies that the *value* applies to the next *n* ages.

Natural mortality	M = 0.2, all ages and years
Growth	K = 0.15; L_∞ = 100; t_o = 0
Length-weight	a = 0.00001; b = 3
Maturity-at-age	4*0.0, 0.3, 0.5, 0.7, 0.9, 0.95, 6*1.0
Recruitment:	
type	Beverton-Holt: $R = S/(\alpha+\beta S)$
parameters	α = 0.67945; $\beta = 5.6621 \times 10^{-5}$ (steepness = 0.7)
variability	Log-normal; CV = 0.6; auto-correlation ρ = 0.5
Selectivity-at-age:	
fishery	0.05, 0.1, 0.3, 0.7, 0.9, 10*1.0
survey	variable depending on assumed mesh size
Survey nominal q	0.001

4 SOFTWARE TOOLS FOR STOCK ASSESSMENT PURPOSES

4.1 Testing, validation and certification of software

Software that is used by ICES to provide advice is generally written and produced by individual scientists or national laboratories. Attempts at ensuring the quality of such software have been made on several occasions in the past; both by this Group, WGMG, as well as by dedicated ICES Study Groups (SGFADS: ICES 1998/ACFM:9). WGMG discussed the proposal from the SGFADS at its previous meeting and endorsed the proposal for an acceptance procedure with some minor modifications. The proposal included specification of standards for documentation and minimum tests that the software runs properly. Furthermore, after passing this stage, it was recommended that ACFM be responsible for endorsing assessment software and the nomination of two reviewers (ICES 2002a).

Since the last meeting of WGMG, several programs have been submitted to ICES for evaluation. A number of these, as well as prototype methods still under development, were presented to this meeting of WGMG and are further described in Section 4.2.

The evaluation process in ICES has not yet proven to be fully effective, and currently may represent an obstacle to the implementation of improved software or the provision of solutions to known problems with existing software used by stock assessment Working Groups. There is a trade-off between ensuring the quality of new (and existing) methods and the need for methodological development. It may also become a problem, in the not too distant future, that ICES will have to continue to apply standard, approved software in cases where these are known to be deficient; when better methods have been developed but not approved.

WGMG discussed this topic during this meeting and identified that the approval of a new method includes at least three components:

- Checking that the program can handle the problems it is supposed to handle. This would include testing that the program can reproduce artificial data when all assumptions are correct, as well as robustness to noise. This may be done by requiring documentation of results using appropriate artificial data sets.

- Evaluation of the method as such. This includes the way inferences are made from the data, assumptions and constraints, and what kinds of problems the method is supposed to be able to handle and under which circumstances one should expect it to fail, as well as strengths and weaknesses when compared to other methods. This kind of evaluation naturally belongs in a forum like that of WGMG. However, the WGMG does not have the resources to fully investigate each candidate new method during its meetings. Doing so could require dedicated meetings, perhaps by a sub-group of the membership of WGMG.

- Certification of the software to commercial standards includes extensive testing and validation to ensure that it is free of errors and *bugs*, that the program code is in accordance with the documentation and that the code meets international standards. This task definitely is outside the remit of WGMG. ICES needs to reflect upon how far such a process should go. Stringent requirements for certification may preclude further methodological and software development.

In conclusion, there is a strong need for ICES to have in place a formal process whereby software is tested, evaluated and approved for general use. However, such a process may fall short of formal certification of software.

In any case, approval of a method should not be taken as a guarantee that it can be used uncritically. Each method will have its own limitations, and if methods are used on stocks where underlying assumptions are violated, the results will potentially be misleading, even though a method may have been approved for general use. It is naïve to expect that any method will be universally applicable and this implies that stock assessment Working Groups will need a wide range of methods and software at their disposal.

To alleviate a number of these concerns, it is proposed that WGMG work inter-sessionally before its next meeting to **draft guidelines on the formal procedures to be adopted by WGMG for the testing, evaluation and validation of software for use by ICES stock assessment Working Groups**.

4.2 Programs presented to the Working Group

4.2.1 AMCI

Model	**AMCI**
Version	**2.2 (year: 2002)**
Model type	A separable model is applied to the whole assessment period. Selection can be allowed to change slowly according to the signal in the catches. The rate of change is determined by the user by specifying a gain factor for the influence of the current catch data. One extreme is then to keep the selection fixed (as in ICA). The population is projected forwards in time.
Selection	The selection at one age can be specified as the average over some other ages, but this specification cannot include any multiplier. The selection at oldest age is estimated unless it is linked by the user to some other age.
Estimated parameters	Recruitment, initial stock numbers, annual fishing mortalities, selection-at-age by year, catchability-at-age (and year), natural mortality, quarterly distribution of fishing, quarterly distribution of stock by area. The user decides upon which of these to estimate; the remainder are kept at fixed values.
Catchabilities	Catchabilities are in principle modelled as separable, but the age factor can be allowed to vary slowly using the same principle as for the selection-at-age in the catches. In practise, it will most often be kept fixed, and it then behaves as it does in ICA. Proportionality between index and stock abundance is always assumed. The proportionality can be fixed to the value one.
Plus group	The plus group is modelled as a dynamic pool. The fishing mortality assumed for the plus age can be estimated, or linked to some younger age. The fit of the modelled plus group is included in the objective function unless specified otherwise.
Objective function	There is a variety of objective functions available but most often, the weighted sum of squared log residuals is used. Weighting is decided by the user. AMCI does some implicit weighting internally which implies that the weights assumed in ICA and AMCI are not directly comparable.
Variance estimates/ uncertainty	'Variances' of the parameter estimates can be derived from the Hessian, which is computed directly. There are also options for estimating uncertainty by parametric or non-parametric bootstrapping.
Other issues	AMCI allows the incorporation of tagging data and SSB indices as additional sources of data. It allows for multiple fishing fleets and multiple areas, defining local partial fishing mortalities. Distribution by area is specified as parameters but there is no migration model yet.
Program language	FORTRAN 77. No external libraries required.
References	Draft manual available but no formal publications yet.

4.2.2 ISVPA

Model	**ISVPA**
Version	**Year:2002**
Model type	A separable model is applied to one or two periods, determined by the user. The separable model covers the whole assessment period
Selection	The selection at oldest age is equal to that of previous age; selections are normalized by their sum to 1. For the plus group the same mortality as for the oldest true age.
Estimated parameters	
Catchabilities	The catchabilities by ages and fleets can be estimated or assumed equal to 1. Catchabilities are derived analytically as exponents of the average logarithmic residuals between the catch-derived and the survey-derived estimates of abundance.
Plus group	The plus group is not modelled, but the abundance is derived from the catch assuming the same mortality as for the oldest true age.
Objective function	The objective function is a weighted sum of terms (weights may be given by user). For the catch-at-age part of the model, the respective term is: • sum of squared residuals in logarithmic catches, or • median of distribution of squared residuals in logarithmic catches MDN(M, fn), or • absolute median deviation AMD(M, fn). For SSB surveys it is sum of squared residuals between logarithms of SSB from cohort part. For surveys; for age- structured indexes it is SS, or MDN, or AMD for logarithms of N(a,y).
Variance estimates/ uncertainty	For estimation of uncertainty parametric conditional bootstrap with respect to catch-at-age, (assuming that errors in catch-at-age data are log-normally distributed, standard deviation is estimated in basic run), combined with adding noising to indexes (assuming that errors in indexes are log-normally distributed with specified values of standard deviation) is used.
Other issues	Three error models are available for the catch-at-age part of the model: • errors attributed to the catch-at-age data. This is a strictly separable model ("effort-controlled version") • errors attributed to the separable model of fishing mortality. This is effectively a VPA but uses the separable model to arrive at terminal fishing mortalities ("catch-controlled version") • errors attributed to both ("mixed version"). For each age and year, F is calculated from the separable model and from the VPA type approach (using Pope's approximation). The final estimate is an average between the two where the weighting is decided by the user or by the squared residual in that point. Four options are available for constrainting the residuals on the catches: 1. Each row-sum and column-sum of the deviations between fishing mortalities derived from the separable model and derived from the VPA-type (effort controlled) model are forced to be zero. This is called "unbiased separabilization" 2. As option 1, but applied to catch residuals. 3. As option 1, but the deviations are weighted by the selection-at-age. 4. No constraints on column-sums or row-sums of residuals.
Program language	FORTRAN 77.
References	Vasilyev, D.A. (2001). Cohort models and analysis of commercial bioresources at information supply deficit. VNIRO Publishing: Moscow.

4.2.3 LTEQ

This is a computer program for calculating long-term equilibrium distributions of SSB and yield as a function of the realised fishing mortality; taking into account natural variations in recruitment, growth and maturation.

LTEQ uses an iterative procedure to arrive at the stationary distributions. This has the advantage that there is no risk of transient effects, but it precludes inclusion of feed-back from the population to the growth and maturity; e.g. by density dependent growth. The program allows for two independent fishing fleets in order to illustrate the trade-off between conflicting fisheries. It screens over a range of fishing mortalities that produces output that can be converted to the kind shown in Figure 4.2.3.1.

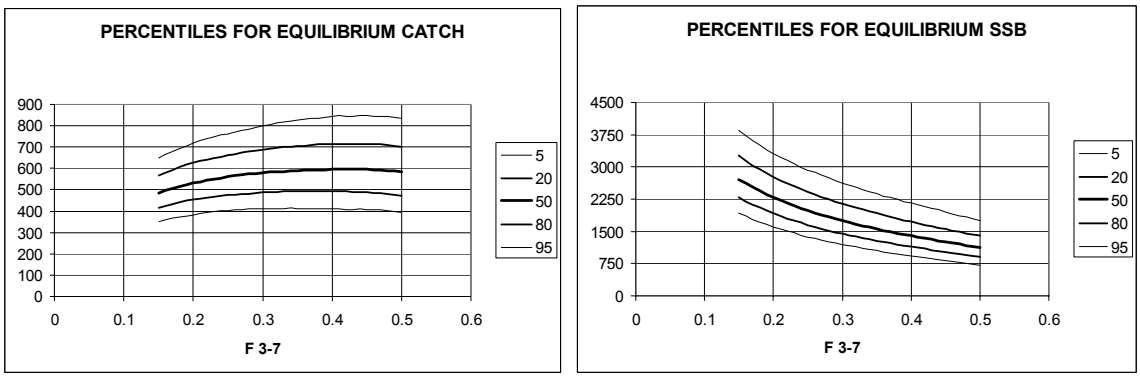


Figure 4.2.3.1 Long-term equilibria calculated using the LTEQ software: percentiles (5, 20, 50, 80 and 95) of catch and SSB as a function of fishing mortality (F_{3-7}).

4.3 Software development of stock assessment tools

At the last meeting of WGMG (ICES 2002a), the Group endorsed the further development of TSA (Time-series Analysis) into a usable FORTRAN 90 subroutine by FRS Aberdeen, the likelihood-based development to XSA (Extended Survivors Analysis) by CEFAS Lowestoft and the development of a new methodological tool for medium-term projections (MedAn) by FRS Aberdeen. During the past year the development of these software packages has progressed, but not to the extent that completion, testing and release of the envisaged programs are possible yet.

In this Section, the details of the on-going and future developments of these programs are presented; together with the details of the development of the software tool CADAPT. These are presented merely for completeness and information and do not imply their endorsement by WGMG.

4.3.1 Current and future developments to TSA

The relevant individuals at FRS Aberdeen intend to make a fully documented FORTRAN 90 subroutine available within the next year, that will fit a standard class of TSA models. They envisage that this will address mainly the following points 1 through 9.

1. Combine catch-at-age code with landings and discards-at-age code to provide a single TSA subroutine. This will allow landings and discards-at-age data to be combined with several surveys - at present several surveys are allowed with catches-at-age, but only one survey with landings and discards-at-age.

2. Change the error structure of the observation equations to assume a constant coefficient of variation, as proposed at the last meeting of WGMG 2001 (ICES 2002a). At present (Fryer 2002), the observation equation for the catch is given by

$$C_{observed}(a,y) = C_{true}(a,y) + \varepsilon_{catch}(a,y)$$

where

$$C_{true}(a,y) = \frac{F(a,y)}{Z(a,y)}\left(1-\exp(-Z(a,y))\right)N(a,y)$$

The $\varepsilon_{catch}(a,y)$ are assumed to be NID (Normal Independent Deviate) with zero mean and standard deviation $\sigma_{catch}B_{catch}(a)q_{catch}(a,y)$ and represent measurement errors in estimating the catch. The $B_{catch}(a)$ are first taken to be unity, but can be adjusted if the measurement errors associated with some ages are larger than for others. The $q_{catch}(a,y)$ are pre-determined by the catch data as described by Gudmundsson (1994). The main reason for this approach in the Gudmundsson's original implementation was to make parameter estimation more robust. This it does do, but only at the expense of using data twice – catch data are used to determine *q*, which then influences the fit of the model to the data. This circularity is hard to justify.

There are several other disadvantages to using the $q_{catch}(a,y)$ as a pre-determined variance component; they are difficult to interpret, they can be imprecisely estimated when e.g. there are missing years of catch data, and they can be susceptible to outliers in the catch data.

A simpler alternative would be to assume that the $\varepsilon_{catch}(a,y)$ are NID with zero mean and standard deviation $\mathrm{cv}_{catch} B_{catch}(a) C_{true}(a,y)$; i.e. to assume that measurement errors are distributed with constant coefficient of variation cv_{catch}. The $B_{catch}(a)$ would still allow the coefficient of variation to vary with age if necessary, and it will still be possible to down-weight individual points to decrease the influence of outliers. Similar changes will be made to the error structure of the observation equations for survey, landings and discards data.

The option to pre-specify the variance of the measurement errors will be retained.

3. Provide a more general way of modelling discard data. This will be useful when a logistic discard curve is not appropriate, either because the logistic curve does not describe the data or because there are too few ages to fit the logistic curve to. The motivation for this comes from Division VIa cod, where only ages 1 and 2 are discarded in significant quantities.

To replace the logistic ogive, it is proposed that the proportions discarded at age $P(a, y)$ will be assumed to evolve in a manner analogous to the evolution of fishing mortalities. Adapting the notation in Gudmundsson (1994) and Fryer (2002) by using a superscript P to denote state variables and variances associated with discard proportions:

$$\text{logit } P(a,y) = U^P(a,y) + V^P(y) + \text{NID}\left(0, \left(H^P(a)\sigma_P\right)^2\right)$$

$$U^P(a,y) = U^P(a,y-1) + \text{NID}\left(0, \sigma^2_{U^P}\right) \quad a \le a_{d2}$$

$$\text{with the constraint that } \sum_1^{a_{d2}} U^P(a,y) = 0$$

$$V^P(y) = Y^P(y) + \text{NID}\left(0, \sigma^2_{V^P}\right)$$

$$Y^P(y) = Y^P(y-1) + \text{NID}\left(0, \sigma^2_{Y^P}\right)$$

- *the logit of the proportion discarded is separated into an age component* $U^P(a,y)$ *and a year component* $V^P(y)$*, both of which can evolve over time,*
- a_{d2} *is the age above which discarding is negligible,*
- *the variance* $\sigma^2_{Y^P}$ *induces persistent changes in the overall level of discarding (through the year component* V^P*),*
- $\sigma^2_{V^P}$ *induces transitory changes in the overall level of discarding (through* V^P*),*
- $\sigma^2_{U^P}$ *induces persistent changes in the pattern of discarding (through the age component* U^P*),*
- σ^2_P *induces transitory changes in discarding around the separable model* $U^P + V^P$*,*
- $H^P(a)$ *allows the variability in discarding to be age dependent, and*
- *the constraint on the* $U^P(a,y)$ *is necessary for identifiability.*

The observation equations for the discards $D(a, y)$ and landings $L(a, y)$ are now

$$D(a,y) = P(a,y)\, C_{true}(a,y) + \varepsilon_{discards}(a,y)$$

$$L(a,y) = \left(1 - P(a,y)\right) C_{true}(a,y) + \varepsilon_{landings}(a,y)$$

where $\varepsilon_{discards}(a,y)$ and $\varepsilon_{landings}(a,y)$ are assumed to be NID with zero mean and standard deviations $\mathrm{cv}_{discards} B_{discards}(a) P(a,y) C_{true}(a,y)$ and $\mathrm{cv}_{landings} B_{landings}(a)\big(1-P(a,y)\big) C_{true}(a,y)$ respectively.

This new approach will increase the estimation load. In practice, it will probably be difficult to estimate σ_P^2 separately from $\mathrm{cv}_{discards}$ and $\mathrm{cv}_{landings}$ (unless the data are unbelievably good) and the effect of transitory changes in age-specific discarding will have to be absorbed into the estimates of measurement variability.

Initial attempts have been made to implement this model for Division VIa cod (Needle & Fryer 2002) but there was limited opportunity to validate either the code or the fitted model.

4. Extend TSA to allow for a fishery with two fleets with catches-at-age or landings and discards-at-age. The motivation for this comes from Subarea IV whiting where it would be desirable to separate the catches attributable to the human consumption and industrial fisheries. This should be achievable by writing $Z(a,y) = F_1(a,y) + F_2(a,y) + M(a,y)$, where $F_1(a,y)$, $F_2(a,y)$, the fishing mortalities of the two fleets, are allowed to evolve according to the usual state equations. In principle, changes in the fishing mortalities of the two fleets might be correlated, but assuming independence between fleets will be a good first step.

5. Tidying-up for general use: error-trapping and documentation.

6. Alter code to produce standard ICES output, such as SEN and SUM files.

7. Provide a facility for retrospective runs.

8. Provide standard errors or profile likelihood regions for the model parameters.

9. Provide a module that will give reasonable initial estimates of the model parameters. This could take the form of a simple cohort analysis or separable model.

10. Provide a Windows front-end for general ease of use, including output plots and a diagnostic tracking facility.

The intended time-scale for these developments is:

- to implement points (1), (2), (3) and (6) for the Working Group on the Assessment of Northern Shelf Demersal Stocks [WGNSDS] which next meets in May 2003, building in obvious error-traps at the same time, and drafting some preliminary documentation;

- to implement point (4) for the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak [WGNSSK] which next meets in September 2003;

- to tidy-up the documentation following the experiences of both WGNSDS and WGNSSK, and distribute for testing to potential users; and

- to address points (7), (8) and (9) and redistribute as time permits.

The implementation of point (10) is desirable but of a lower priority at the moment.

The Kalman filter method on which TSA is based is inherently conservative. Because the filter acts as a time-series smoother, rapid changes in (say) fishing mortality in the stock being assessed may take some years to become apparent in the corresponding TSA assessment. We have noted this in the context of a TSA assessment of cod in the NAFO 3Ps area (Needle 2001). The fishery for that stock was closed for several years during the 1990s, yet the TSA estimates of fishing mortality did not fall to zero for several years. It is possible to impose a rapid change of fishing mortality, with the extent of the change estimated, but we must still specify when the change takes place. In the current ICES situation, potentially stringent management measures may lead to a rapid fall in fishing mortality, in which case we would have to modify TSA accordingly. On the other hand, we may not want to assume such a fall if enforcement of management measures is not rigorous, in which case we would not modify TSA. The way in which TSA is used in the future is therefore very dependent on what we believe to be happening in the fishery. **It would be beneficial to evaluate more**

precisely the response of TSA to rapid fishery changes, using the kind of simulated datasets described elsewhere in this report.

4.3.2 Current and future developments to XSA

The Extended Survivors Analysis (XSA) algorithm used to fit the XSA model developed by Shepherd (1999) is currently undergoing modification at CEFAS, Lowestoft. In addition to the current algorithm specification, the developments currently being programmed and tested are:

1. Independent application of the fishing mortality shrinkage constraint across years and ages. This development has particular relevance in the current fisheries climate and has been incorporated into the XSA algorithm.
2. Fleet based catchability at age models allowing more flexibility in the fitted model structure. This will replace the one-model-fits-all approach currently applied to fleet catchability.
3. Inverse variance weighting by index series. XSA estimates from models fitted to noisy data can be dominated by spurious fits to particular ages. In such model structures weighting by the standardised index series standard error can provide a more robust model fit.
4. The use of index data collected subsequent to the final year of catch data removing the need for the RCRT program which uses the same regression algorithms.
5. Minimisation of the objective function using numerical search algorithms, which allow the estimation of the parameter variance covariance matrix.
6. The use of alternative objective function distribution assumptions; e.g. Quasi-likelihood.
7. Non-parametric bootstrap algorithms for deriving distribution of parameter estimates and confidence intervals of output metrics.
8. Bias correction methods.

The development work that was first described at last year's meeting is nearing the testing phase (using the simulated data sets described elsewhere in this report). Progress will be reported at the next meeting of this Working Group.

4.3.3 StockAn, RecAn and MedAn

The last meeting of the ICES Working Group on Methods of Fish Stock Assessment (ICES 2002a) highlighted the need for replacements for the current medium-term projection software, namely WGMTERM (Working Group Medium-Term) and the Aberdeen Suite. There are three main reasons for this:

1. The Aberdeen Suite is cumbersome, inflexible and difficult to use.

2. The projection software takes no account of stock-based biological processes governing growth and reproduction. For example, weights-at-age and maturity-at-age are assumed fixed, and recruitment is modelled as a function of spawning stock biomass (which may be a poor proxy for reproductive potential).

3. Random variation in recruitment is implemented by bootstrapping model residuals, which ignores any time-series structure in the historical dataset.

A potential replacement for WGMTERM, called MedAn 1.0, was presented to the previous WGMG meeting. This modelled time-series variation in recruitment residuals using autoregressive moving-average (ARMA) techniques, and thus addressed point 3 above. WGMG recommended that this work be continued, and that particular attention be paid to the projection of weights-at-age and maturity-at-age.

Shortly after the 2001 meeting of WGMG, the third and final meeting of the ICES Study Group on the Incorporation of Process Information into Stock-Recruitment Models [SGPRISM] (ICES 2002e) endorsed this recommendation and proposed further that a dedicated Study Group be set up to facilitate this work. Accordingly, the ICES Study Group on Growth, Maturity and Condition in Stock Projections [SGGROMAT] (ICES 2003a) was established and met for the first time in December 2002. The findings of this Group are discussed in more detail in Section 7.1.1 of this Report. In preparation for the meeting, scientists from FRS (Aberdeen, Scotland) and IMR (Bergen, Norway) worked together during 2002 to develop process–based models of growth and reproduction, using Northeast Arctic cod as a data-rich case study. Work on improved medium-term projection software proceeded in parallel.

In the course of this collaboration, it became clear that any process-based projection models would have to be accompanied by models of the same processes in the historical dataset, along with a robust and flexible recruitment

model-fitting component. Therefore, the software implementation is currently being developed in three separate modules, namely StockAn (historical processes), RecAn (recruitment modelling), and MedAn (projections). These are described below in the relevant sub-sections of this Report – Sections 4.3.3.1, 4.3.3.2 and 4.3.3.3, respectively. We should note that the processes addressed in this work are concerned with growth and reproduction only – the gap between spawning and subsequent recruitment has not yet been addressed, nor is it clear at present how this should be done.

4.3.3.1 StockAn

The aim of StockAn (current version 1.0) is to fit models to historic data of proportion female, weights, proportion mature and fecundity. All these models are fitted on the basis of length, rather than age, as it is believed that growth and reproductive characteristics (in Northeast Arctic cod, at least) are more dependent on length than age. The final version of the software will have to include age-based process models as well, to allow for cases where length data are missing are age data are more reliable. Models of growth have not yet been implemented in the current version, although these will be required for projections, and possible methods of doing this are described below (Section 4.3.3.4).

The StockAn input data specific for the Northeast Arctic cod example are summarised in Table 4.3.3.1. The program will still run if some of these data are missing (in particular, fecundity estimates might be difficult to obtain for many stocks). In these cases, the models that require the missing data are simply switched off. The methods by which the data are collated will vary from stock to stock, and so are not explored further here. Further details on the case study will be reported to the next meetings of both SGGROMAT in 2003 and WGMG in 2004.

The modelling carried out within StockAn can be summarised as follows:

1. *Generalized linear models (GLMs: McCullagh & Nelder 1983) with binomial errors and logit-link functions are fitted to each year of proportion female at length data* $P_{l,y}^{L}$. *In the Northeast Arctic cod case the proportion is assumed to be 0.5 for all lengths less than 85 cm, and a fixed model is assumed for the years 1980 onwards, but these aspects are case-specific and at the control of the user.*

2. *Power models are fitted to each year of survey-derived weight-length data* $W_{l,y}^{L,S}$, *using GLMs with gamma errors and log-link functions.*

3. *GLMs with binomial errors and logit-link functions are fitted to each year of proportion mature at length data* $Mat_{l,y}^{L}$.

4. *For each series of models fitted in steps 1–3, a number of ARMA time-series models are fitted to parameter estimates by maximum likelihood. The AR and MA orders of these time-series models range from 0 to 3. They are compared using Akaike's Information Criterion (AIC: Akaike 1973) goodness-of-fit statistic, and the best-fitting time-series model will be used subsequently to generate stochastic future realisations of each process model.*

5. *A power model using all available years together is fitted to weight-length data* $W_{l,y}^{L,F}$ *from the fecundity dataset, using a GLM with gamma errors and a log-link function. This serves as a global weight-length relationship. Relative condition* $C_{l,y}$ *is derived using* $C_{l,y} = \ln\left(\hat{W}_{l,y}^{L,S} \middle/ \hat{W}_{l,y}^{L,F}\right)$. *That is, relative condition at a given length in a given year is the lognormal ratio of the fitted weight at that length and year (from surveys) to the fitted weight at that length for all years (from a fecundity dataset). Thus relative condition measures how the weight-at-length in a given year differs from a global weight-at-length relationship. In "bad" years* $C_{l,y} < 0.0$, *while in good years* $C_{l,y} > 0.0$.

6. *A GLM with gamma errors and a log link-function is fitted to fecundity data, using ln(length) and relative condition* $C_{l,y}$ *as independent variables.*

7. *Finally, StockAn calculates a series of summary indices, writes files with recruitment and reproductive potential for subsequent input into RecAn, and produces a further transfer file for use in MedAn. The indices used are SSB*

derived from age-based data, SSB derived from length-based data, female-only SSB derived from length-based data, harvestable biomass (total weight of all fish over the minimum landing size), and total egg production (TEP).

The estimation of fecundity for Northeast Arctic cod is described in more detail in the report of the 2002 meeting of SGGROMAT (ICES 2003a). StockAn has a Windows user interface that automatically produces standard output plots for stock assessment Working Group use.

4.3.3.2 RecAn

RecAn (current version 2.20) has been used in ICES assessment Working Groups for some years, having replaced the RECRUIT program as part of the Aberdeen Suite (although RECRUIT is still widely used as well). In essence it is a relatively simple recruitment model-fitting program, with the same user interface as StockAn and able to read in stock-recruitment data from a number of sources (including StockAn). While it is an easy enough task to fit a linearised Ricker model (say), it is more difficult to create code robust enough to fit a Shepherd model (say) using non-linear least-squares to over 500 test-case datasets. Ensuring generality has been the major focus of work on this module.

The version of RecAn which is currently available can fit a wide variety of different recruitment models, assuming different error distributions and fitting methodologies. Models may be fitted assuming residual autoregression, or residual time-series can be characterised using the same ARMA approach that was outlined above for StockAn. Model estimation is repeated for truncated datasets, to test for the robustness of parameter estimates to new data. Finally, a series of supplementary calculations are included, including various management reference points. The principal output from RecAn is a file of reproductive potential and recruitment, identical to that produced by StockAn but augmented by information on the recruitment model and time-series model fits.

4.3.3.3 MedAn

The development of MedAn (current version 2.0) is still in the early stages. As yet, the program does little more than replicate WGMTERMC with the same user interface as used for StockAn and RecAn. In the future, it is envisaged that ARMA time-series methods will be used to generate stochasticity in projections for the variables fitted in StockAn (see above), either with or without reference to biotic or abiotic driving variables.

4.3.3.4 Growth modelling

The new medium-term projection methods described above do not yet include models of growth. For the Northeast Arctic cod case study, it would seem reasonable to assume that growth in length in a cohort-based process, while growth in weight is a year-based process. That is, a fish of a certain length can only get longer in the next year, and it is the extent of lengthening that must be modelled; while a fish of a certain weight can either get lighter (skinnier) or heavier (fatter) in the next year, depending on feeding conditions. Initial work on a suitable growth model has begun, with the following algorithm proposed as one possibility:

1. Collate all the age-length data $N_{a,l}$ for a given cohort.

2. Fit a robust version of the von Bertalanffy growth curve to these data, using nonlinear least-squares estimation (Stratoudakis *et al*. 1997). There is no need to weight the model fit by the inverse-variance of the observations at each age, since raw observations are used. The fitted model is:

$$\hat{L}(a; L_1, L_2, k) = L_1 + \frac{(L_2 - L_1)\left(1 - \mathrm{e}^{-k(a-a_1)}\right)}{1 - \mathrm{e}^{-\mathrm{k}(\mathrm{a}_2 - a_1)}}$$

   where a_1 and a_2 are reference ages (set to 4 and 12 respectively for NEA cod), L_1 and L_2 are the estimated lengths at the reference ages, and *k* is the estimated growth rate.

3. Bootstrap the fitted model to generate simulation envelopes.
4. Fit a linear model through the age-length data. If it lies within the simulation envelope, then there is no significant difference between the growth curve and the linear model. If this applies to all cohorts, the projection of growth would be simplified.

The results of applying this algorithm to the 1943–1946 cohorts of Northeast Arctic cod are summarised in Figure 4.3.3.1. There are two main features in these plots: the fitted von Bertalanffy growth curves are very close to being linear, and the simulation envelopes are extremely tight (due to the size of the datasets). This means that the test for linearity mentioned in point 4 above is probably not appropriate, and work on this model is continuing.

4.3.3.5 Software design

A single implementation of process-based projection models is unlikely to be sufficient. Assessment Working Group members need programs that are robust, consistent and straightforward to use. There is little time within the confines of a Working Group for experimentation with process-model formulation, and perhaps more importantly, ACFM must know the precise basis and methodology behind each projection which they are asked to consider, and be able to replicate the results. These points suggest centrally-managed packages, with tight version control and regular, well-tested and verified updates. Conversely, process modellers trying to evaluate the effect of different models and methodologies in a management framework require programmes that are flexible, open-source and easily modified.

The requirements of the two groups likely to make use of the software are therefore very different, and both should be addressed by developing parallel implementations. The first, the StockAn/RecAn/MedAn collection described above, is being written in FORTRAN, and consists of centrally-managed code suitable for Working Group use. This software will be modularised as far as possible, in order to simplify expansion and modification. The second will be available as a script for the R language (http://www.r-project.org/). R is a freeware version of the S language, which is also used as the basis of S-PLUS. R has similar functionality and syntax to S-PLUS (so that the code is fairly straightforward to modify for those experienced in the latter), and provides a suitable programming environment for process modellers seeking to test hypotheses and develop methodologies.

4.3.4 Current and future developments to CADAPT

CADAPT was presented to WGMG and showed some promise but might be devoloped further for the following reasons: CADAPT has possibilities to estimate uncertainty in the results by bootstrapping, it is convenient to do retrospective runs (Eero BB1), a single input file makes running the program relatively straigthforward, ADMB standard deviation reports and profile likelihood give alternative descriptions of uncertainty in stock parameters and estimates, further development could make use of a growing literature of routines useful for fisheries models, and both CADAPT and R are freeware thus making it relatively easy to build further upon existing methods.

A list of items for future CADAPT work includes:

1. Finish a specification of a model that could serve as the stable version.
2. Develop variants suited to the particular needs of stocks under assessment such as tuning with a mixture of age group indices and lumped indices for SSB, numbers above a certain age. These models might benefit from utilities of the type used in CADAPT of coupling the model with an ‘R’ or ‘S’ environment.
3. Add bootstrap bias corrections, study the difference between bootstrapping raw and residuals standardized/homogenized with the inverse variance case weights currently implemented.
4. Start projecting with ‘cadapt’ including only bootstrapped estimates of uncertainty. A division between the optimizations for estimating log-qjus and log-survivors and those needed in projections into the year of assessment, and further a head is suggested.Options for choosing beetween yield, harvest control rule or F-constraint might be included. In developing such projections in CADAPT, the simplifying assumption of Ockham recruitment might be appropriate.
5. Adding robust log-normal optimization for proportions, as an optional component to the objective function.
6. Parameterize survey selectivity (such as logistic, double-normal) as an option.
7. Deal with migrations (compare with ICCATT VPA2box).
8. Input/output conversions: ‘cadapt2lowestoft’ and ‘lowestoft2cadapt’ routines that could convert inputs with a single command should be developed (single file input for CADAPT, ca. 10 files of input in lowestoft format). Read/write routines for standard format output should also be considered.
9. Develop standard diagnostics and summary plot routines, evaluate whether ‘methods’ and ‘classes’ in ‘R or S’ should be developed. Dynamic links to C++ programs (result of ADMB templates) to the environment of choice should also be considered.

4.4 General guidelines for exploring and comparing assessment methods

A general overview of the methods presented to WGMG 2003 was presented in Section 4.2. The main characteristics of the methods is given in Table 4.4.1.

Each assessment method relies on a set of assumptions about the properties of the stock and of the relation between stock and observations. The choice of method should preferentially be guided by the validity of the assumptions. This may not always be the case – often the choice of method is more due to tradition and what is available, and a justification for the choice of method is not often seen. For example, in ICES, it is hard to find a rationale for the almost universal use of XSA in some working groups, and of ICA in others.

Also, when different methods give conflicting results, it is either because they rely on different assumptions, some of which may be violated, or because of differences in the way they emphasise by conflicting data and noise. The data will influence the final outcome to a variable extent, and it may be very hard to identify intuitively the data that drive different methods in different directions.

Influence diagnostics (Sections 5.4 and 5.5) may be a helpful tool to identify data that have a strong influence on the final outcome. Furthermore, some understanding of the problem can be achieved by simple analyses of the data, to reveal deviations from important assumptions and by identifying crucial parameters and explore how they are determined.

It is useful to have an overview of the most important assumptions when exploring methods. Table 4.4.1 gives this for several methods that have been used in the ICES community to a greater or lesser extent. In brief, these include:

Assumptions about selection-at-age: This can be the assumption that the selection is stable for some period. It also includes the relation between the mortality at the oldest age relative to younger ages, which most often is fixed either of the program or by the user.

How the plus group is treated: This may be important in cases where considerable amounts of fish reach that stage, but may also conceal unrealistic large or small numbers surviving the oldest true age.

Assumptions about catchability in surveys and CPUE: Trends in catchability is a clear candidate as cause of retrospective bias, because it gives the impression that there are trends in the population abundance that may not be real.

Assumptions about the distribution of noise: As the model sees it, any deviation of the data from what the model would imply is treated as noise. The final estimate is one that minimises some measure of this noise, e.g. a likelihood measure. Partly, this will provide a compromise between noisy data, but it will also attempt to adjust the model to account for deviations from the model within the constraints given by the assumptions. Accordingly, the optimisation will both try to compensate for deviations of the data from the model assumptions, and adjust to the noise in the data.

Assumptions about natural mortality: The experience is that natural mortality can in general not be estimated reliably together with the fishing mortality from catch and survey data because the data can be explained almost equally well be a range of each of the mortalities, provided the other is adjusted accordingly.

Assumptions about catch data: It is generally assumed that reported landings are representative of actual catches. There are exceptions, such as the TSA assessments of haddock and whiting in Division VIa (ICES 2002c), but in general the effects of discarding and misreporting are not explicitly considered. Levels of these are thought to have been high for some time, and are likely to become higher with collapsing stocks and stringent management measures. Potential alternatives to catch-based assessment methods are discussed in Section 6, and Appendices A and B.

Detecting mortality signals in the data

The decline in abundance over time within a cohort is generally assumed to be due to mortality. Simple analyses of trends in the data can give valuable insight in the mortality. The exploratory analyses outlined here can easily be made in a spreadsheet.

Plots of catches-at-age by year class: These plots allow for following the decline of the year classes. Changes in selection can be seen directly as the relation taken at young age vs. old age (e.g. Figure 4.4.1). Doing this on a log-scale gives an indication of the shape of the selection patter that is to be expected – a dome shape indicates that mortality rises towards old age (e.g. Figure 4.4.2).

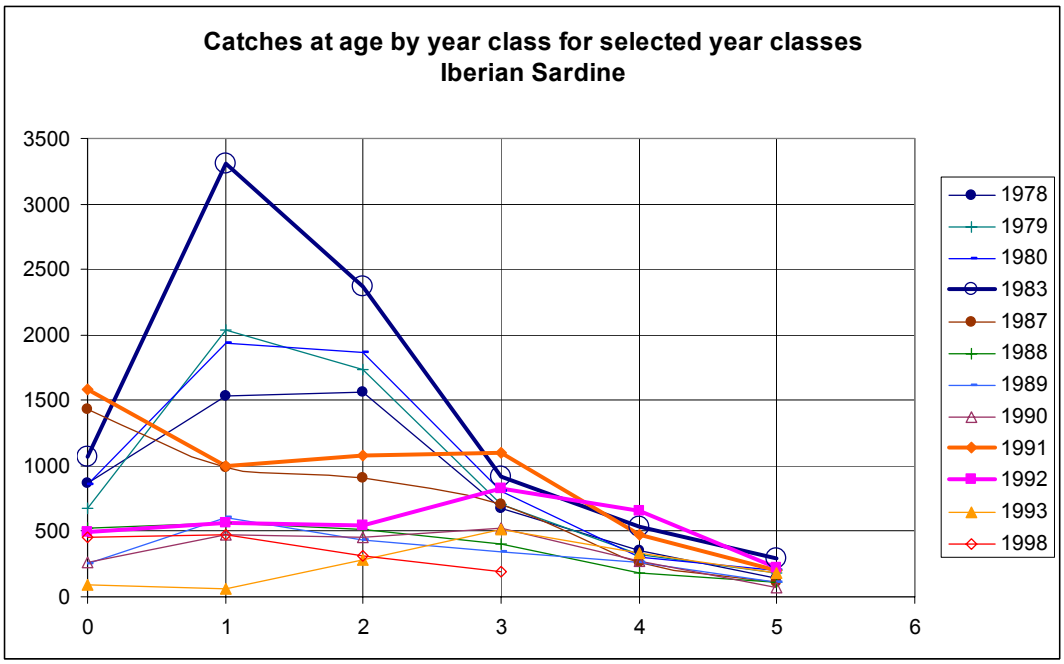


Figure 4.4.1 Plot of catches by year class to show how different year classes have been exploited differently. Some large year classes are highlighted, and some year classes have been omitted for clarity. In this case, there is a clear difference in exploitation pattern between early and recent year classes.

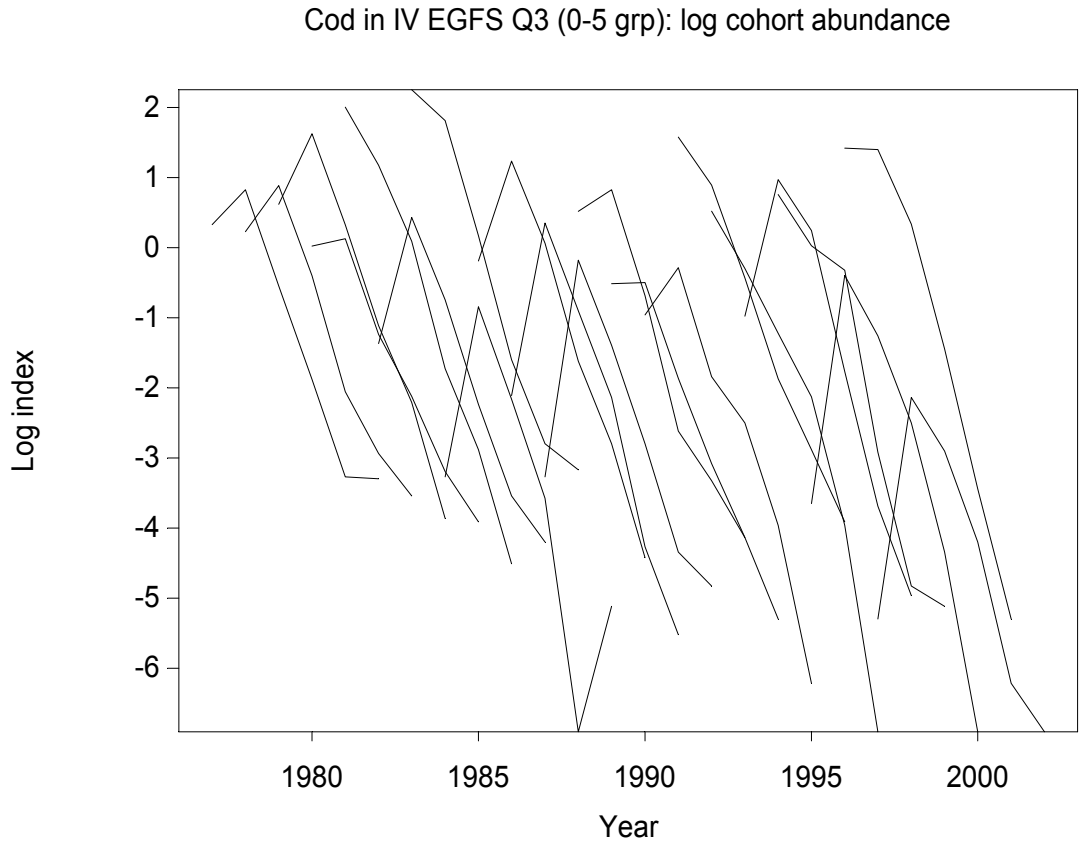


Figure 4.4.2 Log survey indices by year. Each line follows a separate year class through its life span, as measured by the survey. Perfect catchability would result in straight lines, so departures from linearity highlight ages and years for which survey catchability is not 1.0.

Plots of log catch ratios: *LCR = log[C(a,y)/C(a+1,y+1)]*. The LCR is to a fair approximation the mean total mortality in the cohort in the years y and y+1, adjusted for the change in fishing mortality:

$$LCR \sim [(Z(a,y) + Z(a+1,y+1)]/2 + log[F(a,y)/F(a+1,y+1)]$$

The curves may be quite noisy, and it may be useful to apply a moving average to get the patterns clearer (e.g. Figure 4.4.3). If the change in F by age can be neglected (i.e. fully recruited ages), LCR is due only to the level of the total mortality and the change from one year to the next. If the fishing mortality is relatively stable, this gives an indication of the total mortality. Furthermore, if the time courses of LCR deviate for different ages, this is a strong indication that the selection-at-age is not stable.

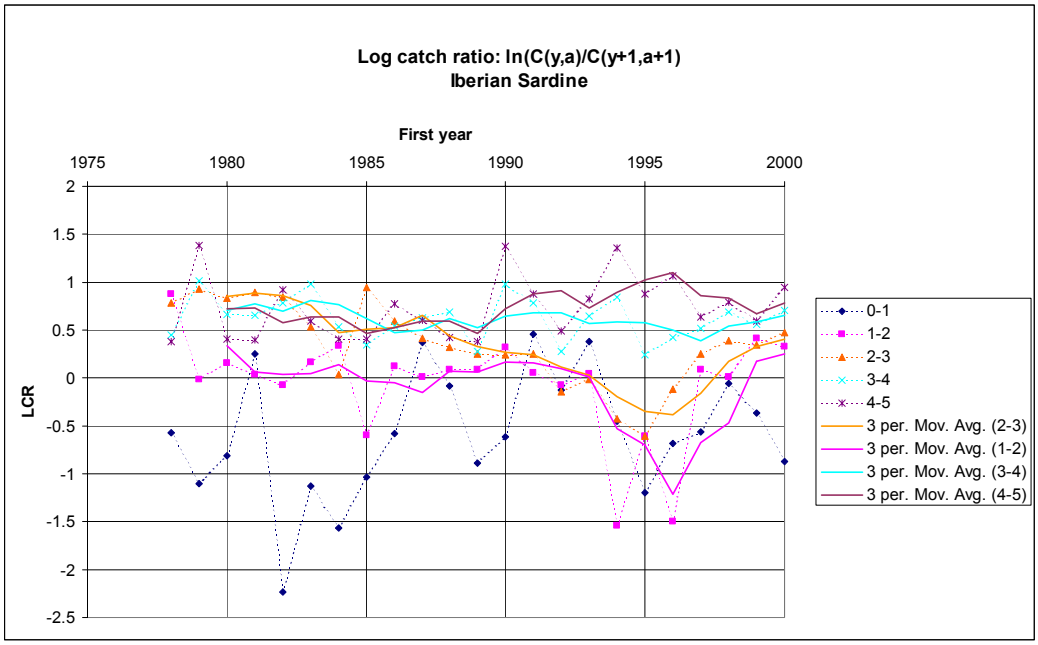


Figure 4.4.3 Plot of log catch ratios (raw values and 3-point moving averages) for Iberian sardine. The time course of the various deviate, indicating a variable selection.

Plots of log index ratios for survey indices: *LIR = log[I(a,y)/I(a+1,y+1)]*. Analogous to the LCR defined above, LIR is a measure of the total mortality in the cohort between the surveys, adjusted for the change in catchability as log[q(a,y)/q(a+1,y+1)]. The time trajectories for various age should be parallel, if catchabilities at age are constant over time. However, trends in catchabilities over time affecting all ages on equal terms will not be revealed. Furthermore, if the catchability is similar at neighbouring ages, the level of the LIR indicates the total mortality. If the signal in the LCR and in various LIR is conflicting, there must be a problem with either catchabilities or the mortality signal in the catches. In practise, however, this measure may often best serve to illustrate the signal to noise ratio in the mortality signal from the surveys, and this ratio can be quite bad (e.g. Figure 4.4.4).

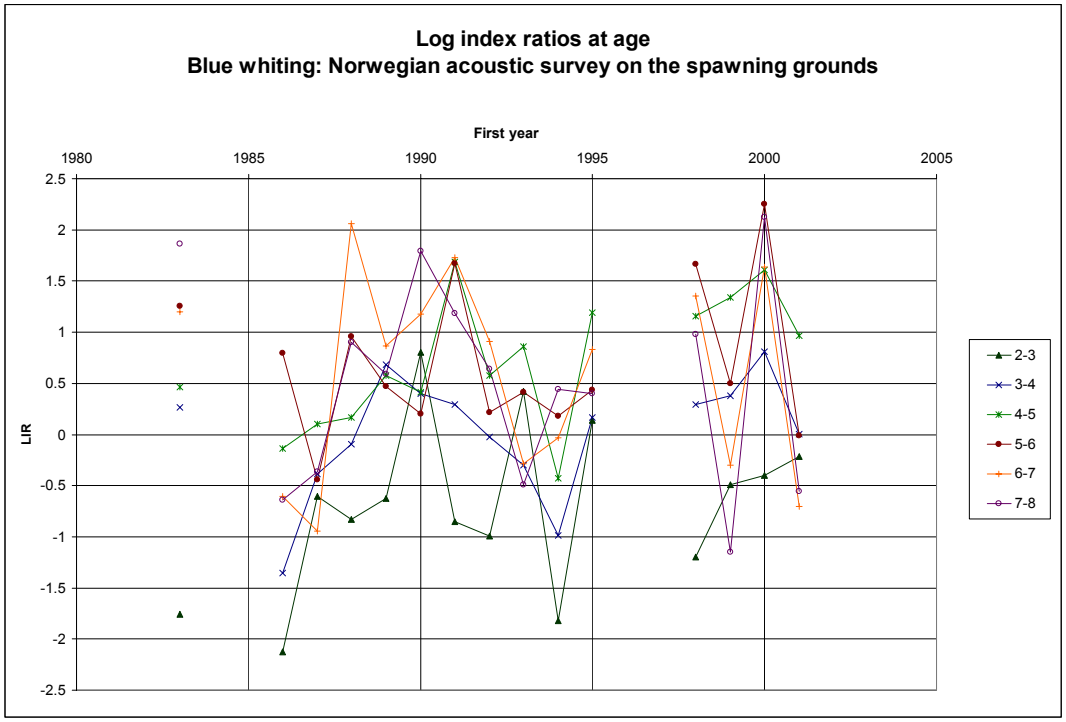


Figure 4.4.4 Log index ratios for an acoustic survey for Blue whiting. Despite considerable noise, there seems to be an increasing trend for most ages, suggesting an increasing trend in total mortality. The data are too noisy to make inferences about trends in catchabilities at age. The large variations in the last year indicates strong year to year variations in the catchability.

Influence of data on parameter estimates

Terminal Fs: In a separable VPA (Pope and Shepherd 1982), there is one solution for the matrices of fishing mortalities and of stock numbers to each set of terminal F and terminal selection. Thus, in this framework, the catches define the stock as a two-parameter family of solutions. In a separable model, this is almost true, although some combinations give a slightly better fit than others. A VPA basically needs one parameter for each cohort. However, by simple constraints like assuming the same selection in the last two years, this also can be reduced to the two parameters terminal F and terminal S.

When the population is fitted to other data as well, the population may deviate from the family generated by the catches alone. However, unless there are very strong signals in the supplementary data, their influence is largely restricted to choosing a best terminal F and terminal S. Then, the rest of the matrices are largely given by the catches. In most models, the terminal S is fixed, in which case the terminal F is the only remaining free parameter. This may be a useful guideline when tracing the influence of various data, because one can get quite far by studying the effect of varying the terminal F and possibly the terminal S on the residuals.

The effect of changing model parameters: The combined effect is complex, because it propagates through the population matrix (e.g. Figure 4.4.5). Adjusting a model parameter will improve the fit in some places and make it worse in others. A simple rule of thumb is that the model will prefer to find the adjustment that is least expensive:

- When fitting the model to the early years the data, the model will adapt to deviations at young age by changing the yearly F, and adapt to deviations at old age by changing N at those ages.

- In the later years, the model will adapt to deviations at young age by changing N at those ages and adapt to deviations at old age by changing yearly F.

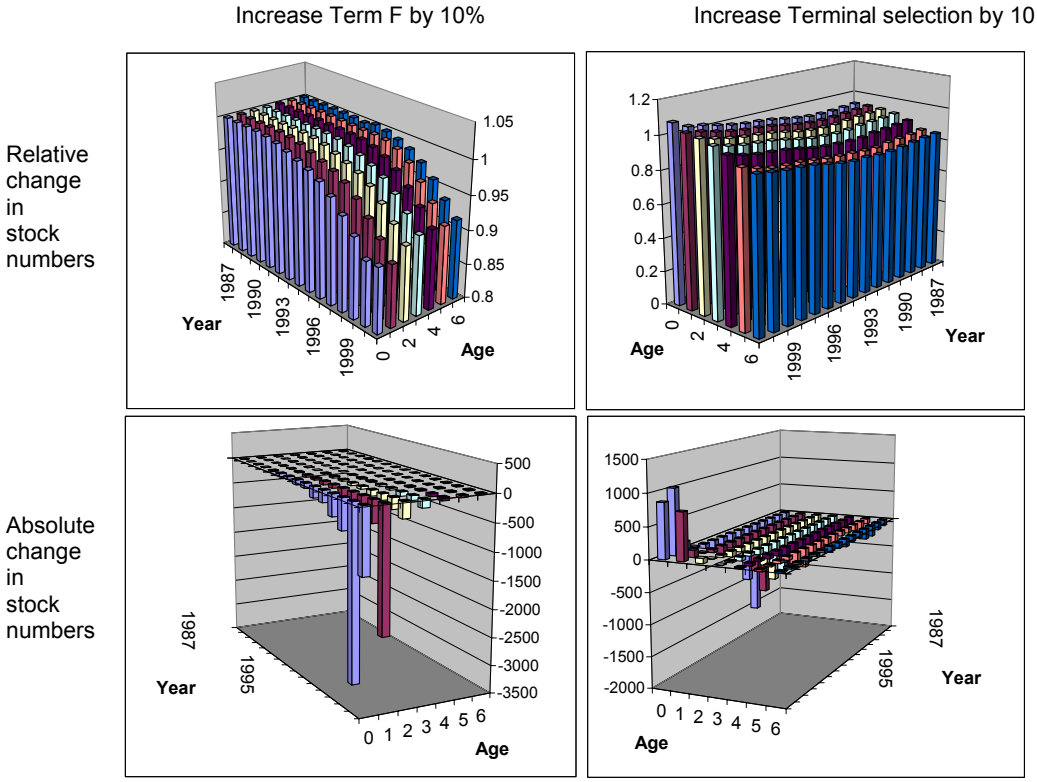


Figure 4.4.5 Change in stock numbers in a separable model fitted to catch data. The figures shows the effect constraining the terminal F and selection respectively, to a 10% increase compared to the optimal fit. The data are for Iberian Sardine.

The change in residuals with changing terminal F: Since the terminal F in many cases will be the remaining free parameter, studying how the squared residuals (i.e. the terms in the objective function) change with small changes in terminal F may give valuable insight in which data are driving the model to the final solution. Such a change will increase some residuals and decrease others, and one may ask the simple question who pay and who gain from this change. In some models this can be done fairly easily. In others, like ICA and XSA, this is not possible. However, the essentials of both models can be captured in relatively simple spreadsheet models.

The analysis of the blue whiting (see Section 8) can be taken as a worked example of the application of the procedures outlined above. For ease of reference, the extensive selection of graphical outputs and diagnostics have been collated at the end of Section 8 for ease of reference.

Table 4.3.3.1 StockAn input data for the Northeast Arctic cod case study.

Variable	*Notation*	*Source*
Numbers-at-age	$N^{A}_{a,y}$	ICES assessment
Numbers-at-length and age	$N^{L}_{a,l,y}$	Averaged proportions from Russian and Norwegian surveys applied to $N^{A}_{a,y}$
Prop. mature at age	$Mat^{A}_{a,y}$	ICES assessment
Prop. mature at length	$Mat^{L}_{l,y}$	Derived using $Mat^{A}_{a,y}$ and $N^{L}_{a,l,y}$
Prop. female at length	$P^{L}_{l,y}$	Commercial and survey data
Mean weight-at-age	$W^{A}_{a,y}$	ICES assessment supplanted with survey information
Mean weight-at-length in surveys	$W^{L,S}_{l,y}$	Survey data
Mean weight-at-length in fecundity dataset	$W^{L,F}_{l,y}$	Fecundity dataset from dedicated surveys
Fecundity at length	$Fec^{L}_{l,y}$	Fecundity dataset from dedicated surveys

Table 4.4.1 Comparison of main properties of assessment models used in Section 8.

	ICA	AMCI	ISVPA	CADAPT
Model type	Separable model applied to one or two periods with backward VPA run for period before separability.	Separable model applied to the whole assessment period. Selection is allowed to change slowly according user specified "gain factor" .	Separable model applied to one or two periods covering the whole assessment period	VPA using Pope's approximation, modified to take account of the timing of fishery.
Selection	Selection at oldest age set relative to the selection at reference age by the user.	Selection at one age can be specified as the average over some other ages.	Selection at oldest age is equal to that of previous age; selections are normalized to sum to 1.	Not applicable
Allowed calibration data	age structured indices, SSB indices	age structured indices, SSB indices, tagging data	age structured indices, SSB indices	One age structured index
Catchabilities	Proportional, absolute or power model for catchability.	Catchabilities are modelled as separable, but the age factor can be allowed to vary slowly according user specified "gain factor". Modelled catchability at age possible.	Catchabilities can be estimated or assumed equal to 1.	Proportional or power model for catchability models. Catchability plateau possible.
Plus group	Plus group is derived from the catch assuming the same mortality as for the oldest true age.	Plus group is modelled as a dynamic pool. The fishing mortality assumed for the plus age can be estimated, or linked to some younger age.	Plus group is derived from the catch assuming the same mortality as for the oldest true age.	Not modelled
Objective function	Weighted sum of squared log residuals of catches and survey indices. Weighting is decided by the user or with iterative reweighting.	Different objective functions available. Weighting is decided by the user. AMCI does some implicit weighting internally.	Different objective functions available for different terms (i.e. catch residuals, age-indices and SSB indices)	Case weights of the elements of the objective function are calculated from the model fit using inverse variance weighting.
Variance estimates/ uncertainty	Variances of the parameter estimates are derived from the Hessian matrix, which is obtained as a by-product of the optimisation routine.	Variances of the parameter estimates are derived from the Hessian, which is computed directly. Options for parametric or non-parametric bootstrapping.	Parametric conditional bootstrap with respect to catch-at-age data and log-normal random noise on indices.	Non-parametric bootstrapping from survey residuals using utility functions written in R.
Other issues		Allows inclusion of multiple fleets and multiple areas.	Three model versions: * strictly separable model ("effort-controlled ") * VPA-type model ("catch-controlled") * "mixed version". Four different constraints on catch residuals available.	Still in development

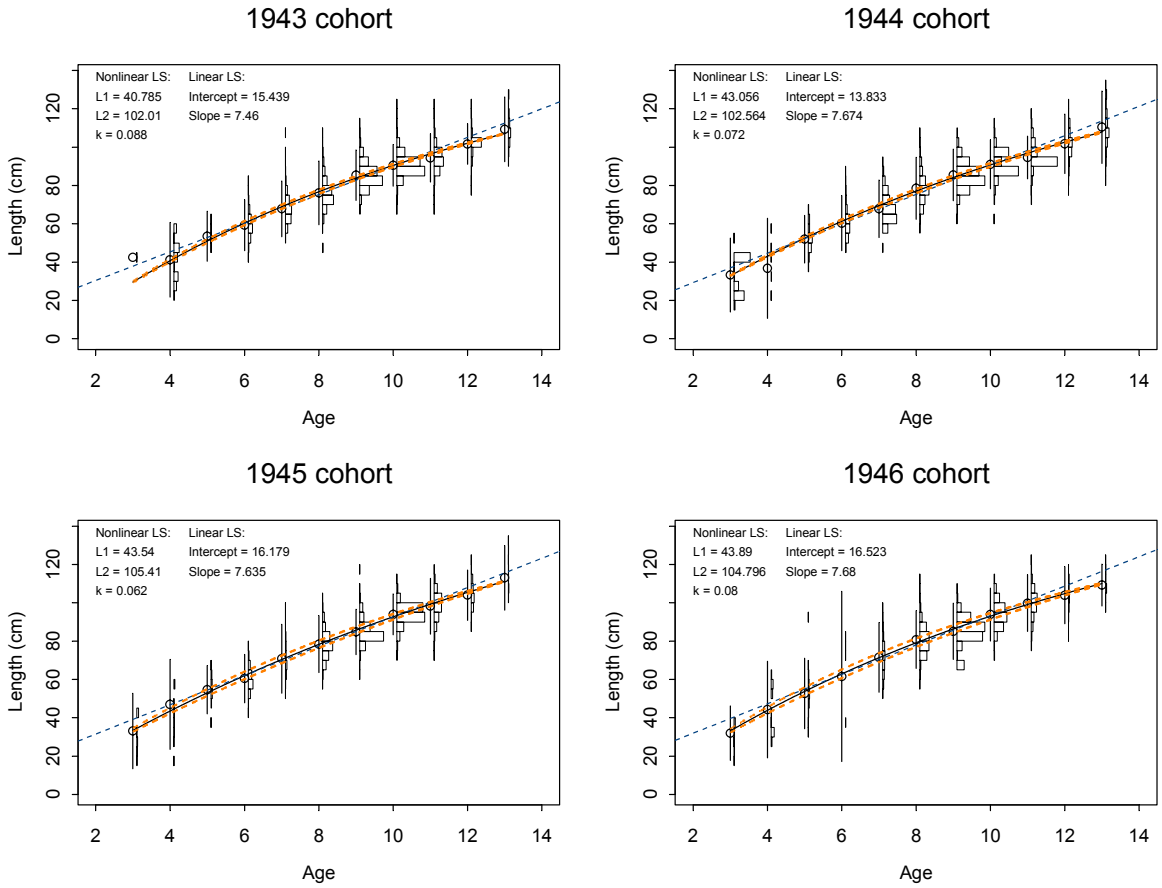


Figure 4.3.3.1 Growth models fitted to Northeast Arctic cod age-length count data. Solid line: robust von Bertalanffy growth curve, with simulation envelope from bootstrapped analyses (this envelope is difficult to distinguish as the envelope is very narrow). Dotted line: linear model fit. Open circles: means of distributions at age, with ±2 standard errors. Histograms of count distributions are also plotted.

5 INFLUENCE DIAGNOSTICS FOR DETECTING DEVIATIONS FROM MODEL ASSUMPTIONS

5.1 Introduction

The working group (ICES 2002a) was interested in investigating further whether local influence diagnostics (LID's) could be utilised to detect mis-specifications in the data inputs or assumptions to SPA. This Section 5 addresses the two ToRs a) and b) (Section 1.2).

The members of WGMG were interested in determining the performance of LID's on simulated data sets from known populations. Before interpreting LID's in case studies of real fish stocks, it is important to understand their performance when applied to known populations. This is the main focus of this section. The Group did not specifically address diagnostics for the cause of retrospective patterns; however, a case study involving the Eastern Scotian Shelf cod stock was presented and discussed that considered this.

The SPA LID's have been developed based on some ad hoc computer programs (referred to as QLSPA) developed by a participant. It is first useful to briefly describe QLSPA.

5.2 Quasi-likelihood sequential population analysis (QLSPA)

The SPA in QLSPA is very similar to the SPA in ADAPT. Populations are reconstructed using Pope's cohort analysis from survivors and numbers at the terminal age. The QL in QLSPA stands for Quasi-likelihood, which is a method for estimation and inference that is widely used, and is utilized in popular statistical packages like GLIM. The Quasi-likelihood is not based on a specific distributional assumption for the SPA errors, and hence it offers some distributional robustness. A wide variety of model error structure can be accommodated by QLSPA. Additional details are noted below.

QLSPA is written in SAS/IML, and is used annually in the assessment of NAFO subdivision 3Ps cod, off the south coast of Newfoundland, Canada.

Model	QLSPA
Version	WGMG 2003
Model type	Similar to ADAPT, but based on Pope's approximation
Selection	The fishery selection at one age can be specified as the average over some other ages, and this specification can include multipliers that are common for groups of years. These multipliers can be fixed or estimated, and the estimation can be penalized (similar to shrinkage in XSA). Only one constraint per cohort is recommended. The constraints are normally applied to approximate the historic numbers at the oldest age in the SPA.
Estimated parameters	Survivors, catchabilities, fishery selection parameters, variance parameters, year effects. Parameters may be freely estimated, or constrained using penalty functions, boundary constraints, or more general nonlinear constraints.
Catchabilities	Proportionality between index and stock abundance is always assumed. The proportionality can be fixed to any value.
Plus group	None
Objective function	Weighted quasi-likelihood methods are available, as well as the weighted sum of squared log residuals. The user can decide weighting, or the inverse-variance principal (i.e. self-weighting) can be used.
Variance estimates/ uncerta	Standard errors are based on an approximate information matrix (hessian of the quasi-likelihood).
Other issues	This software is NOT available for general use. It is intended only to be a development platform.
References	Cadigan, N.G. 1998. Semi-parametric inferences about fish stock size using sequential population analysis (SPA) and quasi-likelihood theory. DFO ATL. FISH. RES. DOC. 98/25

5.3 Summary of local influence diagnostics

The local influence approach to perturbation analyses involves studying the effects of small perturbations to model components using basic concepts in differential geometry. The two main considerations in constructing local influence

are what model components to perturb, and how to measure the effect of the perturbation; that is, what part of the data or model components are assessed for influence, and how is influence measured.

The geometry of the surface of the influence measure versus the perturbations is examined to assess influence. The main diagnostics are simply the slopes of the influence surface. The approach is more fully described in Cadigan and Farrell (2002) and Chapter 5 in ICES (2002a). The local influence approach may be computationally more convenient for perturbation analyses because it does not require re-estimation of the SPA over the very large number of axes in the perturbation space. This is possible because often the perturbation surface of the influence measure around a relevant neighbourhood of the origin is reasonably linear. An important diagnostic is the direction that results in the greatest change in the influence measure. Perturbations in this direction can be examined to assess their plausibility. This is demonstrated in the next section, and also in Appendix C.

5.4 Local influence diagnostic analysis of simulated data sets

5.4.1 Base case fits of the assessment models to exact data

QLSPA and XSA were fitted to the catch-at-age and index series data generated without noise in order to investigate the ability of the models to find the exact solution. Both models reproduced the population and fishery dynamics accurately. Figure 5.4.1.1 presents a comparison of the QLSPA estimates and the simulated population values (see Section 3). The QLSPA estimates are very close to the population values, with small differences resulting from the use of Pope's approximation.

5.4.2 Local influence diagnostics on exact simulation data with model mis-specification

In order to examine the power of LID's to detect model mis-specification a simulated data set was generated with a 4% trend in catchability starting in year 10 (see Section 3); the data were generated without noise.

QLSPA and XSA were fitted to the data and the catchability residuals examined for trends. The diagnostics from the fitted models also had catchability residuals that were close to zero; a disturbing result given the magnitude of the trend in catchability. However, the finding is not unexpected given the low fishing mortality rates during the most recent years of the simulated data set dynamics. The most recent years of the SPA assessment structure are poorly converged at low F ($F <= M/2$) and the assumption of constant catchability throughout the time period over which the model is fitted imposes a constraint that controls (conditions) the fit of the model in the final years where F is low. The estimates of stock and fishery dynamics are biased (Figure 5.4.2.1). This figure demonstrates that the perception of stock size in numbers and numbers of recruits (2 year olds) diverge from the true state following the introduction of the trend in catchability in 1910. The fit of the model to the biased index series without the occurrence of diagnostic patterns in the residuals, highlights the requirement for fishery independent surveys with known / controlled dynamics. The introduction of an unbiased index series allowing contrast of the fit of the model to series with and without the trend in catchability, demonstrated that the bias could be isolated at the low levels of fishing mortality. This has been shown previously by this group (ICES 2002a) and demonstrates the importance for cross-validation in having at least one reliable fishery independent survey.

As with the exact data set the LID's are not informative about model mis-specification when the residual variation is zero or near zero. Therefore, in order to generate residual patterns from this relatively simple one-index assessment for an examination of the utility of influence diagnostics, fishing mortality was increased to induce convergence of the SPA. The time-series of F values were (17 x 0.283, 0.295, 0.31, 0.326, 0.345, 0.367, 0.392, 0.421, 0.454, 0.492, 7 x 0.536, 0.486, 0.443, 0.406, 0.4, 0.4, 0.4, 0.4, 0.4). In addition, the trend in catchability was increased to 7% to introduce patterns to the time-series of retrospective assessments (Section 3).

Fits of the XSA and QLSPA models to the high F simulated data series resulted in residual patterns that indicated departure from the assumption of constant catchability in time. The residual patterns (Figure 5.4.2.2) indicate an increase in catchability after year 10, variable by age, but do not mimic the induced trend completely. The model uses the average catchability of the time-series to estimate the terminal populations, hence the decrease in the residuals in the most recent years of the time-series. However, in practise other possible causes for the residual patterns would also exist, such as errors in catches or assumptions about M.

LID's for the QLSPA fit to the high F data set were examined to see if the method could determine the model mis-specification. The LID's in Cadigan and Farrell (2002) were not designed to detect model mis-specification; rather, they were designed to detect whether important SPA outputs like SSB were sensitive to SPA data inputs or assumptions (i.e. a sensitivity analysis). Nonetheless, the deviance influence measure considered by Cadigan & Farrell (2002) is a measure of model goodness-of-fit and may have some utility to detect model mis-specification. Note that the deviance

is part of the fit function that is minimised in QLSPA to estimate parameters. It is a total measure of residual variation and not residual trends; however, the deviance was the only influence diagnostic that was available to the working group to investigate detecting mis-specifications. A useful direction for future research is to investigate whether other diagnostics exist that could be incorporated into the local influence approach to detect mis-specifications.

Four sources of mis-specification (i.e. the perturbation schemes) were investigated. They are case weight's (CW, the weight given to individual residuals in the fit function), survey catchabilities (Q), catches (C), and natural mortalities (M). If the SPA fit function that is minimized can be written as

$$\Lambda = \sum_i \lambda_i$$

where i indexes the observations then the CW perturbation is

$$\Lambda_w = \sum_i \lambda_i w_i$$

The fit function used by QLSPA also has a variance parameter penalty term, and is commonly referred to as the extended quasi-likelihood (or extended deviance). Note that reducing case weights always reduces Λ_w, but reductions in case weights can increase the QLSPA variance penalty term by changing the variance degrees of freedom; hence, changing case weights can either increase or decrease the extended quasi-likelihood. The rationale for investigating case weights is to assess whether small changes in the weights can remove residual patterns. The catch, M, and Q perturbations were of the form $C_w = C(1+w)$, $M_w = M(1+w)$, and $Q_w = Q(1+w)$.

The deviance LID's are presented in Figure 5.4.2.3 for the four perturbation schemes. The case weight slope is very small compared to the others, and this suggests that changes to case weights have little effect on the extended deviance (i.e. goodness of fit) compared to the other perturbation schemes. The Q perturbation scheme has the largest slope which suggests that improvements to the SPA goodness-of-fit can be obtained using smaller changes to the catchability model than to the catch inputs or M. An implicit assumption in this conclusion is that the scale of the perturbations schemes are comparable. This is difficult to assess, and involves issues like deciding if a doubling of catchability is the same sized perturbation as a doubling of a catch. We further assess the plausibility of the perturbations later in this section.

The utility of the local influence diagnostics depends on the linearity of the influence surface. If this surface is substantially nonlinear then describing influence is difficult because patterns of relative influence will depend on the magnitude of the perturbations. When the influence surface is linear within a reasonable neighbourhood of the origin then the results in Figure 5.4.2.3 will provide a good description of influence for all perturbations within this neighbourhood. To check for this, perturbations were made to the SPA inputs in the direction of maximum slope (s_max) and in some individual directions. The SPA was re-estimated and the perturbed extended deviance was compared with the unperturbed deviance. The results are presented in Figure 5.4.2.4. The influence graphs for the directions of maximum slope are reasonably linear, which is also the case for some of the individual perturbations; however, some of the individual perturbations have substantial nonlinearity. This causes some difficulty when interpreting s_max perturbations. These perturbations bound the effect when the influence surface is linear, but this may not otherwise be the case. Another problem in Figure 5.4.2.4 is that the magnitudes of the perturbations are not all comparable. Case weights were varied between 0 and 1; catchabilities, catches and M were adjusted respectively by factors between 0.7 and 1.3, 0.5 and 2, and 0.9 and 1.3. The lack of comparability was unintended.

Perturbations were applied in the directions shown in Figure 5.4.2.3 to reduce the fit function. The amount of perturbation applied was subjectively chosen in an attempt to remove the residual patterns indicated in Figure 5.4.2.2. The resulting residuals are shown in Figure 5.4.2.5. Included in this figure are the un-perturbed QLSPA residuals, which are very similar to those in Figure 5.4.2.2. The perturbations were of the form $w = 1 - h \times s\_max$, where h=-1, -1, and –1.5 for the Q, C, and M perturbations, respectively. The residual patterns based on the Q perturbation are somewhat preferable to the others. Note that some residual trends are still apparent in the Q perturbation analysis. The perturbations, particularly those based on Q, increased the residual variation somewhat. This is analogous to the bias-variance trade-off that commonly exists in many estimation problems. This trade-off involves increased variance for less biased estimators, which is similar to the pattern of increased variability for residuals with smaller trends which is apparent in Figure 5.4.2.5.

The fit of the Q perturbed QLSPA was the same as the unperturbed QLSPA fit. This is in contrast to Figure 5.4.2.4 that shows a decrease in fit for perturbations in the negative direction of s_max from the Q perturbations; however, this decrease was not sustained when $h<<-0.3$ so that when h=-1 the perturbed and un-perturbed fits were almost identical.

This indicates nonlinearity in the influence surface in the direction of s_max when |h|>0.3. Hence, there may exist other Q perturbations of a similar magnitude to h=-1 that result in a better fit plus better residual patterns.

To better understand the Q perturbation residuals patterns we present them in separate panels shown in Figure 5.4.2.6. For comparison purposes we also present the un-perturbed residuals in Figure 5.4.2.7. Bearing in mind the different scales of the vertical axes in these figures, the trends for ages 3-6 are smaller with the Q perturbations; however, the trends at the older ages do not seemed to be improved and are worse at age 11.

The conclusion from the residual analyses is that the Q perturbations were only partially successful in removing trends in residuals.

The plausibility of the perturbations also needs to be assessed in addition to assessing the ability of the perturbations to correct residual trends. The perturbed estimates of catchability are presented in Figures 5.4.2.8 and the perturbed catches and M's are presented in Figure 5.4.2.9 and 5.4.2.10. The perturbations to Q and M appear to be of a similar magnitude; however, the perturbations to catches seem smaller. None of the perturbations seem to be implausibly large.

With simulated data the true population size is known. We can therefore compare the perturbed and un-perturbed estimates of stock size with the true values. The results of such comparisons are presented in Figures 5.4.2.11 and 5.4.2.12. All estimates are very similar; however, they diverge substantially from the simulated true population values in the recent years. For the Q perturbed QLSPA this is mainly because the perturbed values of Q's have not have not detected the magnitude of the trend in Q's for ages 3-6, and have not detected the trend at all for the other ages (see Figure 5.4.2.8).

Our main conclusion from the local influence analyses of the simulated data set with high fishing mortality and a trend in survey catchabilities was that the diagnostics successfully indicated that the mis-specification involved catchabilities and not catches or M. We did not find perturbations to catches or M that could remove the residual trends to the extent that perturbations to catchabilities could. However, the perturbed catchability estimates were substantially different from the simulation values, as were the perturbed estimates of stock size. The diagnostics correctly pointed to the mis-specified component, but other methods may be required to accommodate for the mis-specification. For example, adding a smooth trend for all ages in the catchability model might result in improved estimates of stock size. Also, considering LID's based on a reduced perturbation space might yield improved diagnostics and estimates. For example, if we used only annual perturbations to all ages in the catchability model then our results might be improved. Fortunately this is simple to do with LID's, and **is useful to investigate for the next meeting of the working group**.

5.5 Influence diagnostics to diagnose the cause of retrospective patterns

The retrospective problem involves systematic differences in sequential population analysis (SPA) estimates of stock size or some other quantity in a reference year. The differences occur as successively more data is used for estimation. The most common cause of differences is structural biases that result from a mis-specification of the SPA; e.g. see Section 3 and 4 in ICES (2002a). It is usually difficult in practise to determine which of the causes are more likely. The working group (ICES, 2002a) is interested in whether local influence diagnostics (LID's) can be utilised to find small changes or perturbations to SPA input components that remove or reduce retrospective patterns. If so, the plausibility of the perturbations can be used to assess which component is the likely source of the retrospective pattern.

The retrospective patterns (see Figure 5.5.1) in QLSPA estimates were small even for the high F data set, and because of that we decided that a local influence analysis of these patterns was of secondary importance compared to the extended deviance analysis presented above. A metric of the retrospective pattern may be a better diagnostic of model mis-specification than the deviance. Hence, it would be useful to assess the extent to which local influence perturbations that remove retrospective patterns could also indicate the form of the SPA mis-specification. An illustration of how to use LID's to diagnose the cause of retrospective patterns is provided by the case study of Eastern Scotian Shelf cod presented in Appendix C. This was the same stock considered by Mohn (1999). We present conclusions from this study in the next Section 5.5.1.

The working group encountered some difficulties in generating simulated data sets with significant retrospective patterns by introducing a trend in the catchability model, although in ICES (2002a, Section 4.2.1) a trend in catchabilitity was identified as an important source of retrospective patterns. **This finding supports the conclusion made in ICES (2002a) that the lack of a retrospective pattern cannot be taken as *proof*' that the model is valid**.

5.5.1 Local influence diagnostics for retrospective patterns in Eastern Scotian Shelf cod

LID's were applied to an example SPA for the eastern Scotian Shelf (ESS) cod stock, which has a severe retrospective pattern. The local influence methods were used to find changes to model assumptions or data that reduced or removed the retrospective patterns. The analyses suggested that reasonable changes in assumptions about model errors are not a likely source of the retrospective pattern; however, additional information seemed necessary to discriminate between the likelihood's that the source of the retrospective patterns were catches, natural mortality, or assumptions about survey catchability. This was the same conclusion reached by Mohn (1999), although he decided that catchabilities were the most likely source of the retrospective patterns based on additional information.

Note that the M perturbation scheme used in Appendix C was different than that used in Section 5.4.2. An additive perturbation was used for the analysis in Appendix C, whereas a multiplicative perturbation was used in Section 5.4.2 to make the perturbation scheme more comparable to the multiplicative catch perturbation scheme that was used in both Section 5.4.2 and Appendix C.

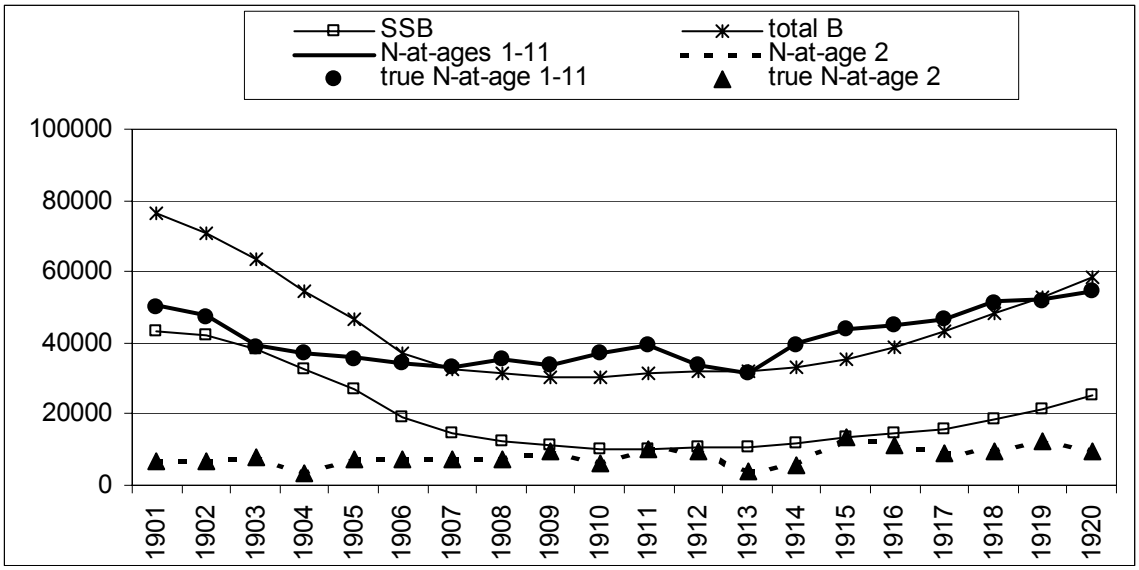


Figure 5.4.1.1 QLSPA estimates of the stock dynamics for the "exact" simulated data set with no trend in q.

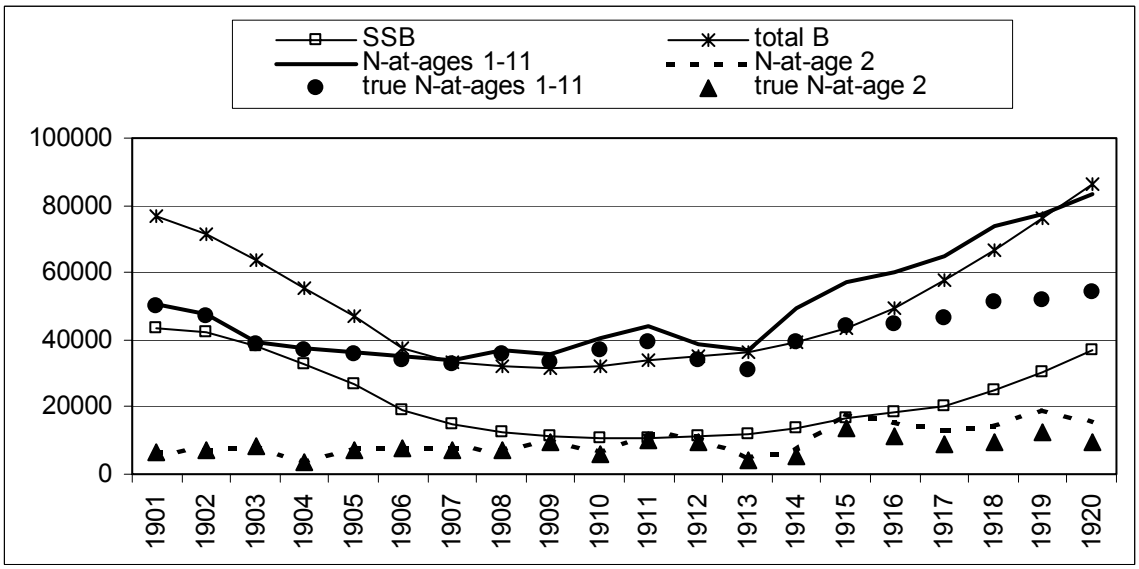


Figure 5.4.2.1 QLSPA estimates of the stock dynamics for the q-trend simulated data set. A.

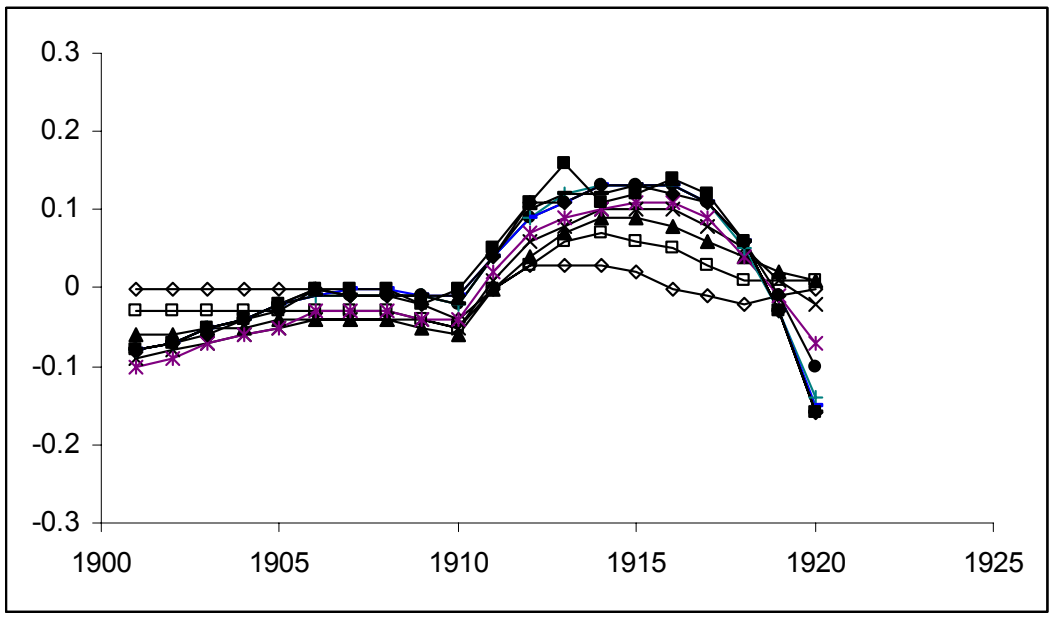


Figure 5.4.2.2 The log catchability residual patterns at age resulting from the fit of the XSA model to the high F simulated data set with a trend in catchability starting in year 1919. Similar residual patterns were derived from the QLSPA model fit.

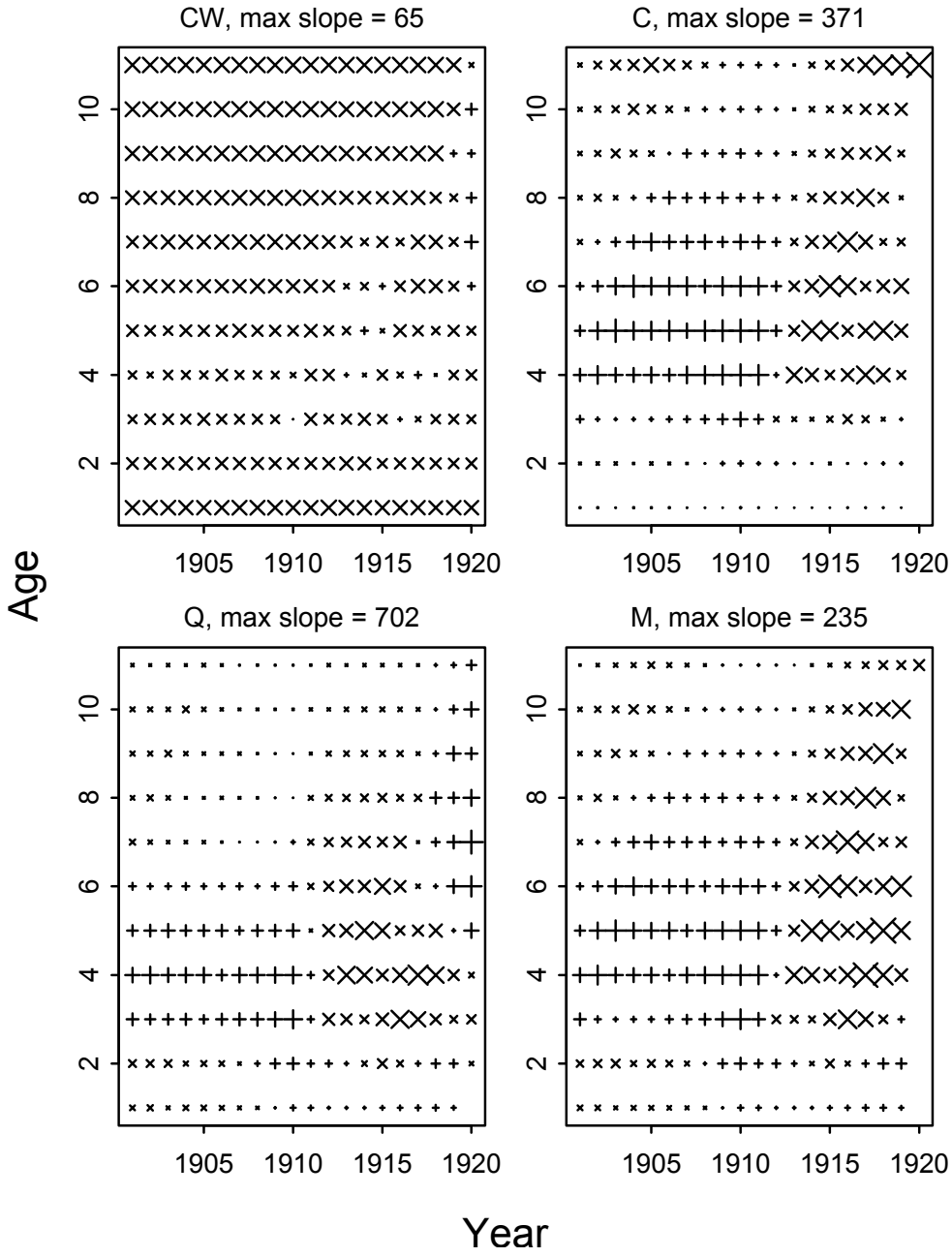


Figure 5.4.2.3 Local influence results for the extended deviance fit function. Each panel shows the elements of the direction of maximum local slope of the influence surface for four perturbation schemes: CW – case weights, Q – survey catchability, C – catch, M – natural mortality. The maximum slope is indicated at the top of each panel. The size and type of the plotting symbols are proportional to the absolute value and sign of the elements, respectively. Negative is denoted by an ×.

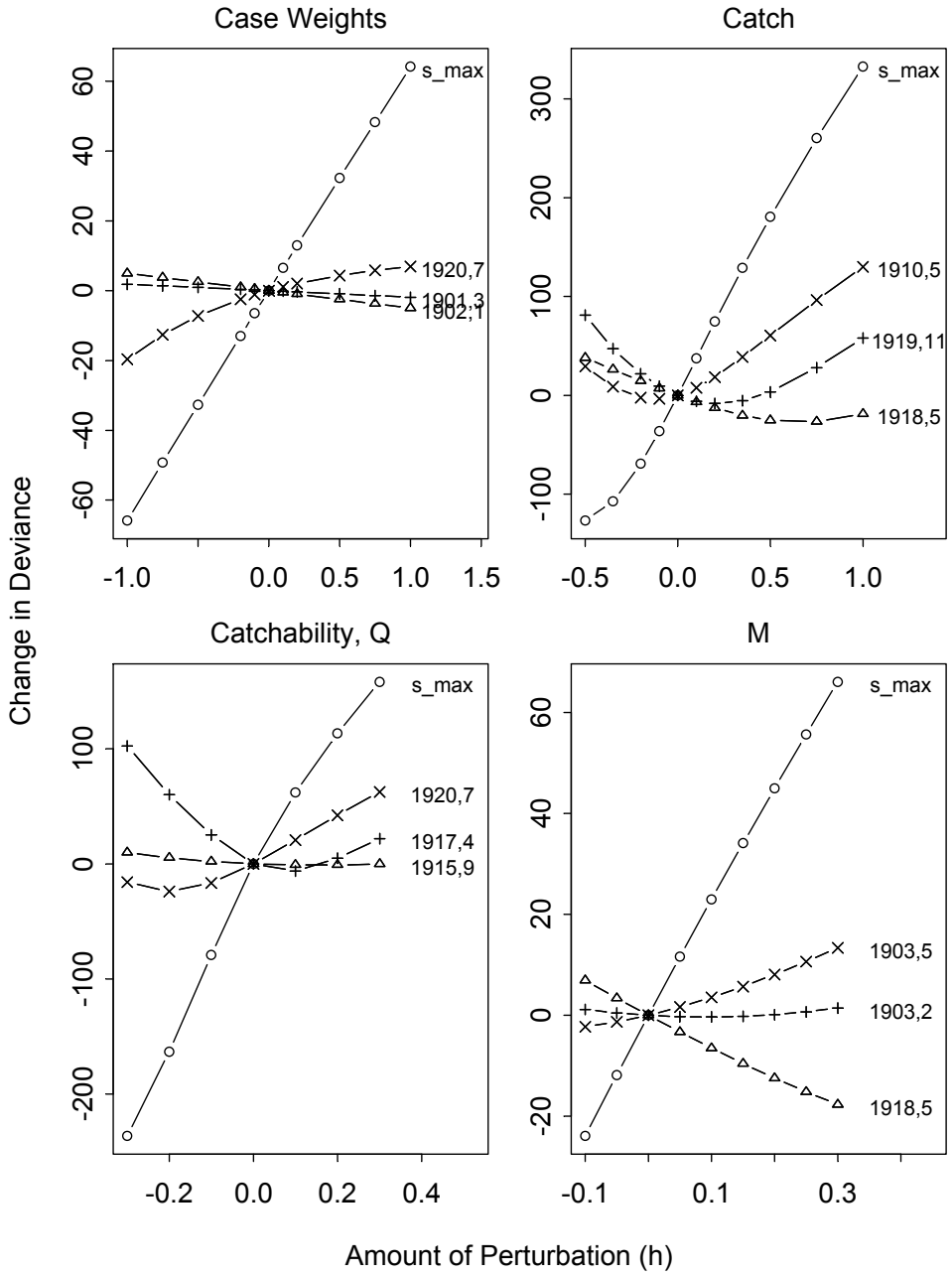


Figure 5.4.2.4 Displacement in QLSPA extended. The (year,age) indicate perturbations of individual components. The s_max perturbations are in the corresponding directions shown in Figure 5.4.2.3. Note that the scales of the horizontal axes differ.

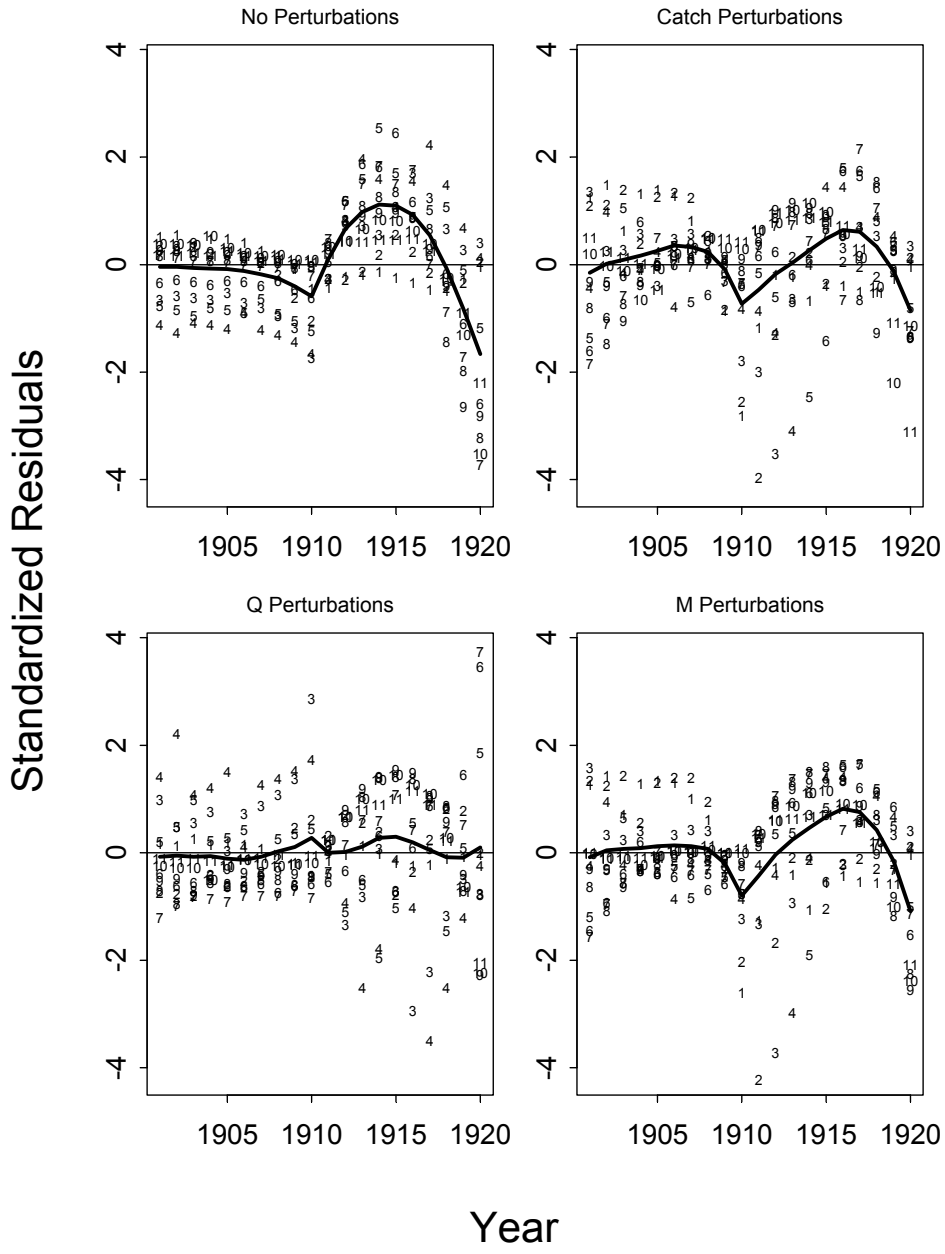


Figure 5.4.2.5 Annual time-series of variance standardized residuals from s_max perturbed and un-perturbed QLSPA analyses. The plotting symbols are the ages corresponding to the residuals. The heavy solid line connects the average residual each year.

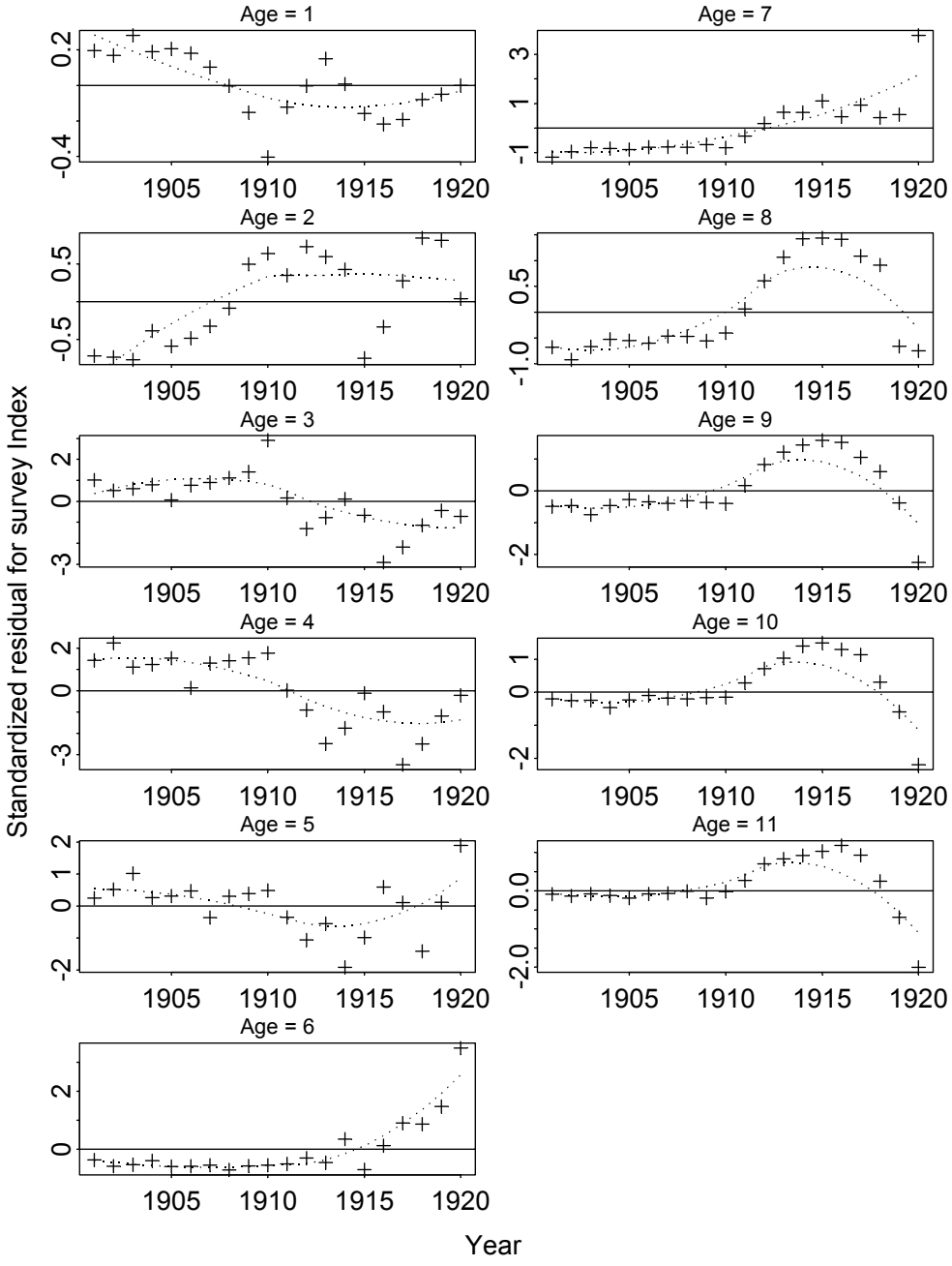


Figure 5.4.2.6 Annual time-series of variance standardized residuals from the s_max Q perturbed QLSPA. The dotted solid line connects the average residual each year.

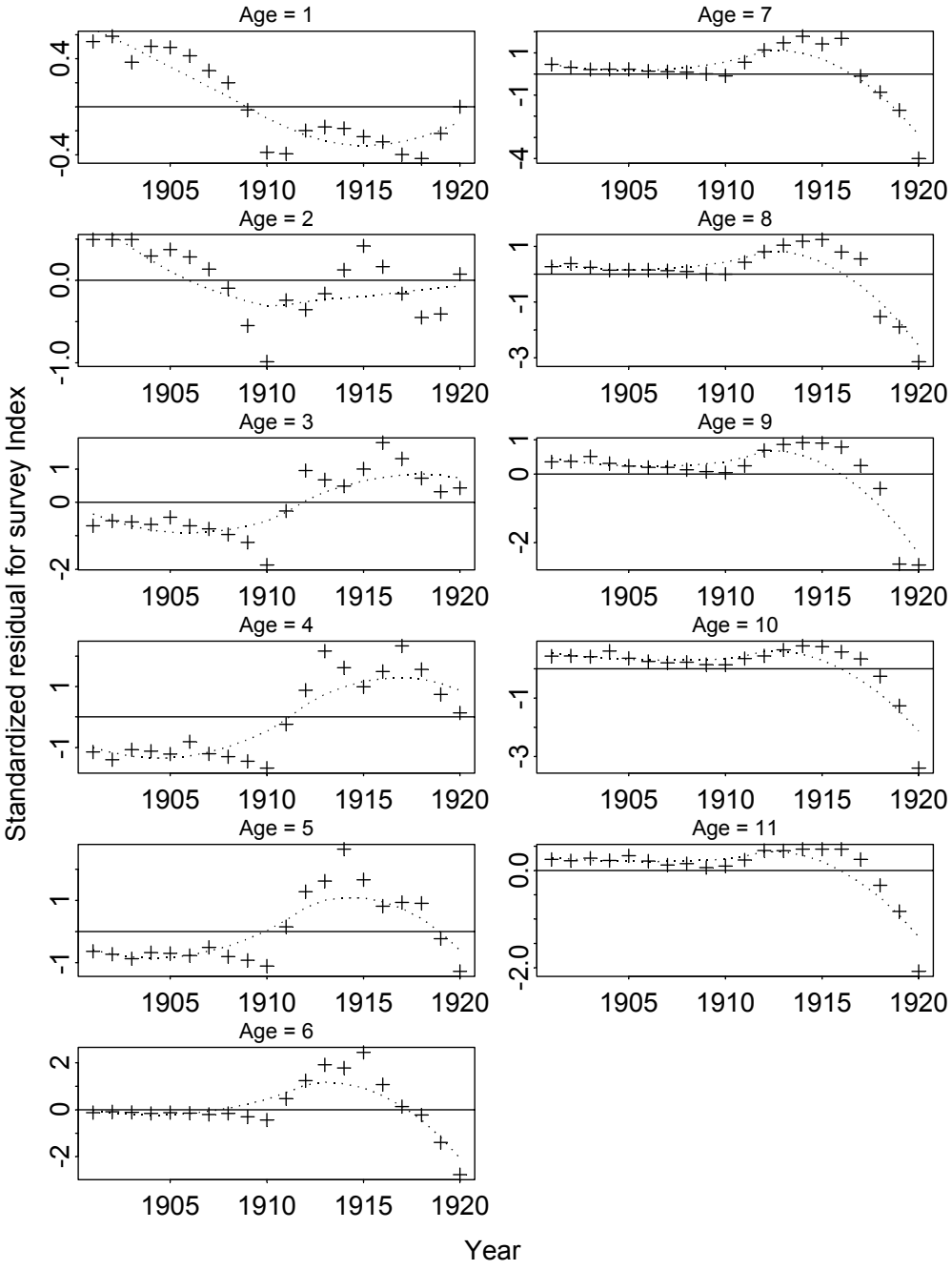


Figure 5.4.2.7 Annual time-series of variance standardized residuals from the unperturbed QLSPA. The dotted solid line connects the average residual each year.

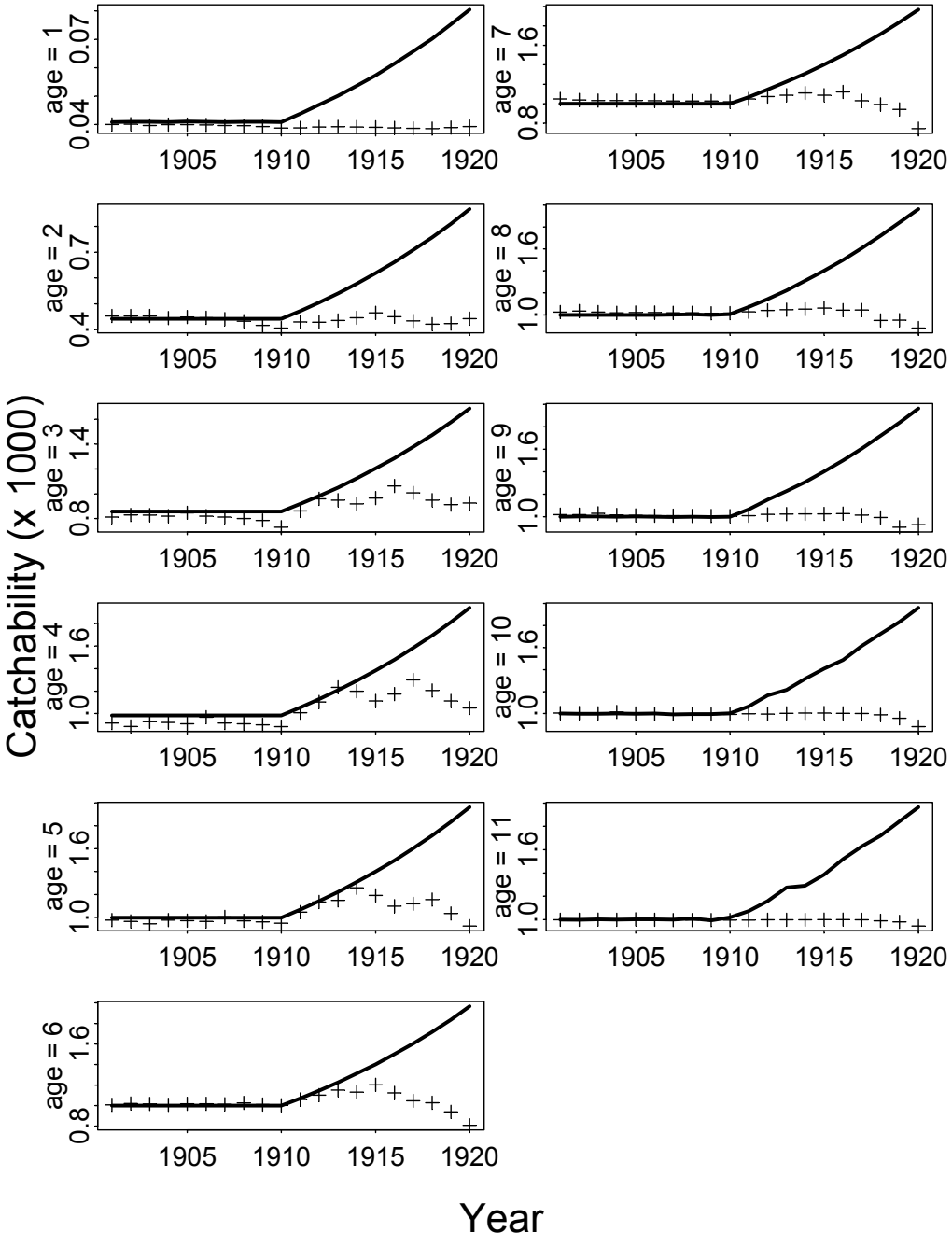


Figure 5.4.2.8 Perturbed estimates of survey catchabilities (+'s). The perturbed Q's are of the form $Q_{a,y} = Q_a$ x (1 + $w_{a,y}$), and Q_a is estimated. The solid line connects the values of Q's used to generate the simulated survey data.

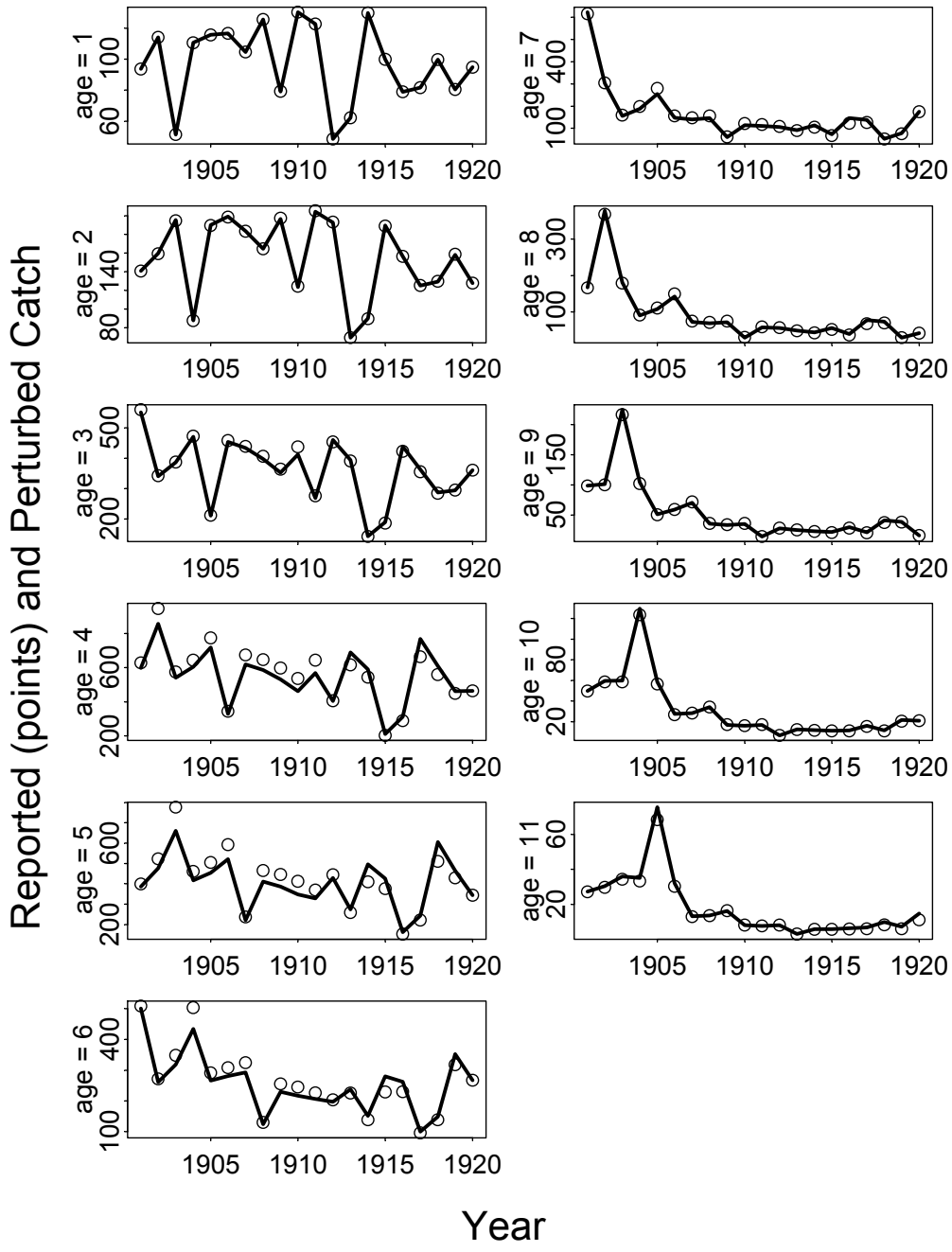


Figure 5.4.2.9 Perturbed catches, shown as solid lines. The o's indicate the un-perturbed catches.

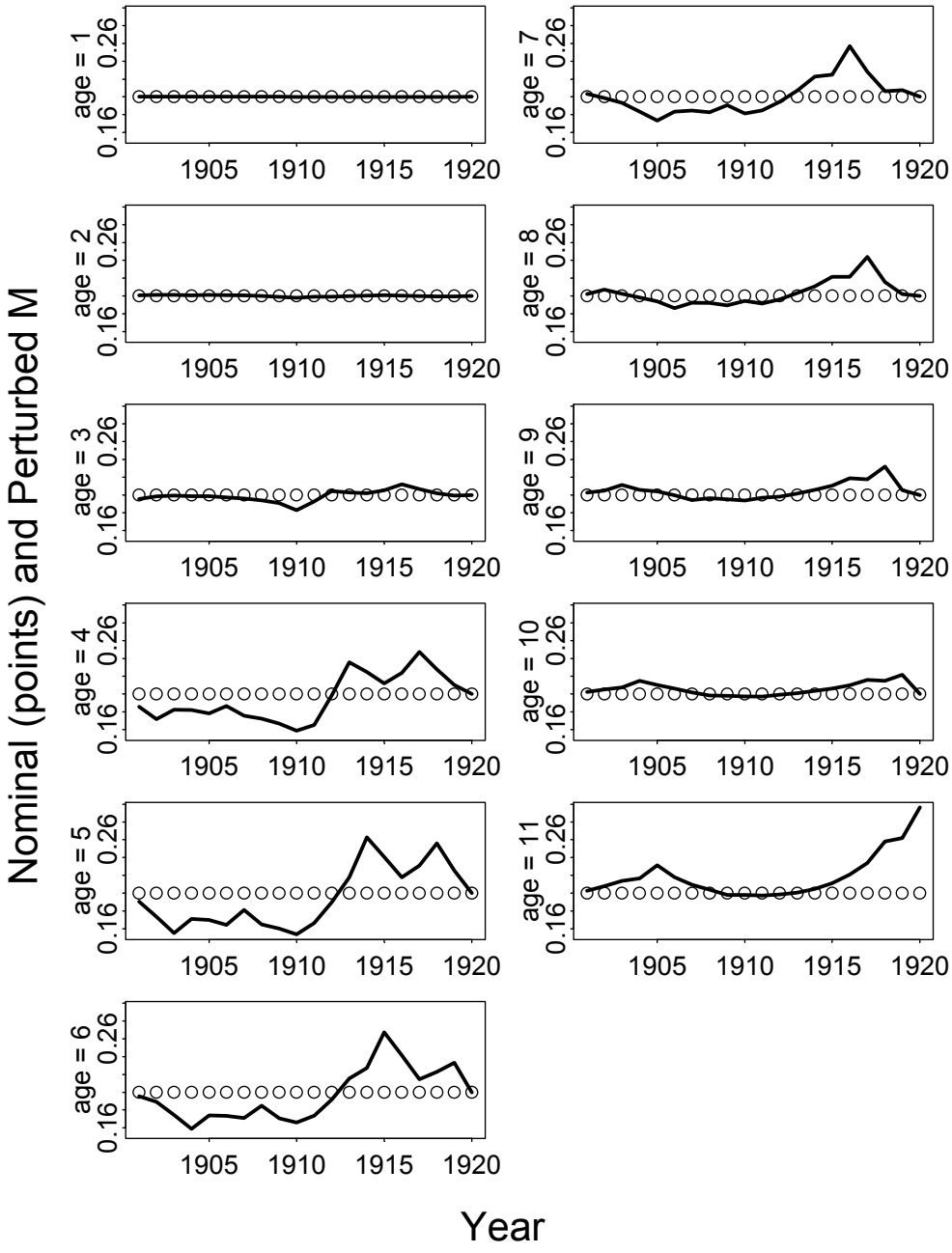


Figure 5.4.2.10 Perturbed natural mortalities, M, shown as solid lines. The o's indicate the un-perturbed M's.

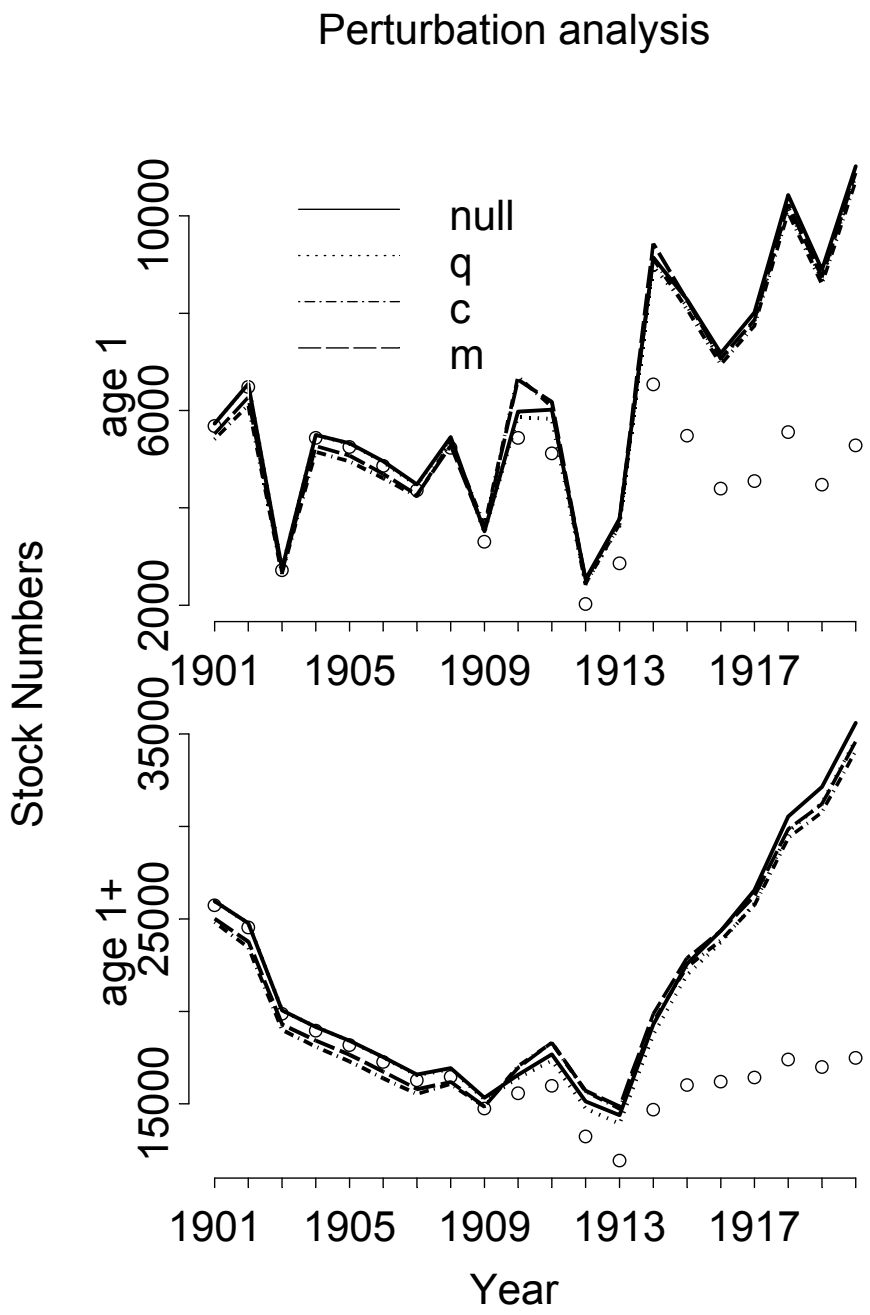


Figure 5.4.2.11 Estimates of numbers-at-age 1 and total numbers-at-ages 1-11 from the un-perturbed QLSPA (null), and the catchability (q), catch (c), and M perturbed QLSPA's. The simulated true values of stock size are plotted as o's.

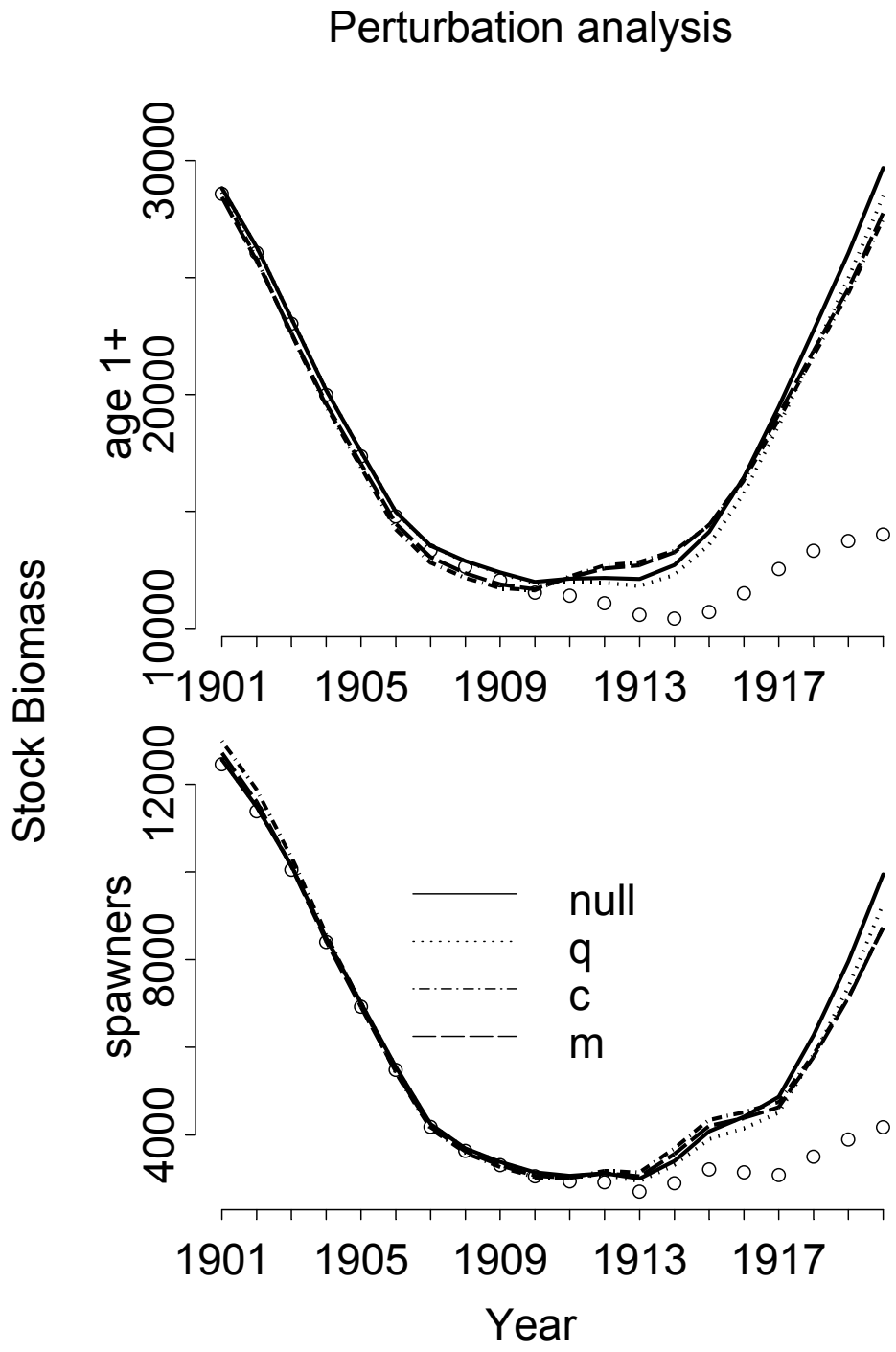


Figure 5.4.2.12 Estimates of spawner biomass and total biomass at ages 1-11 from the un-perturbed QLSPA (null), and the catchability (q), catch (c), and M perturbed QLSPA's. The simulated true values of stock size are plotted as o's

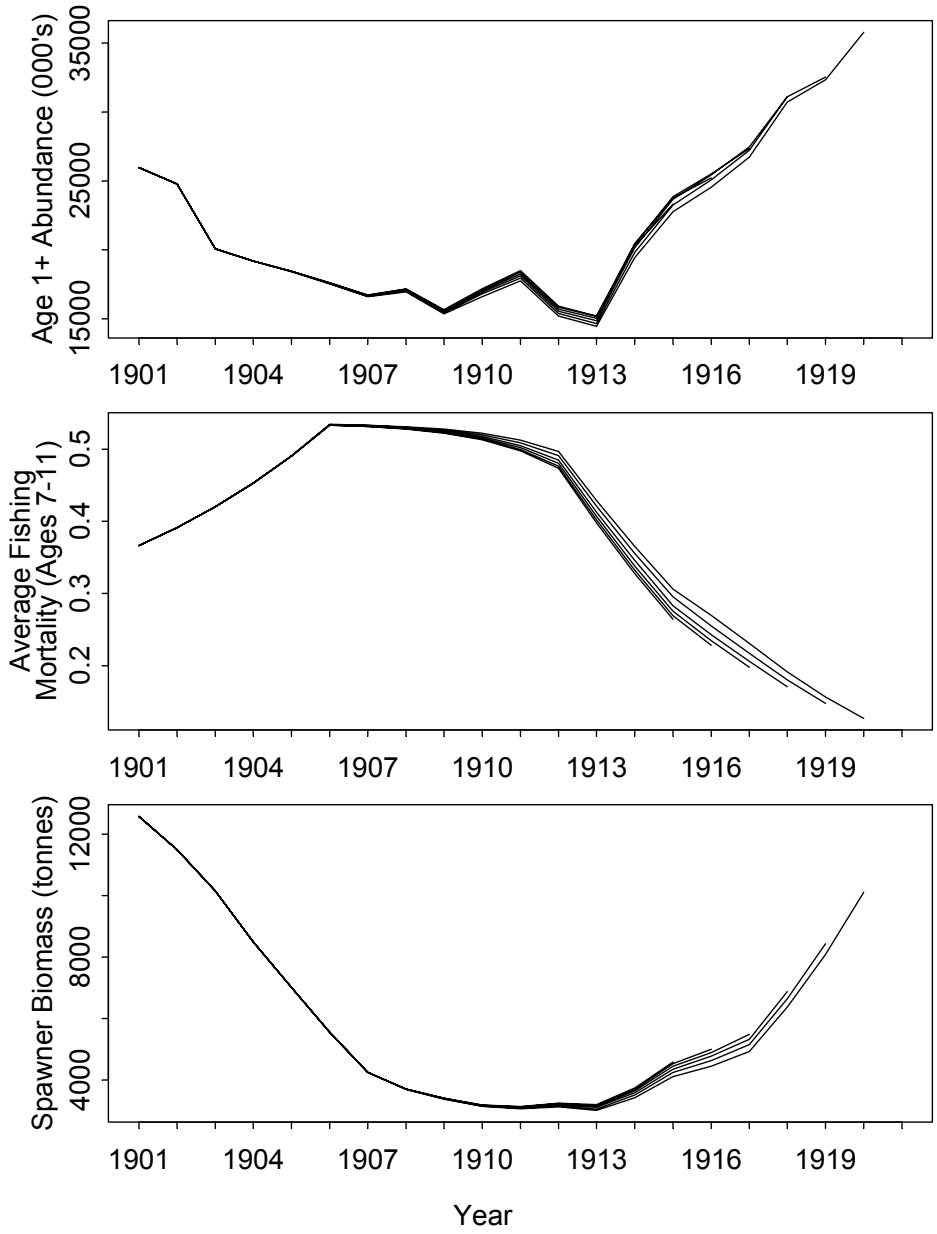


Figure 5.5.1 Retrospective estimates of 1+ abundance, fully recruited F, and spawner biomass from the high F simulated data set.

5.6 Other approaches

5.6.1 Introduction

The Working Group members' thought it useful to summarize some of the methods and tools used for producing sensitivity analysis and influence diagnostics currently in use. These are given in Sections 5.6.2, 5.6.3 and 5.6.4. There is also included a short summary of a paper presented at the ICES Statutory meeting in 2002 dealing with the identification of structural uncertainty in age-structured fish stock assessments (see Section 5.6.5). WGMG had little opportunity to discuss the latter during its meeting but consider that the approached proposed in Patterson (2002) is an interesting one that is worthy of further investigation by the Group members'.

5.6.2 Sensitivity analysis in stock projections

Cook (1993) describes two methods of sensitivity analysis to quantify uncertainties in stock projections. These are a local linearised analysis (from Prager & MacCall 1988) and the Fourier Amplitude Sensitivity Test (from Cukier *et al*. 1978). The first method is implemented in WGFRAN4 program with output plotted using the SENPLOT program and in the windows program WGFRANSW. The results are dependent on assumptions about the uncertainty in the input parameters (numbers-at-age, weights-at-age, natural mortality, selectivity pattern, maturity-at-age and recruitment). This program produces three types of diagnostic output in addition to the forecast:

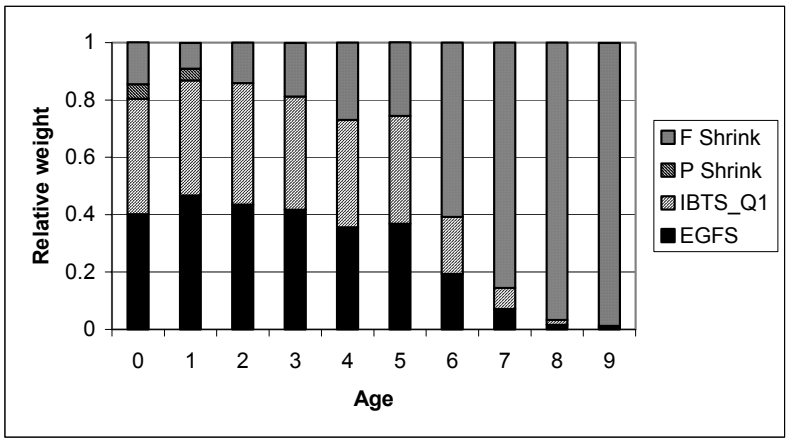
- The rate sensitivity coefficients for each input parameter is ranked and plotted in a histogram. This is a useful tool for identifying the input having the biggest influence on the forecast.

- Partial variances are calculated assuming negligible covariance and presented in pie graphs as percentages of the total variance. The partial variance is depending on both the rate coefficient and the assumed variance in the input parameter.

- Probability profiles of yield versus the probability of F exceeding the status quo fishing mortality and the probability that the SSB will fall below a certain level in the forecast period.

5.6.3 Relative weight of tuning data used as influence diagnostics

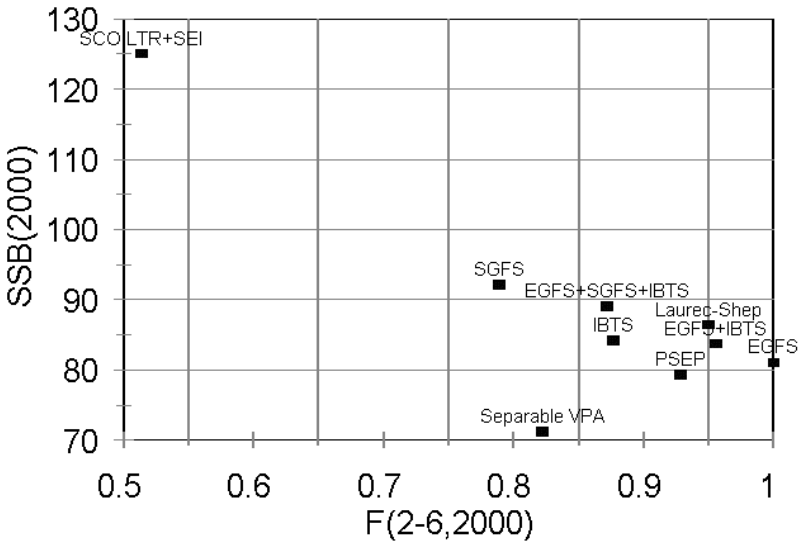
A traditional method of producing influence diagnostics is the commonly used practice of plotting the relative weights given to the different tuning series in the current year prediction. The following figure is an example of an influence diagnostic from a XSA using two tuning fleets and allowing shrinkage. The usefulness of this diagnostic is rather limited and the figure is only showing the relative contribution from the different tuning fleets and "shrinkage" in the prediction.



5.6.4 Choice of tuning series and their influence

Another approach shows the influence of including and excluding particular tuning series. The figure below is an example on how the results (estimates of SSB and $\overline{F}_{2-6}$) vary with varying the choice of tuning input while the remaining input and assumptions are kept constant. This approach can be useful when looking for conflicting signals from different surveys. But one should keep in mind that the single fleet tuning results may not represent the single fleet

contribution in a combined tuning. And it should also be noted that the different fleets usually represents different age groups and one would usually have chosen different model assumption for single fleet assessments.



5.6.5 Quantifying structural uncertainty in age-structured models

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 (ICES 2002a).

Recently, Patterson (2002) made the following listing of some constraints that may be imposed to obtain acceptable solutions to age-based VPA type models:

1. Constraining F at the oldest age to a fixed proportion of an average of F over younger ages, as in ADAPT and Laurec-Shepherd methods;

2. Constraining F at the oldest age to a fixed proportion of an F at a single younger age, as in CAGEAN and Fournier-Archibald based models;

3. Constraining ("shrinking") F at the oldest age towards a weighted average over younger ages, as in XSA;

4. Constraining some F in the last year as a function of other F (e.g. shrinkage in XSA);

5. Constraining the standard deviation of the logged tuning index to any fixed value, either estimated externally or assigned assumed values;

6. Estimating the standard deviation subject to constraints, which may be year-dependent (Cleveland taper weighting) ;

7. Constraining catchability to be equal within a specified range of ages up to the oldest age, in XSA or ADAPT; and

8. Choosing a functional form for catchability (e.g. as dependent on N, on time, or as a random walk in a time-series model).

Constraints 1 to 3 reflect perceptions of the extent to which fishing mortality is stable across different age-groups (either within the whole fishery, or for a particular fishing gear). Similarly, constraint 4 reflects perception of the extent to which fishing mortality is stable across some years. For given tuning index, constraints on fishing mortality are equivalent to constraints on catchability.

Constraints 5 and 6 reflect perceptions of the relative accuracy with which different surveys or fishing fleets provide information about stock abundance.

Constraint 7 reflects the perceptions that for some kinds of fishing gear, fish can be considered "fully recruited" above a certain age, *ie* the catchability remains constant with increasing age.

Constraint 8 reflects perceptions that the efficiency of a fishing gear or fishing fleet may alter according to some variate, which can be estimated, and that this dependency can be modelled.

Patterson chooses to further investigate an approach were he lets the catchability be estimated through a kernel smoother restraining the rate of change with two parameters. One parameter restraining the rate of change in the age pattern, while the other is restraining the year to year change in catchability. Allowing for more variation (and “loss of df”) introduces a penalty to the likelihood term (relative penalised likelihood factor). The robustness of the estimate of interest to alternative model choices could then be evaluated by looking at a range of smoothing parameters. This can be used as a guide to finding the “best” model. Furthermore it can serve as a sensitivity test, quantifying how much results depend on assumptions about how fast catchability may change.

5.6.6 Suggestion for further work

The WGMG propose the following ToR for their next meeting:

To investigate and implement statistical approaches that identify and quantify uncertainty due to conditioning choices in fish stock assessments.

6 ALTERNATIVE STOCK ASSESSMENT METHODS

Declining fish stocks and stringent management measures mean that commercial catch data have become more and more unreliable over recent years. Most commonly-used assessment methods are based on such data, and there is therefore an increased likelihood of biased fisheries management advice. This means that there is a growing need for stock assessment methods which do not use commercial catch data, or which do not require such data in the standard age-disaggregated form. Work on such methods is important so that we will be able to provide valid alternatives when fisheries managers finally acknowledge that commercial fisheries data are unreliable in the assessment context.

6.1 Analyses of simulated data with SURBA

6.1.1 Introduction

SURBA 2.0 is an implementation of a separable model using survey-index data. The original model is presented in Cook (1997). The method is described in Appendix B, and has been used to generate supplementary assessments for several stocks in the North Sea and Northern Shelf areas (ICES 2002cd). However, the properties of this implementation of the model have not previously been tested using simulated data sampled from a generated population about which the true state is known. It has therefore not been clear whether the model is performing as expected. The following analyses are intended to fill this gap. They are based on the simulated datasets described in Section 3, which are denoted here as follows:

Dataset 1: Exact data, low simulated *F*, no catchability trend, mesh size 20 mm, 11 ages over 20 years (the plus-group had to be removed from this and datasets 2 and 5 as there are currently problems with the modelling of plus-groups in SURBA).

Dataset 2: Exact data, low simulated *F*, catchability trend (4% each year from the tenth year onwards), mesh size 80 mm, 11 ages over 20 years.

Dataset 3: Noisy data, low simulated *F*, no catchability trend, 14 ages over 25 years.

Dataset 4: Noisy data, low simulated *F*, catchability trend (4% per year over the last 10 years), 14 ages over 25 years

Dataset 5: Exact data, high simulated *F*, catchability trend (7% per year over the last 10 years), 11 ages over 20 years.

For each dataset, a SURBA input file was created containing the survey index values, natural mortality-at-age, proportion mature-at-age, and stock weights-at-age. Each file also contained default age weightings (which were all set equal to 1.0 for these simulated datasets), estimated catchabilities-at-age, and an estimated catch-at-age selection pattern from the commercial fishery. The estimated catchabilities were $\mathbf{q} = [1.0,1.0,\ldots,1.0]$ (dataset 1), $\mathbf{q} = [0.036,0.4,0.8,0.95,0.99,1.0,1.0,\ldots,1.0]$ (dataset 2), $\mathbf{q} = [0.05,0.1,0.45,1.0,1.0,\ldots,1.0]$ (dataset 3), $\mathbf{q} = [0.045,0.1,0.25,0.6,1.0,1.0,0.9,0.9,\ldots,0.9]$ (dataset 4) and $\mathbf{q} = [0.03,0.36,0.8,1.0,1.0,\ldots,1.0]$ (dataset 5).

The reason for including the commercial selection pattern was as follows. The key difficulty at the moment with SURBA, and indeed with all exclusively survey-based assessment methods, is the specification of catchability for the survey. The temporary solution used during this meeting was to fit the SURBA model with all catchabilities fixed at 1.0. We then fitted a separable VPA to the corresponding catch-at-age data from the simulated dataset using the Lowestoft VPA package (Darby and Flatman 1994), assuming equal weighting on all years, a selection of 1.0 on the oldest age, and a terminal F from the first SURBA model fit. We scaled the fitted selection pattern from the catch-at-age separable VPA so that the plateau was at the same approximate level as the plateau from the first SURBA fit. This scaling is necessary because SURBA does not fix the values of any age-effects (selectivities), relying instead on restrictions to year-effects (temporal trends). We then manipulated the catchabilities until the fitted SURBA age-effect was "close enough" to the scaled catch-at-age separable VPA age-effect.

This procedure gives an empirical procedure for estimating survey catchability, but it is clearly not very satisfactory. Catchability remains the central problem with the use of survey-based methods in generating fisheries management advice. One possible way might be to include the sum-of-squares error between survey and catch-at-age selectivity in the overall SURBA minimisation, using the catchabilities as parameters. Of course this makes use of catch data, the avoidance of which (due to perceived increases in misreporting) is the main justification for considering survey-based assessment at all. However, the impact of catch-data problems might be lessened by block-bootstrapping the years over

which the catch-at-age separable model is fitted, possibly with a bias towards earlier years about which we have more confidence. This would have two benefits: firstly, it would shift attention away from the later years of catch data; and secondly, it would provide a simulation envelope and would allow us to evaluate the sensitivity of the method to assumptions about catchability. On the other hand, it may not be feasible to estimate separable-model parameters and catchabilities simultaneously.

6.1.2 Results

We applied the procedure described above to each of the simulated datasets in turn, and compared the estimated SSB and mean *F* time-series with those from the true, underlying populations. For each of the five datasets there were four SURBA runs, as follows:

Run 1: All age-weightings set to 1.0, *F*-smoother set to 1.0.

Run 2: Inverse-variance age-weighting, *F*-smoother set to 1.0.

Run 3: All age-weightings set to 1.0, *F*-smoother set to 0.5.

Run 4: Inverse-variance age-weighting, *F*-smoother set to 0.5.

Plots comparing the outputs from each run with the true state are given in Figures 6.1.1 (dataset 1), 6.1.2 (dataset 2), 6.1.3 (dataset 3), 6.1.4 (dataset 4) and 6.1.5 (dataset 5). In each Figure the summary plot for the relevant SURBA run 1 is also given.

For dataset 1, there are good fits to both relative SSB and mean *F*, although there is a strong year-effect pattern in the residuals. Estimated relative SSB for dataset 2 shows a considerable skew, which is a direct result of the imposed trend in catchability. Mean *F* is still well-estimated for dataset 2, and the catchability trend shows up clearly in the SURBA residual plot. For dataset 3, there is some variation in relative SSB estimates, but they all still reflect well the underlying true state. It is likely that all such estimates would lie within a common simulation envelope if we were to carry out bootstrap residual runs, and this is something that needs to be done in the future. Mean *F* estimates from dataset 3 are very noisy, although again we can still make out the underlying pattern. The SURBA age-effects (selectivities) show large departures from the expected smooth curve, whilst the residuals are well-distributed with no obvious trend. The imposition of a trend in catchability (dataset 4) does not lead to great changes in the residual pattern, nor do the age-effects look any more reasonable. This dataset also shows some skew in relative SSB estimates, but the effect is not as noticeable as it is for the clean data with a catchability trend (dataset 2). The final figure shows the results for dataset 5, which has the same skew in SSB estimates, but which also underestimates mean *F* in the recent period. This may be a result of the imposed trend in catchability in that dataset.

6.1.3 Discussion and further work

As we have discussed elsewhere (Section 3), the simulated datasets used in these analyses are probably not ideal in terms of determining the reasons for occasional poor performance in models which are well-tested and documented. However, for models such as SURBA which have only recently become generally available and which have not yet been widely used, the datasets provide a good opportunity to check model specification, and to ensure that the model is performing as we would expect it to. The catchability difficulties in SURBA have long been known. The analyses described above have demonstrated that relative trends in abundance (and therefore SSB) are estimated well by SURBA, but that fishing mortality estimates are not so well estimated in the presence of noise. At least, the overall pattern in *F* is returned, but with a great deal of unwarranted variation caused by the interaction of age-effects (selectivities) with uncertain catchabilities. The imposition of a trend in catchability has the effect of increasing SSB in recent years, but this is expected and would be observed in most assessment models. In general, we should be more confident in assuming constant survey catchability than in assuming that current catch data are representative.

Potential avenues for further work on SURBA in particular, and survey-based assessment methods in general, were explored by Needle (WC1). To summarise:

- The need for the user to specify catchability-at-age for each survey is a serious problem with the method. We have already discussed a potential method to do this empirically using catch-at-age separable model outputs, but this approach is moving away from the ideal of an exclusively survey-based model. The derivation of catchabilities is a key area for future research.

- The assumption of a constant age-effect is also difficult to defend. It is not the selectivity of the survey which is being measured by the age-effect, but that of the underlying fishery, and this can be expected to change through time. It may be possible to allow the age-effect to vary empirically through time, using an iterative residual-fitting process or a Kalman filter, although if we are simultaneously trying to estimate catchabilities we may run into over-parameterisation difficulties.

- The model does not currently allow for enough uncertainty in specification. A sensitivity analysis with bootstrapped model residuals would generate simulation envelopes, thus allowing us to determine if the various SURBA "features" (noisy *F* estimates, for example) are significant.

- We have found that the model does not work particularly well with certain datasets, particularly when there are missing data in the last year. An example of this is Rockall haddock (ICES 2002c), which is surveyed only once every two years and which has a lot of missing data as a result. The reason for this problem is not clear, but could be addressed with the use of simulated datasets with missing data.

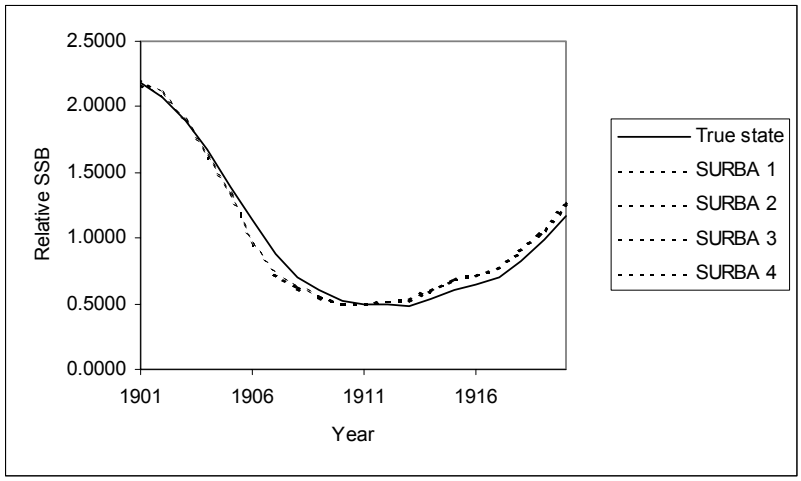
The Northern Shelf WG (ICES 2002c) made a recommendation that a Study Group of the Methods Working Group (WGMG) should be established, to look into the field of survey-based assessment in general. On reflection, it now seems that this need would be better served within the remit of WGMG itself. To this end, WGMG propose the following ToR for their next meeting:

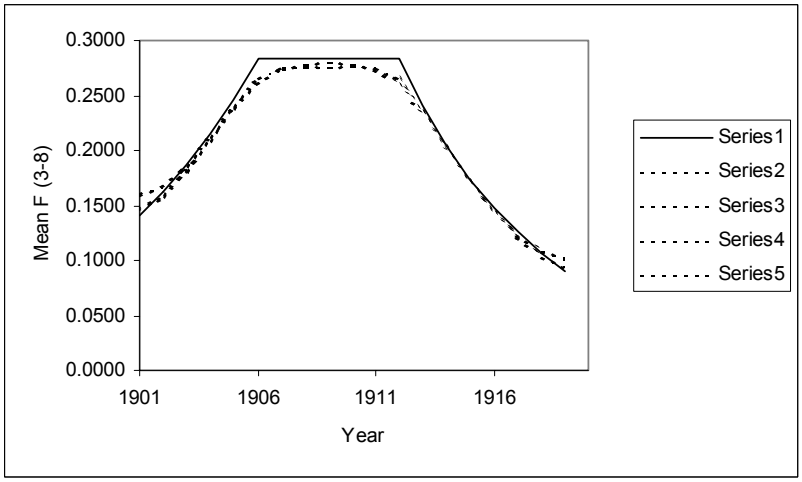
To develop fishery-independent assessment methods, measures of uncertainty, and appropriate diagnostics, with particular attention to data-poor situations and the estimation of relative catchability.

In this context, we intend that "data-poor situations" should encapsulate cases where data are limited in either extent (such as elasmobranch and deep-sea fisheries) or reliability (such as catch data of low quality) or both. This suggestion is similar in nature to the ToRs a) and c) addressed at the 1995 meeting of WGMG (ICES 1995).

6.1.4 Conclusions

It seems clear that landings data have become increasingly biased and unreliable as stocks have declined further and fisheries management measures become increasingly punitive. In order to use survey-based assessment methods, we must assume that survey indices are unbiased (though variable) representations of stock trends. With good survey design, this will be true to a first approximation. The validity of a particular survey in the assessment of a given stock has to be evaluated separately from the assessment process, and we have not attempted this here. However, if we assume that there is an answer to the catchability problem we would contend that survey-based assessments are perfectly feasible and, indeed, are the methods we are likely to have to use in the future. While development work is still needed, the methodology implemented in SURBA 2.0 is beginning to approach that which would be required for a genuinely unbiased assessment approach. One advantage of this model is that it is easy for assessment Working Groups to use – it has a straightforward user interface and produces appropriate plots automatically, although it remains to be seen whether it is doing what we think it is doing. In any case, we should encourage work on a variety of such methods.





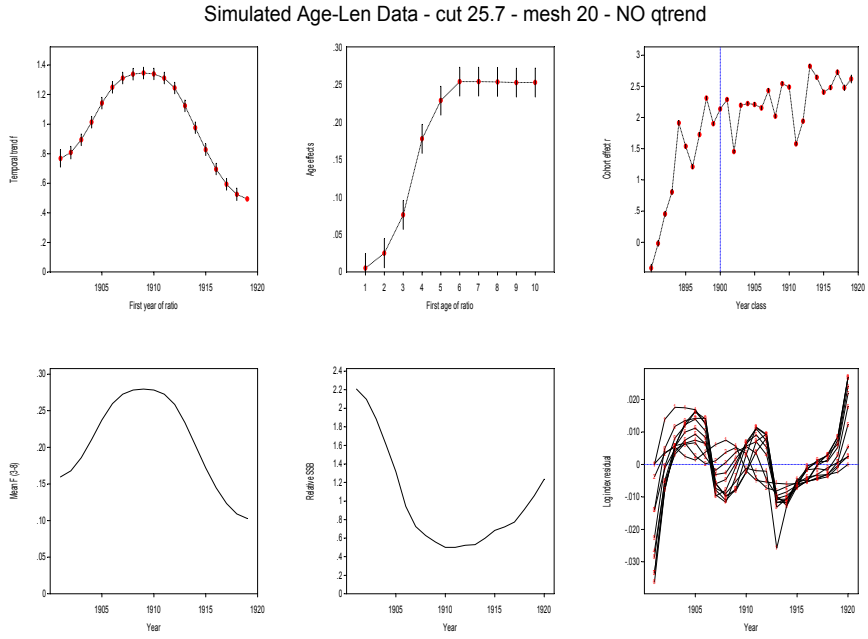
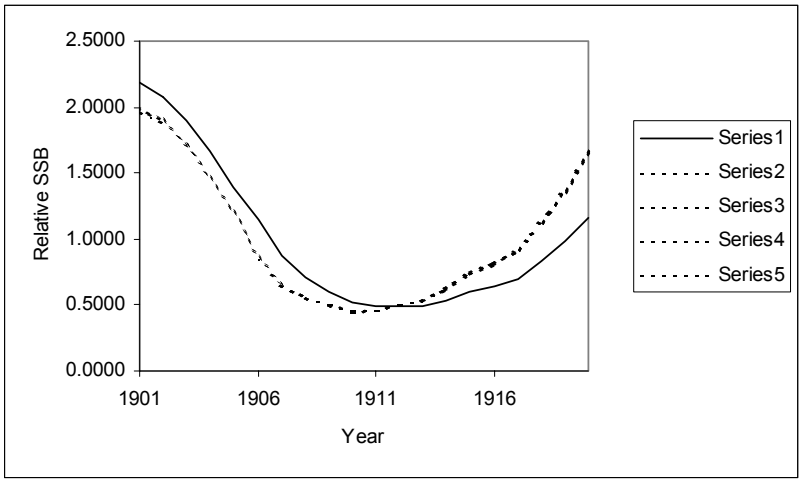
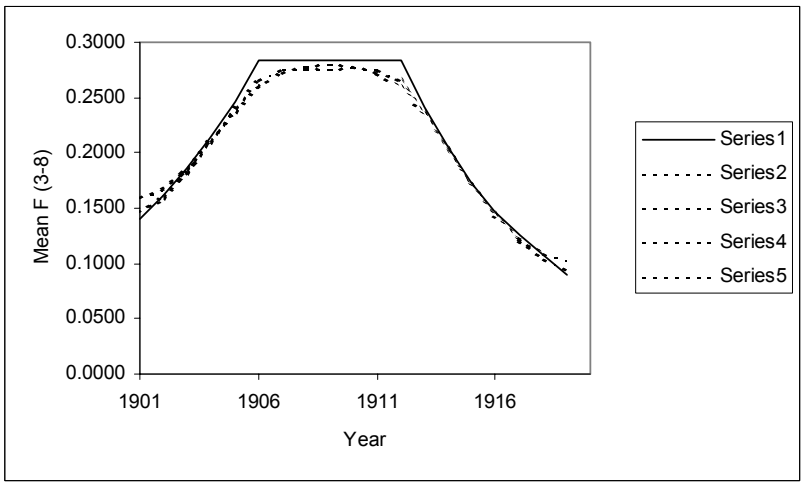


Figure 6.1.1. Comparative output plots for four SURBA model runs with dataset 1, along with the true state. See text for details of dataset and run settings. Upper panel: relative (mean-standardised) SSB. Middle panel: mean *F* (3–8). Lower panel: SURBA summary plot for run 1.





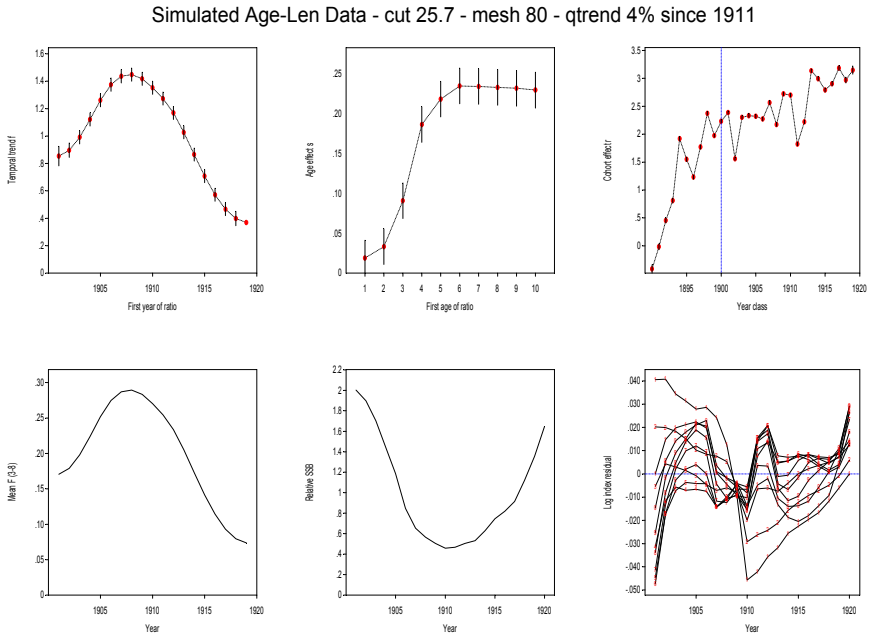
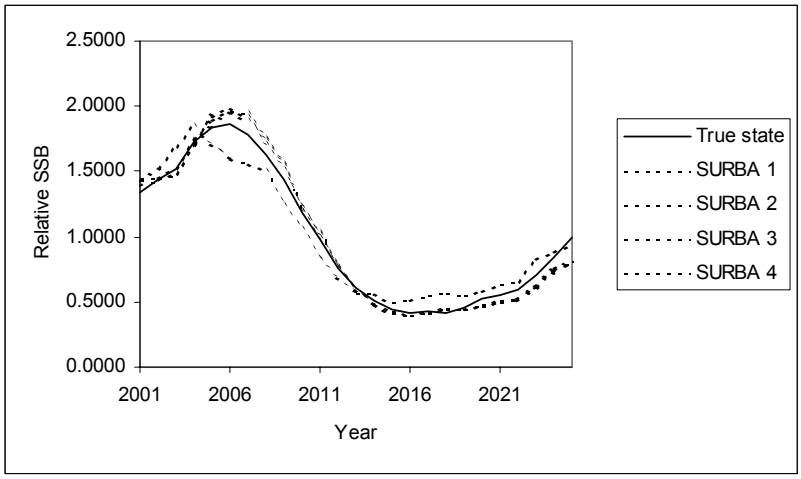
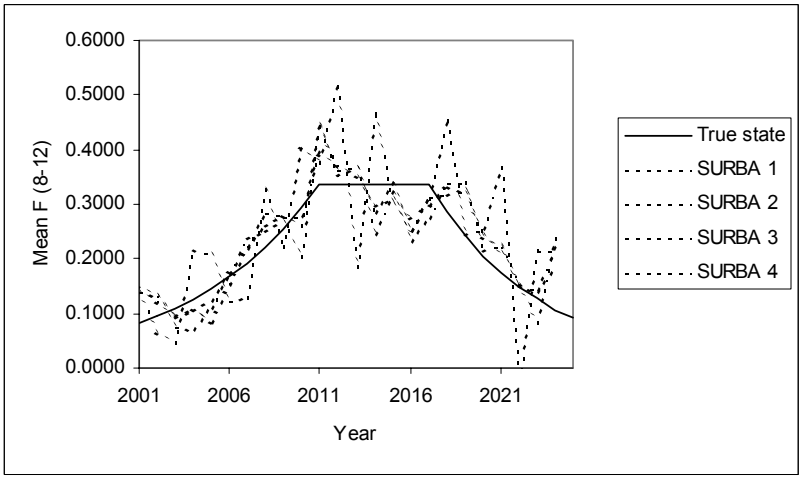


Figure 6.1.2. Comparative output plots for four SURBA model runs with dataset 2, along with the true state. See text for details of dataset and run settings. Upper panel: relative (mean-standardised) SSB. Middle panel: mean *F* (3–8). Lower panel: SURBA summary plot for run 1.





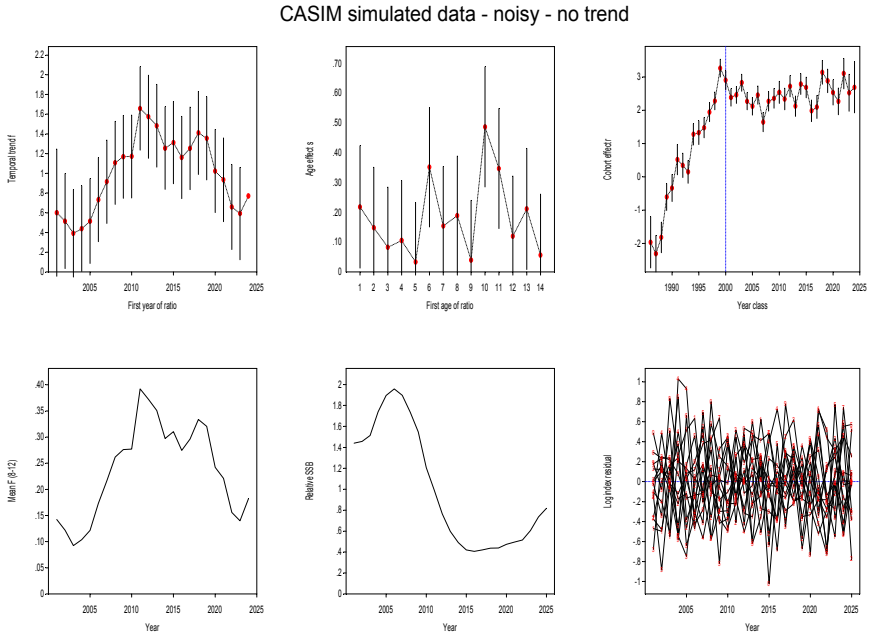


Figure 6.1.3. Comparative output plots for four SURBA model runs with dataset 3, along with the true state. See text for details of dataset and run settings. Upper panel: relative (mean-standardised) SSB. Middle panel: mean *F* (3–8). Lower panel: SURBA summary plot for run 1.

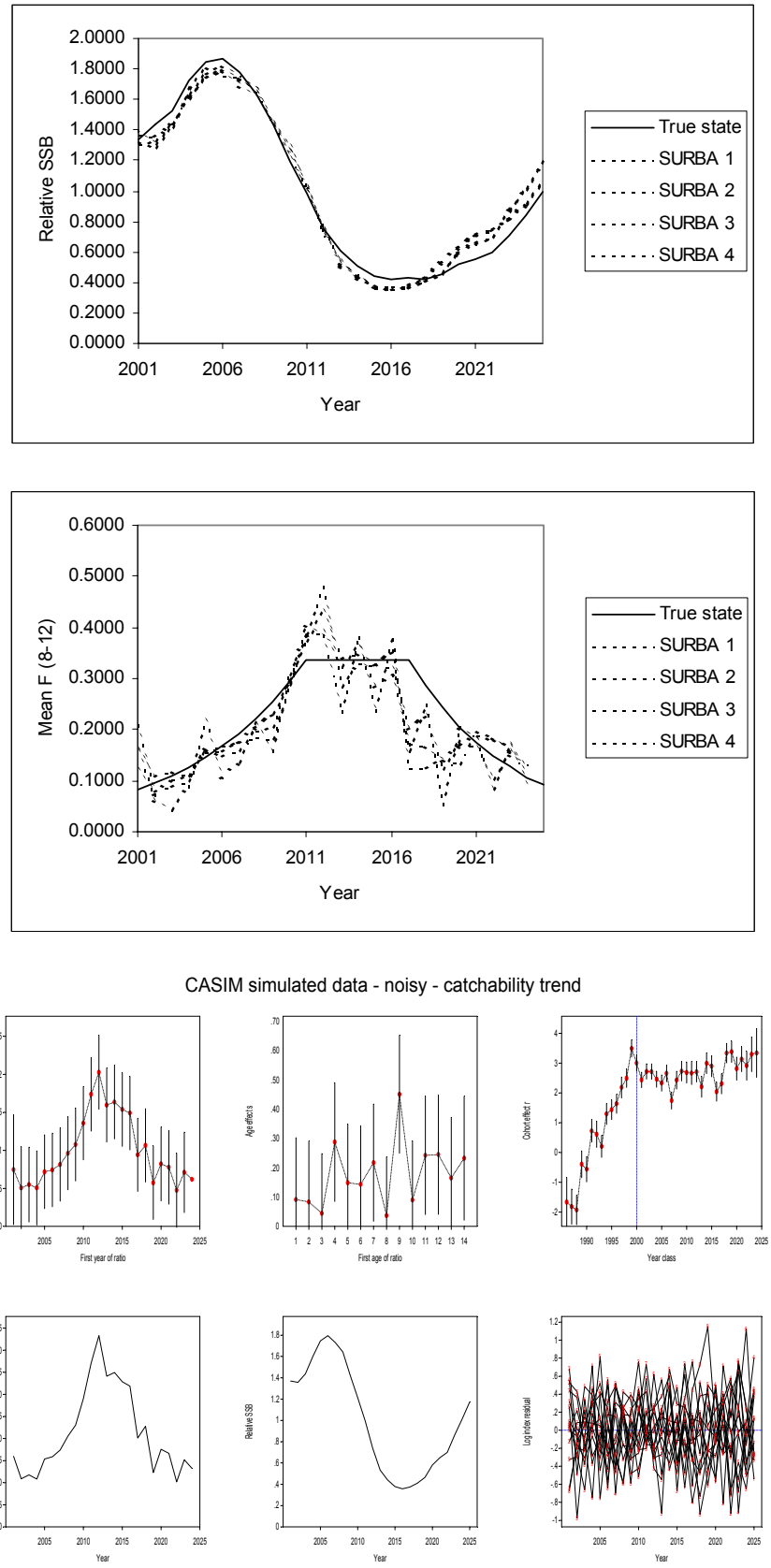


Figure 6.1.4. Comparative output plots for four SURBA model runs with dataset 4, along with the true state. See text for details of dataset and run settings. Upper panel: relative (mean-standardised) SSB. Middle panel: mean *F* (3–8). Lower panel: SURBA summary plot for run 1.

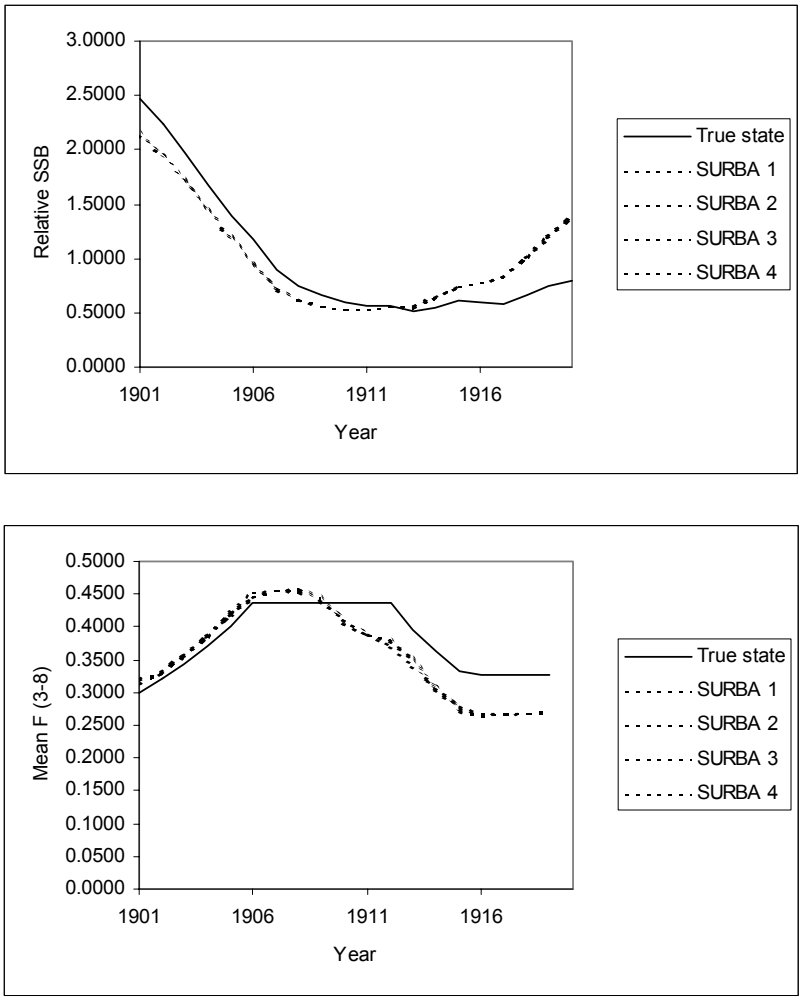


Figure 6.1.5. Comparative output plots for four SURBA model runs with dataset 5, along with the true state. See text for details of dataset and run settings. Upper panel: relative (mean-standardised) SSB. Middle panel: mean *F* (3–8). Lower panel: SURBA summary plot for run 1.

6.2 Analyses of simulated data with CSA

An overview of the Catch-Survey Analysis (CSA) method and of its data needs is provided in Appendix A of this Report. Some aspects of the sensitivity of stock abundance estimates obtained using CSA are addressed in the limited literature dealing with this method. For example, Collie & Kruse (1998) show that estimates of fully-recruited catchability in the survey are negatively correlated with assumed natural mortality: underestimating M translates into over-estimation of catchability and of harvest rate, and under-estimation of stock abundance (i.e. qualitatively the same effect as in VPA). They also show that the parameter, which defines when catch is discounted within the year, acts as a scaling factor: catchability estimates are scaled down and stock size estimates scaled up by the factor exp(M); e.g. if M = 0.2, recruitment or biomass estimates assuming = 0.5 are inflated by 10.5% compared to assuming = 0. They indicate that small errors in allocating animals to either stage result in moderate bias, but that bias may be more serious when staging errors are large and asymmetric. The importance of staging fish correctly is also emphasised by Cadrin *et al.* (1999) and Cadrin (2000). However, the literature is unclear about the effects of errors in setting the catchability ratio s. The Working Group therefore focused on this particular issue. Catchability trends in the tuning fleets were among the concerns raised during the 2001 meeting and in recent assessments, and this issue was also explored.

6.2.1 Application to clean data

The results discussed in this section were obtained using the all-observation error version of CSA on the simulated age- and length-structured data presented in Sections 3.1 and 3.2. Comparisons with true population values were only made for recruits and fully-recruited stock sizes in number, since biomass estimates might also be affected by differences between true and estimated mean weights.

Given that the length composition of the survey catch in each year is provided by the simulator, an attempt was made to estimate the s parameter using the mean length in each stage and the known selection curve for each mesh size. This procedure was successful for the 20-mm mesh survey, in which even the smallest fish are retained, leading to s estimates which are all close to unity, as in the simulation (results not presented). However, it did not work satisfactorily for the 50-mm mesh survey, giving s about 0.58 on average whereas the true average is 0.51. The effect on estimates is shown in Figure 6.2.1. Figure 6.2.2 shows a similar comparison for the 80-mm survey under two s assumptions, where s = 0.05 was obtained with the procedure above whereas s = 0.055 is somewhat larger than the mean of the true annual s in the simulation (0.052). It is apparent from both Figures that CSA can capture the general trends in recruitment and stock abundance, even though it uses rather crude data, but that the absolute values are quite sensitive to the choice of the s parameter which acts as a scaling factor: the higher s is set, the larger the CSA estimates of abundance are. The sensitivity analysis reported in WB2 indicates sensitivity coefficients in the range 1-2 for recruitment estimates, and in the range 2-3 for fully recruited numbers or total biomass in this particular set of simulations. This means that an error of +10% (say) in s causes errors of +10-20% in recruitment and of +20-30% in total biomass, associated with errors of –20-30% in survey q.

With the simulated data, it is possible to know the true catchabilities by stage (ratios of indices divided by population sizes) and therefore the true s. Even though the survey selectivity is kept constant, the true catchabilities of the fully-recruited and true s vary in time due to variations in the age structure of the underlying population. However, CSA runs where the true s by year were input resulted in stock estimates that tended to be biased low. As shown by Jeremy Collie (University of Rhode Island, USA, pers. comm.), this bias can be corrected if the CSA equation is modified to account for changes in the true catchability of the fully-recruited from one year to the next. Unfortunately, this correction would be impossible in real assessments where, by definition, the true q's are unknown.

6.2.2 Application to data with q trend

In principle, one should not expect that surveys would have a trend in efficiency. However, in data limited contexts, users might not have survey data at hand and might have to use commercial catch-per-unit-effort (CPUE) data in which, if no correction can be made in the effort data, catchability trends are likely to be involved. Figure 6.2.3 illustrates the effects upon CSA estimates of a 7% q trend in the tuning data starting in year 10 (high F scenario). As expected, since CSA involves a constant-q assumption, a q trend results in an overly optimistic perception of stock development in recent years when it cannot be accounted for in the processing of indices. A similar, albeit less dramatic, effect was seen with a 4% trend in the low F scenario and it is qualitatively similar to the effect upon VPA estimates. A bizarre behaviour of CSA, however, is that it estimates a lower time-series average q when an increasing trend is simulated compared to the base case without trend.

Figure 6.2.4 shows the results of retrospective CSA runs for the high F and strong trend scenario. Although this data set was chosen to produce retrospective patterns, there is hardly any deviation between successive CSA estimates, except in

the period when the q trend starts (Mohn's rho for these runs is a small –0.032). The figure also shows that, contrary to VPA, CSA does not benefit from a convergence property for past estimates.

The lack of pattern is problematic since it indicates that, contrary to earlier expectations, **retrospective analyses are inadequate to detect the kind of model mis-specification assumed in this data set**.

6.2.3 Application to noisy data

As explained in Section 3.2, this data set involves log-normal observation errors with CV of 20% for catches and of 40% for indices. Moreover, one of the fleets providing indices (fleet 1) has a positive catchability trend. To start with, only data for the second fleet, without trend, were considered and the effects on CSA results are shown in Figure 6.2.5. Here, s was assumed to be 0.1, a value which was arrived at in an heuristic manner by searching for the s value which produced the lowest minimum for the objective function. Even though CSA still captures the overall trend, there are local discrepancies in the sign and magnitude of year to year changes. This is consistent with findings by Cadrin (2000) who concludes that CSA could be misleading, even for trends, when observation error exceeds 30-40%. He points out, however, that such errors do not necessarily increase the bias in CSA estimates (contrary to surplus-production estimates).

Figure 6.2.6 shows a similar comparison for the fleet with trend in addition to noise, assuming an s of 0.08 corresponding to the lowest minimum in the objective function. CSA is often in trouble to correctly identify the relative strength of year classes but roughly captures the trend in fully-recruited stock size, although all estimates are biased low. As indicated in Figure 6.2.7, this bias is somewhat eliminated by assuming a higher s value (0.1). However, there would be no objective procedure to select this value in practice.

6.2.4 Conclusions regarding CSA

A primary virtue of CSA is that it makes limited demand in data compared to age-based methods. It just requires that a recruits stage can be distinguished from older fish; e.g. when a clear cut-off size is visible for the youngest age component in length composition plots.

In trials with *clean* simulated data from a fully age-structured population with 15 ages, CSA performed very well to track relative changes in stock abundance and recruitment. Performance deteriorated when observation errors were added to catches and survey indices, but the overall trends were still reasonably captured. These findings reinforce the conclusions drawn from the 1995 meeting of WGMG (ICES 1995 - Section 3.7) that the method had considerable promise to monitor stock trends.

In view of the robustness of the relative trends in stock abundance indicated by CSA, **this method offers ACFM a chance to give management advice when data are insufficient for VPA**; e.g. when catches in number are known but no time-series of reliable age data is available.

However, absolute CSA estimates are still problematic as they are very sensitive to one key parameter, the ratio of recruits to fully-recruited survey catchabilities, that needs to be set by the user. An objective basis for this choice, given the data commonly available, is missing. A procedure was tried, based on mean lengths by stage and the selection curve of the survey gear, but was not found reliable in all cases for setting s in absolute, although it is acceptable to indicate year-to-year variations in this parameter if one wants to take these into account. The problem of constraining the catchability profile is not unique to CSA, however. Similar requirements can be found in VPA (e.g. age of q plateau for surveys, shrinkage of fishery's F at older age), in separable models (ratio of older age F to F at reference age), and in SURBA (see Section 6.1.1).

If it can be safely assumed that recruits and older fish are equally susceptible to being caught by the survey gear, the *s* ratio is bound to be close to unity and the errors in estimated stock size may be small. This suggests that the method should preferably be used when abundance indices are from small-mesh research surveys. A more compelling reason to be cautious with commercial CPUE data is that, like a number of assessment methods which assume some stationarity in q, CSA will give an over-optimistic perception of stock states when there is an increasing trend in catchability due to gains in fleets' efficiency, that is not accounted for in the compilation of effort data.

To the extent that the *s* ratio can be set close enough to its true value, CSA can compare favourably with VPA. This is all the more remarkable that it only uses a fraction of the extensive information required by VPA. VPA results are influenced by-catch- and CPUE-at-age data over the whole age range, and when the age compositions over the life time of extant cohorts are uncertain, for whichever reason, it is likely that CSA would outperform VPAs using such deficient

data. **Not only can CSA be suggested as a substitute for VPA in such circumstances, but it is also a good candidate for analysing data from a different perspective and verifying VPA-based assessments**.

Although the method can be traced back to the early 1980's, it is still in its infancy if we judge by the amount of literature – about a dozen references, mostly grey, over the past 20 years. More research is obviously needed to better appreciate the properties, good or bad, and limitations of CSA. Some issues to address are suggested below:

- Investigate procedures to estimate the s ratio and/or to formally incorporate uncertainty in this parameter in assessments and advice.

- Run tests on a wider variety of simulations mimicking the dynamics of different real stocks and the characteristics of associated surveys. A clear limitation of this study is that it used exact catch and indices data, and a population with good contrast. Simulations with various levels of noise should also be conducted (see Cadrin 2000).

- Develop appropriate diagnostics to detect violations of model assumptions, and to help objective choice between alternative runs. Diagnostics currently available (including bootstrap and retrospective analyses) do not meet these requirements.

- Elaborate precautionary reference points. Unless one is fortunate that most or all so-called fully recruited fish are indeed mature, CSA cannot sort out SSB from total biomass. This raises the question of how to introduce stock-recruitment processes. Appropriate measures of fishing pressure are required for developing analogues of F-based BRP's; although perhaps harvest rate is good enough.

- Implement software for forecasts. Given the simple population structure, this should be easy even with spreadsheets. CSA provides similar information for predictions as analytical models (except perhaps SSB and S/R-R). A multiple-fleet version can be envisaged, as well as assessments of mesh changes.

Note that for VPA, the resolution of similar problems was the result of decades of developments and thousands of research articles. Compared to that, the attention given to CSA so far has been infinitesimal.

In brief, CSA appears to deserve more resolute consideration within ICES. Consistent with its recommendations in 1995, this meeting of **WGMG encourages its members to explore this method further**.

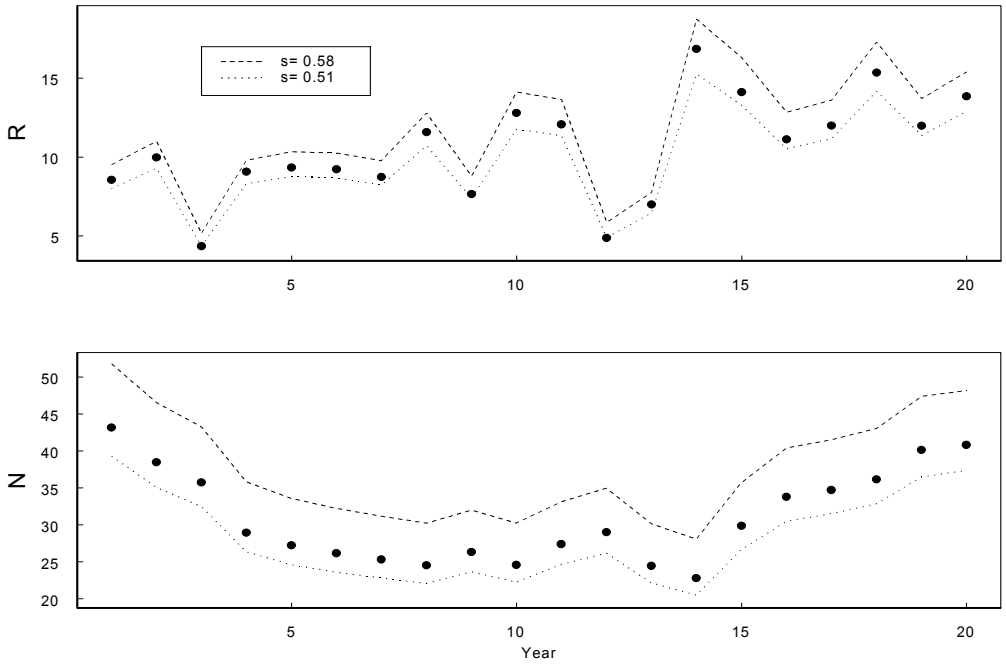


Figure 6.2.1 Comparison of CSA estimates of recruitment and fully-recruited population size under two s (0.51, 0.58) assumptions, and with the true values (solid circles) – 50-mm mesh survey.

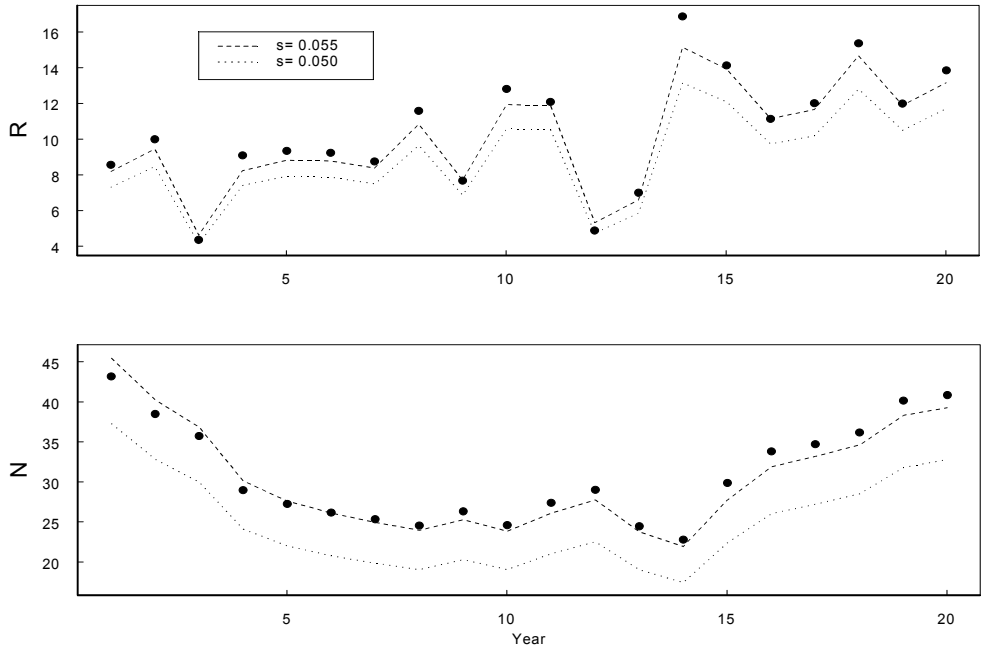


Figure 6.2.2 Comparison of CSA estimates of recruitment and fully-recruited population size under two s (0.050, 0.055) assumptions, and with the true values (solid circles) – 80-mm mesh survey without trend.

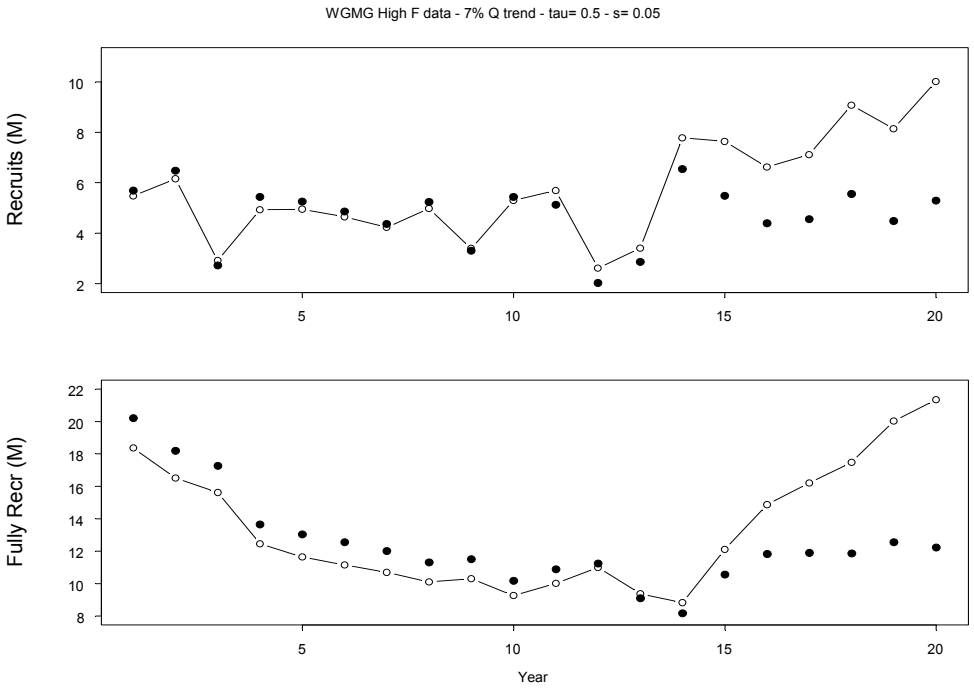


Figure 6.2.3 Comparison of CSA estimates of recruitment and fully-recruited population size with the true values (solid circles) – 80-mm mesh survey with 7% trend, high F scenario – assumed s = 0.05.

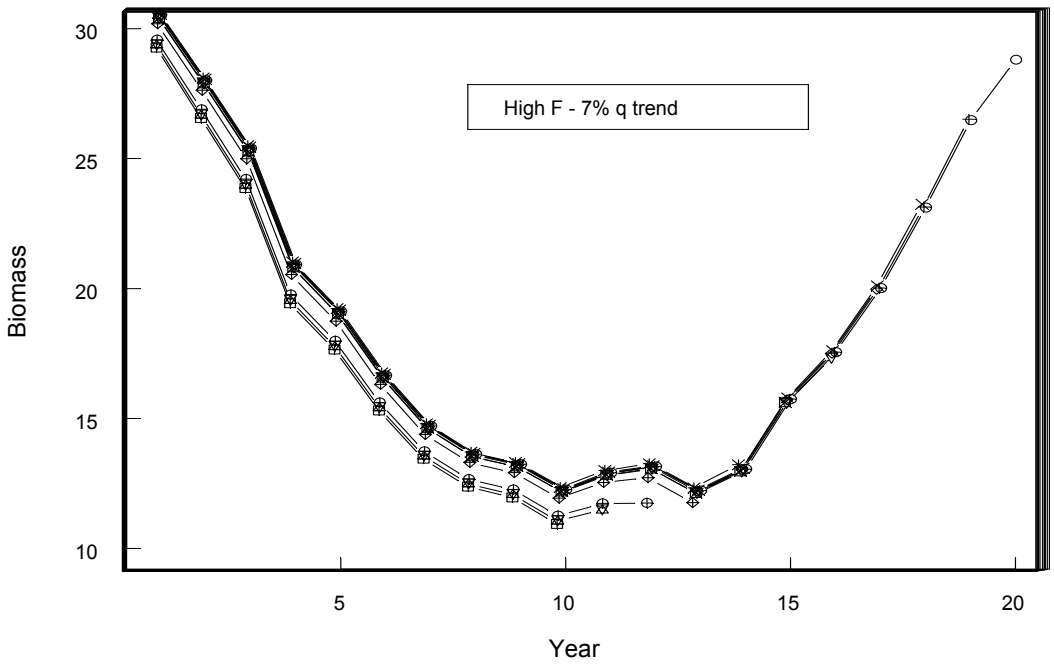


Figure 6.2.4 Retrospective CSA estimates of total biomass – 80-mm mesh survey with 7% trend, high F scenario – assumed s = 0.05.

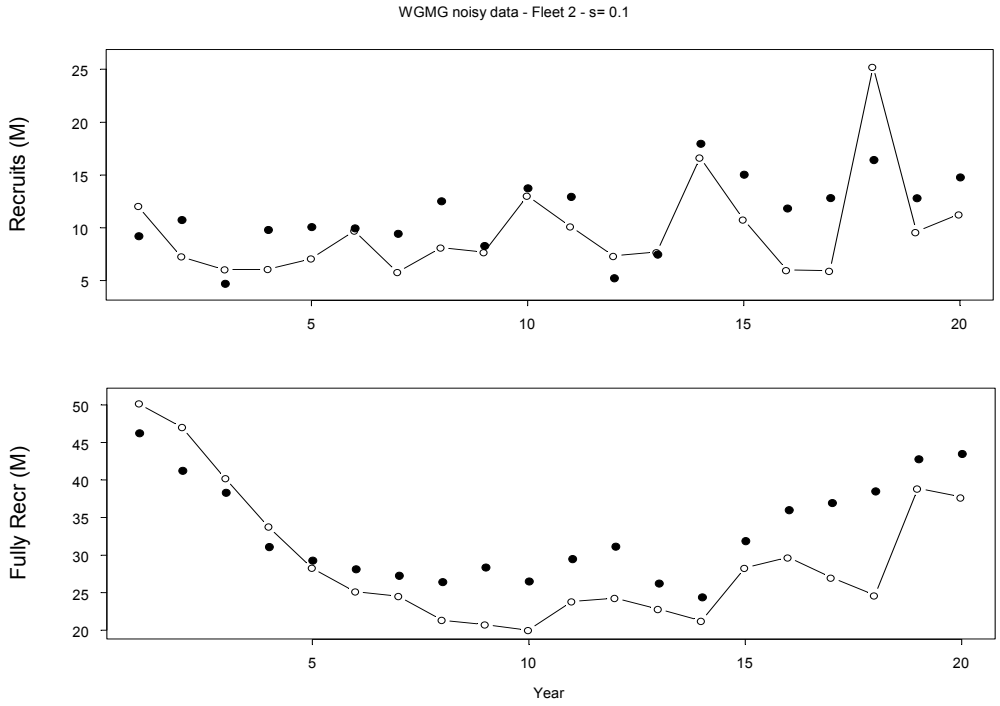


Figure 6.2.5 Comparison of CSA estimates of recruitment and fully-recruited population size with the true values (solid circles) – noisy data, fleet without trend – assumed s = 0.1.

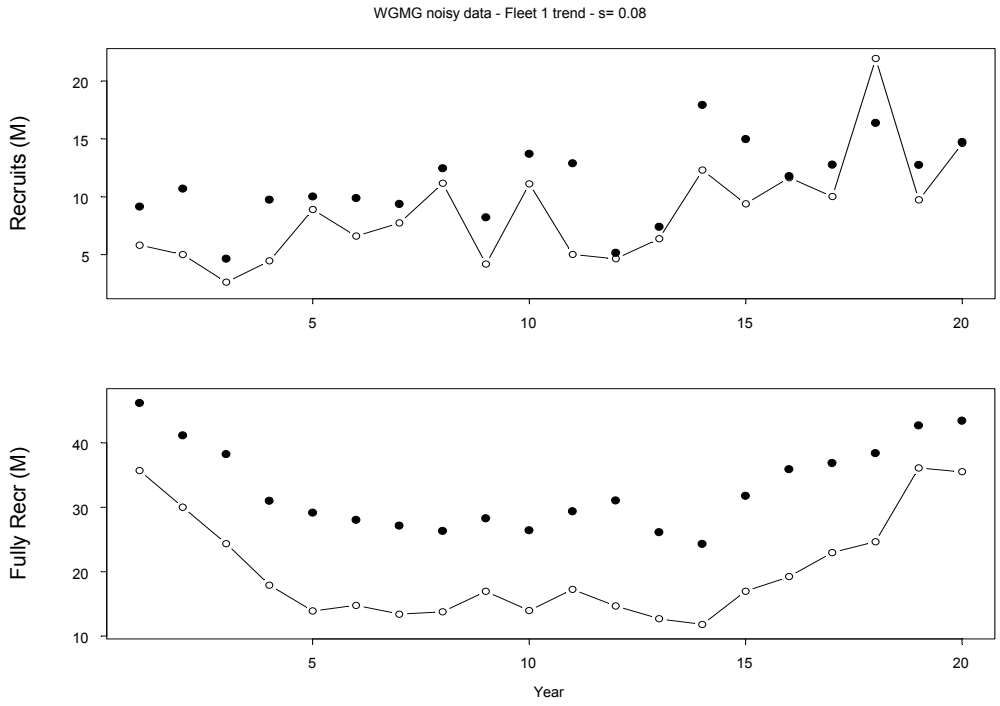


Figure 6.2.6 Comparison of CSA estimates of recruitment and fully-recruited population size with the true values (solid circles) – noisy data, fleet with trend – assumed s = 0.08.

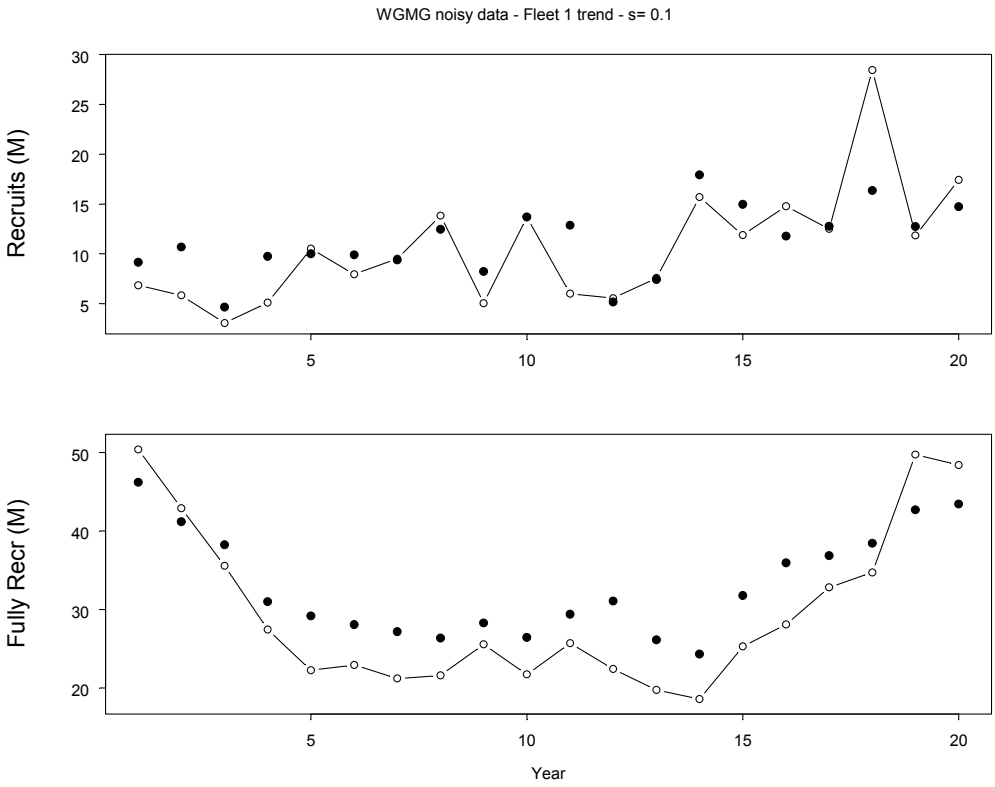


Figure 6.2.7 Comparison of CSA estimates of recruitment and fully-recruited population size with the true values (solid circles) – noisy data, fleet with trend – assumed s = 0.1.

6.3 Detection of inconsistencies in different sources of information – applying Benford's law to fisheries stock assessment

In many sets of numerical data relating to natural phenomena, the first non-zero digits are not uniformly distributed as one might expect but follow a particular logarithmic distribution (Newcomb 1881; Benford 1938; Hill 1995). This distribution, which has become known as Benford's law, has recently been recognised as a powerful and relatively simple tool for pointing suspicion at frauds in commercial accounting data (Nigrini 1996), and for its potential application to other numerical data sets, including scientific notebooks (Matthews 1999). Principally, this is because when humans fabricate numerical data, the resulting frequency of digits rarely follows Benford's law which states that in listings, tables and statistics the digit 1 tends to occur with probability ~ 0.3; much greater than expected from one digit out of 9. Regarding fisheries data this approach could also be a promising tool to look for mistakes in large number data sets such as commercial log-books or official landings.

In a fisheries context, there are indications that the distribution of landed weight over many hauls or trips may follow Benford's law. If one considers a haul as a sample from a distribution, the distribution's parameters and possibly even the appropriate distribution will depend on a range of factors. Hence the landings by haul or trip may be samples from a mixture of distributions (O'Brien *et al*. 2000). Hill (1995) proved that if distributions of numbers are selected at random and a random sample taken from each distribution, the first non-zero digits of the combined sample will converge to Benford's law. Nigrini (2000) states that practice has shown that for a data set to conform to Benford's law it should: i) describe the sizes of similar phenomena; ii) have no pre-set maximum or non-zero minimum; iii) not be assigned numbers (e.g. telephone numbers); and iv) have more small values than large values, in general. A noteable feature of the law is scale invariance. A data set that follows Benford's law will still follow it after multiplying all the values by a constant. This means that the units used to record the data will not affect agreement with the law.

Two papers were tabled at this meeting of WGMG (Maxwell & Dunn WC2; Azevedo WC3) describing investigations of Benford's law to fisheries data. The first (WC2) investigates the applicability of the law to catch data from the English groundfish survey and the second (WC3) considers routinely derived quantities (catch-, survivors- and fishing mortality-at-age) from ICES stock assessments. **Overall, the evidence from these papers for the law to be applied to model-derived quantities (e.g. catch-, survivors- and fishing mortality-at-age) is not as compelling as for fisheries data but it could, nonetheless, still be a useful component of quality assurance checks which screen large numbers of data sets**.

7 APPLICATION OF METHODS TO FISHERIES MANAGEMENT ADVICE

7.1 Medium-term projections

Medium-term projections are a useful component in the provision of advice to fisheries managers. This is particularly true in the current situation of rapidly declining demersal stocks in many ICES areas. These declines are likely to lead to stringent management measures of one kind or another, all with the intention of reducing effort in the fisheries concerned. Projections can be used to indicate levels of exploitation which are likely to lead to sustainable stocks. In addition, any move to fisheries management by multi-annual TACs would have to be supported by a reasonable awareness of what the likely future development of the stock would be under the proposed TAC. These points both imply that medium-term projections are required, and that they should be unbiased with plausible estimates of uncertainty.

The need for improvements to the current medium-term projection methodology has been outlined in Section 4.3.3, along with methods intended to implement such improvements (StockAn – RecAn – MedAn). However, there are several additional points which need to be raised in connection with projections, and these are discussed below.

7.1.1 Drivers of variation

One possible way to improve medium-term projections would be to incorporate process-based models of growth and reproduction. Once the decision to do this has been taken, we must then determine which are the most important factors that would drive changes in growth and reproduction for particular stocks. Although not reflected in the final report (ICES 2003a), the 2002 SGGROMAT meeting spent a considerable amount of time debating the advantages and disadvantages of incorporating biotic or abiotic drivers of change in projections. Biotic drivers are aspects such as stock structure or prey availability, while abiotic drivers are environmental factors such as temperature, salinity and current Flows. The conclusions of SGGROMAT can be summarised as follows.

Both biotic and abiotic factors influence fish stocks, and the degree of influence varies widely across stocks and with time. Therefore, there is no methodological reason why attention should be focused on one set of factors at the expense of the other. However, SGGROMAT concluded that attention should be directed towards biotic factors in the first instance, and the reason for this was purely logistical. It would seem appropriate to concentrate on those factors which will yield the most benefit for the available effort, and for most ICES stocks these will be biotic. There are two reasons for this. Firstly, fisheries management is capable (in theory, at least) of directly affecting biotic factors such as stock structure and prey availability, so it makes sense to provide advice along the lines of "if such-and-such a stock structure is maintained, then such-and-such a population is likely after n years". Fisheries management cannot affect environmental factors and the environment cannot be predicted easily, so projections based on a specified environmental signal may not help very much. Scenario modelling is a possibility, but would have to be presented in a very precise way to as to prevent managers from simply selecting the scenario they prefer. Secondly, the current population structure of the stock will persist for some time at least, since a proportion of the fish alive today will survive into the future. A large year class at age a is likely still to be a large year class at age a+1, and this information can be used in projections.

In conclusion, given the limitations in manpower and resources within ICES that can be devoted to work on projection methodology, the greatest benefit for the most stocks in the least time will be gained by concentrating initially on biotic drivers of change. This is not to say that abiotic drivers will not be important or that they should be ignored. There are already particular examples where a lag or lead-in time in the projection would permit the rapid inclusion of such data, such as Northeast Arctic cod and American plaice, which recruit at ages 3 and 5 respectively, giving at least that many years worth of environmental data to be incorporated. There are also examples where environmentally-driven process-modelling is in an advanced state of development, such as Baltic Sea cod, and the results of such work may influence software development.

7.1.2 Biological projection or management simulation

Projection simulations can be constructed in one of two ways. They can be biological projections, in which the future structure of the population is stochastically simulated under a range of imposed exploitation patterns (which might include agreed harvest control rules or TAC constraints). Alternatively, they can be management simulations, in which the behaviour of the complete fishery management cycle is modelled.. The two types of projection can be ordered hierarchically, with a biological projection underlying a management simulation. The types of simulations we have been discussing here are biological projections. This is because the available data (from fisheries, surveys etc.) do not allow us to determine what future management action is likely in response to population change. These decisions appear to be driven by economic, social and political considerations at least as much as by biological ones, and the

evaluation of these aspects is not within our remit as members of WGMG. However, work on management simulations is proceeding elsewhere (Punt *et al*. BEH1, BEH2, BEH3; EU-MATES project FISH/2001/02). It would be appropriate for WGMG members to bear management simulations in mind when developing biological projections, as that is one use to which such projections could be put.

7.1.3 Testing projection methodology

Projection methodologies need to be subjected to the same specification and sensitivity testing that has been presented above for historical stock assessments. Development work on projections should be accompanied by three types of analysis. Firstly, we should explore the sensitivity of projections to variations in starting populations and assumptions. Secondly, we should run hindcast projections (starting from some point in the past) to see if agreement with the estimated historical population is reasonable. Thirdly, we should evaluate whether the range of abundances in the projections is similar to the range of historically-observed abundances. These analyses would form a *reality check* on the projection procedures, allowing us to test whether the resultant projections are misleading. A start was made in this area at the December 2002 meeting of the Study Group on the Further Development of the Precautionary Approach to Fishery Management [SGPA] (ICES 2003c), during the EU Concerted Action on Estimation of Uncertainty (Patterson *et al*. ICES CM 2000/V:06), at the Working Group on the Assessment of Mackerel, Horse Mackerel, Sardine and Anchovy (ICES CM 2003/ACFM:07), and during the recent meeting of the Study Group on Biological Reference Points for Northeast Arctic Cod [SGBRP] (ICES 2003b). **Further work on projection methodology testing is to be encouraged.**

7.2 Inconsistencies in the North Sea cod short- and medium-term projections

7.2.1 Introduction

Recent studies undertaken to examine the mixed fishery interactions in the North Sea and for North Sea cod quota negotiations have highlighted inconsistencies between the short-term and medium-term projections for the North Sea cod stock. The inconsistencies arise from the data sets used by the Working Group, and model structural assumptions made within the software package WGMTERMC. The divergence of methodological assumptions and data results in very different projections for the probability distributions of SSB and catch. Potentially, the differences could lead to confused messages to stock managers.

7.2.2 Data set assumptions

Different vectors of weight-at-age data were used by the North Sea Working Group (ICES 2002d) for the short and medium-term projection series. In the short-term projection the average weight of the most recent three years of the assessment time-series is used for the projected weight at each age. In the medium-term projection the time period over which the average is derived is the most recent 10 years of assessment data. In general, this approach is used so that the short-term projections are consistent with the most recent changes in the weight-at-age and the medium-term projections incorporate a wider range of uncertainty.

In recent years weight-at-age in the North Sea cod has decreased and the ten-year average used for the medium-term projection is significantly greater that that used in the short-term (Figure 7.2.1). At the ages contributing the greatest proportion of the stock and spawning stock biomass (3-5) the difference in weight is ~10%. The difference in the estimate of spawning stock biomass in the final year, calculated using the two weight-at-age vectors, is 10%. When the percentiles of the first three years of the short-term projection are compared with the medium-term projections, there is a noticeable difference in the projection results.

Given the strong gradient of the stock and recruitment curve over the region, in which the spawning stock is now estimated to lie, the difference in the magnitude of the estimates of spawning stocks results in faster rates of recovery using the optimistic weights taken from the longer time-series.

7.2.3 The WGMTERMC algorithm

WGMTERMC, the software used to carry out medium-term projections, does not allow for intermediate year cuts in fishing mortality resulting from quota reductions. This results in an inconsistency in the starting points for medium-term projections and short-term projections. As discussed above the medium-term projections are more optimistic that the short-term.

It is straightforward to bring the assessment estimates of population abundance forward by one year using a deterministic projection and then to start the stochastic simulations one-year later. However this is not an ideal solution, which would have a stochastic simulation for the first year after the assessment.

7.2.4 Discussion

A consistent approach to short and medium-term projections is required in order to prevent confusion in the advice given to managers. When Working Groups use the two projection methodologies, the results should be equivalent. Differing vectors for the expected value of the input data should be avoided. Rather than using two separate procedures, it would be better to produce a single forecast. The first two or three years of this would be used as the short-term forecast, while the full forecast would be used as the medium-term projection. This could be achieved by allowing recruitment in the early years of the forecast to be determined by VPA survivor's estimates or juvenile surveys, with recruitment models used for later years. The projection output would also need to include catch option tables, to ensure relevance for quota setting.

The forecasting of future weight-at-age values for use as input to medium-term projection programs is being studied by SGGROMAT (ICES 2003a). Models for changes in weight-at-age are being developed that should account for cohort, time-series and age effects. Until the results of the studies are made available it is recommended that a similar approach to that applied to the vector of fishing mortality-at-age, used as input to short-term forecasts, is applied to weight-at-age.

If there are no obvious cohort effects in the time-series, noise in the weight-at-age data can be smoothed by taking an average over a period of the most recent years of data. If a trend in the most recent years is apparent, scaling to the final year would be appropriate. This approach would help to compensate for time-series effects that can induce an over- or under-estimation bias to estimates of potential catch and spawning stock biomass (Darby WA2 at ICES 2002a, Appendix B). If cohort effects are apparent in the time-series more complex models are required, similar to those described by SGPRISM (ICES 2002e).

Where weights-at-age are considered to be historically more variable, coefficients of variation derived from longer time periods will be more appropriate in medium-term projections. These should be applied to the expected values from the smoothed, shorter, time-series calculated over the most recent years. Using such an approach will bring consistency to the short- and medium-term forecasts.

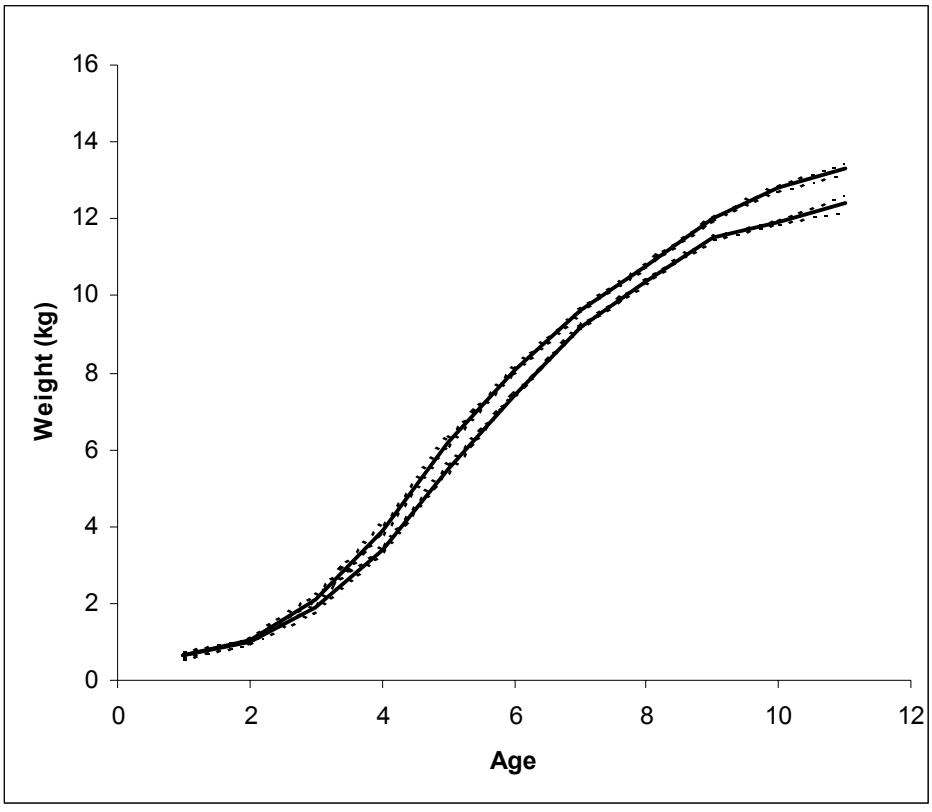


Figure 7.2.1 The weight-at-age used in the short-term (lower curve) and medium-term (upper curve) projections as used at the 2002 North Sea assessment Working Group. In both series the solid line illustrates the mean weight and the broken lines one standard deviation intervals.

7.3 Recruitment of Northeast Arctic cod

7.3.1 Recruitment models with spawning stock structure

A simplified variant of a model of cod recruitment has been developed to cover a fairly long historical period (1979-2000; see WE1). It is based on the Ricker model, to which is added the assumption that different ages of mature fish have different contributions to recruitment. We suppose that the number of mature fish at age *t* in year *y* participating in spawning for the k^{th} time is given by

$$N^{p}_{t,k,y} = N_{t,y} \cdot (\delta_{t-k+1,y-k+1} - \delta_{t-k,y-k}), \qquad (7.3.1)$$

where $N_{t,y}$ is the total abundance of fish and $\delta_{t,y}$ is the proportion of mature fish at age t in year y (see Table 7.3.1).

Applying this approach to Northeast Arctic cod reveals three structural groups with different qualitative compositions of spawners:

1. early maturing spawners having an age of first spawning of 4-9 years and participating in spawning 1-3 times (identified by $i = 1$);

2. early maturing spawners having an age of first spawning of 4-9 years and participating in spawning 4-9 times (identified by $i = 2$); and

3. late maturing spawners having an age of first spawning of 10-12 years (identified by $i = 3$).

The total abundances of fish in each structural group are used as indices of the reproductive potential of the population. We use the following stock-recruitment relationship to do this:

$$\widehat{R}_j = \sum_{i=1}^{3} \alpha_i \cdot N_{i,j} \cdot \exp\left(-\beta_i \cdot N_{i,j}\right) \qquad (7.3.2)$$

where *i* is the group identity number, $N_{i,j}$ is the total abundance of mature fish in each group in the j^{th} year, and α_i and β_i are group-specific Ricker parameters to be estimated..

Much recent work (e.g., SGGROMAT, ICES 2003a) has concentrated on generating improved indices of reproductive potential using total potential fecundity or hepatosomatic or morphometric indicators of condition. Tretyak (WE1) has shown that, while these are necessary, they are unlikely to sufficient descriptors of reproductive potential, and that stock-structure considerations of the type proposed in Equation 7.3.2 will also be required (see Figure 7.3.1).

There are at least three conditions which must be met before the determination of the number and identification components of the spawners is possible. Firstly, recruitment in the time interval considered should be a stationary random function. That is, the expectation and variance of recruitment should be constant in this interval. Secondly, the correct underlying stock-recruitment relationship should be chosen. The Ricker model was used in the Northeast Arctic cod case study because it is thought to encapsulate perceived density-dependent mortality in that stock. Thirdly, the mean age of mature fish should not show pronounced downward or upward trends. Such trends are caused by variations in maturity ogives and lead to the appearance of inter-annual qualitative heterogeneity of mature fish and nonstationary of recruitment. Such variations can be caused by intensive fishing mortality, and by the influence of the positive inverse feedback which arises in the mechanism of homeostasis of population abundance. This feedback magnifies the effect of decreasing density on the process of structural-composition change in mature fish. Further increases in fishing intensity would make this process irreversible, and lead the spawning population further and further away from its initial structural composition.

This method of modelling the effects of stock structure on recruitment looks promising, and it would be useful to apply it to different stocks. The implications of the modified Ricker model for management would best be explored through projection simulations. While not directly within the remit of WGMG, this activity would be relevant to the work of SGGROMAT and would benefit from being presented there.

Table 7.3.1 Proposed stock structure of cod spawners.

Age t, years	Abundance of mature fish at age t	Abundance mature fish at age t, spawning the k^{th} time										
		k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11
3	N^p_3	$N^p_{3,1}$										
4	N^p_4	$N^p_{4,1}$										
5	N^p_5	$N^p_{5,1}$										
6	N^p_6	$N^p_{6,1}$										
7	N^p_7	$N^p_{7,1}$	$N^p_{7,2}$	$N^p_{7,3}$	$N^p_{7,4}$	$N^p_{7,5}$						
8	N^p_8	$N^p_{8,1}$	$N^p_{8,2}$	$N^p_{8,3}$	$N^p_{8,4}$	$N^p_{8,5}$	$N^p_{8,6}$					
9	N^p_9	$N^p_{9,1}$	$N^p_{9,2}$	$N^p_{9,3}$	$N^p_{9,4}$	$N^p_{9,5}$	$N^p_{9,6}$	$N^p_{9,7}$				
10	N^p_{10}	$N^p_{10,1}$	$N^p_{10,2}$	$N^p_{10,3}$	$N^p_{10,4}$	$N^p_{10,5}$	$N^p_{10,6}$	$N^p_{10,7}$	$N^p_{10,8}$			
11	N^p_{11}			$N^p_{11,3}$	$N^p_{11,4}$	$N^p_{11,5}$	$N^p_{11,6}$	$N^p_{11,7}$	$N^p_{11,8}$	$N^p_{11,9}$		
12	N^p_{12}		$N^p_{12,2}$	$N^p_{12,3}$	$N^p_{12,4}$	$N^p_{12,5}$	$N^p_{12,6}$	$N^p_{12,7}$	$N^p_{12,8}$	$N^p_{12,9}$	$N^p_{12,10}$	
gr+	N^p_{gr+}			$N^p_{gr+,3}$	$N^p_{gr+,4}$	$N^p_{gr+,5}$	$N^p_{gr+,6}$	$N^p_{gr+,7}$	$N^p_{gr+,8}$	$N^p_{gr+,9}$	$N^p_{gr+,10}$	$N^p_{gr+,11}$

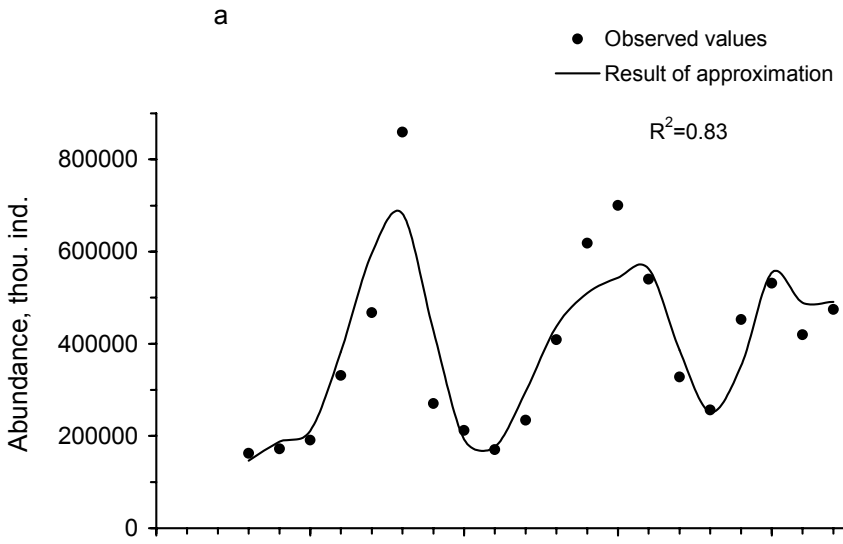


Figure 7.3.1. Observed and fitted recruitment, using modified Ricker model. The *x*-axis denotes the 1976–1997 year classes

8 SPECIAL REQUEST ON BLUE WHITING AND NORWEGIAN SPRING SPAWNING HERRING

8.1 Background of the problem

During the coastal state meeting on blue whiting in Oslo on 7-8 November 2002, the assessment and management advice given by ICES was presented and discussed. Assessments and predictions for blue whiting and Norwegian Spring Spawning herring based on the ISVPA model (Vasilyev 2001, WF3) was presented in the Oslo meeting and was compared to the output from the models presented by ACFM in 2002 (AMCI and SeaStar, respectively). The results with ISVPA gave substantially different estimates of the stock size of both stocks. It is known that both these models are used by ICES in *The Northern Pelagic and Blue Whiting Working Group* along with other available tools.

The parties have requested ICES to "evaluate the two assessment models with respect to Norwegian Spring Spawning herring" and to "extend these evaluations to also include assessment of blue whiting" (c.f. Section 1.4).

Unfortunately, the evaluation of the SeaStar model could not be carried out during this working group meeting, because the main author of the model could not be present, and therefore the group was not able to evaluate the Norwegian Spring Spawning herring assessment. The evaluation of the blue whiting assessment is presented below in the Section 8.4.

In Section 8.2, a general description will be presented of the models that were evaluated for blue whiting. In Section 8.3 results of the model evaluation using simulated data are presented (see Section 3.2 for a description of the simulated data). Section 8.4 presents the results of the evaluation using the blue whiting data from the most recent ICES Northern Pelagic and Blue Whiting Working Group (ICES 2003b). General conclusions to the analysis are presented in Section 8.5.

8.2 General descriptions of models investigated

The evaluations of different assessments models that could be applied to blue whiting are described in Section 8.3 using simulated data and Section 8.4 using the blue whiting data. The following models have been included in the testing process:

- AMCI
- ISVPA
- ICA
- XSA
- CADAPT (for exploratory purposes only)

The general descriptions of models used are presented in Sections 4.2.1 (AMCI), 4.2.2 (ISVPA) and 4.3.4 (CADAPT). An overview of methods is presented in Table 4.4.1. AMCI, ICA and ISVPA all belong to the general class of separable models. AMCI fits a separable pattern to the catch data but allows for small changes in selection between years. ICA fits a separable pattern over a usually shorter period and applies a traditional VPA over the earlier years of the catch data. ISVPA can be used both in a strictly separable manner ("effort controlled") or in a VPA-type manner ("catch controlled") or a mix of the two. In addition ISVPA allows additional constraints to be placed on the characteristics of either the catch residual matrix or the matrix of residuals between values of F calculated from the catch-at-age and estimates of abundance, and their "theoretical" values, based on a separability assumption.

XSA and CADAPT are implementations of calibrated VPA models which both take the catches to be true instead of the modelled catches as in a separable models. XSA is widely used to assess a broad range of stocks in the ICES area and CADAPT is in an early stage of development.

An extensive number of graphical outputs have been produced from the assessment models investigated and in contrast to the other Sections of this report, the figures have not been interspersed within the text of this Section 8 but have been collated at the end of Section 8 for ease of reference.

8.3 Results of stock assessments on simulated data

The simulated datasets are described in Section 3.2 and include random noise on the catch-at-age data (CV=20%) and random noise on the survey data (CV=40%). In addition, one of the surveys included a trend in catchability of 4% per year from the middle of the simulation period. In general, the level of fishing mortality over the whole period was

relatively low (i.e. between 0.1 and 0.35). The simulated catch data are shown in Figure 8.3.1 and the survey data in 8.3.2.

Figures 8.3.3, 8.3.4 and 8.3.5 compare estimates of SSB, F_{8-12} and recruitment at age 1, respectively, as estimated by the various assessment methods (AMCI, ISVPA, ICA, XSA and CADAPT).

8.3.1 Model settings and general results

AMCI

Only the noisy data set was explored. AMCI was set up as one normally would do for a first attempt with a new data set, without knowing the true data. For the first year, a flat selection from age 11 to agee 15 was assumed. For the other years, the selection was allowed to vary with a small gain of 0.1, since there were no indications in inspection of the data and the log catch residuals gave no indication aof marked shifts in selection. Yearly fishing mortalities, recruitments and stock numbers-at-age were estimated parameters, except for the recruitment in the last year, which was fixed at a guesstimate of 15 000. The caches were assumed to be evenly distributed over the year. Natural mortality was set to 0.2. Survey catchabilities at age were estimated, assuming they were constant over time. Weighting was the standard one, where each whole survey gets the same weight as each year of catch data, which implies that most weight is given to the catches.

The objective function to be minimised was:

$$S = w_0 \Sigma_{a=1,15} \Sigma_{y=2001,2025} \ln(Cobs(a,y)/Cmod(a.y))$$

$$+ \Sigma_{fl=1,2} * w_{fl} \Sigma_{a=1,15} \Sigma_{y=2001,2025} \ln(Iobs(y,a,fl)/Imod(y,a,fl))$$

where $w_0=1/15$ and $w_1 = w_2 =1/(15*25)$

Three runs were made, with both surveys and only one of the surveys respectively.

Largely, the results were in accordance with the true population. The selection had a dip around age 8 in the early years, which probably was a response to noise in the data. In the late years, there was a slight rising trend in selection towards old ages. The residuals were generally scattered, although there was some trend in the residuals in the survey with a trend in catchability (Figures 8.3.7 and 8.3.8).

ISVPA

For ISVPA stock assessment was undertaken using catch-at-age only and by including the auxilliary information. In Figure 8.3.10 it can be seen that the objective function shows a clear minimum for the catch data only, but when the two surveys are included, the minimum is much less pronounced.

The objective function was the same as that for AMCI given above, except that equal weights were taken; namely,

$$w_0 = w_1 = w_2 = 1.$$

An attempt was made to estimate natural mortality within the model formulation. Results of the minimization profiles in terms of fishing mortality and natural mortality are shown in Figure 8.3.11. Natural mortality estimates were between 0.10 and 0.20 depending on the data used.

ICA

ICA was initialized with a separable period of 6 years, a reference age of 4 and the selection at the oldest true age to be equal to the selection at the reference age. Three separate run were performed: one with both simulated fleets incorporated, one with only fleet 1 (with q trend) and one with only fleet 2 (without q trend). The catchability models were all assumed to be linear and equal weighting was applied to each survey fleet with full correlation between ages.

The objective function used for the ICA runs was:

$$\sum_{a=1,y=2020}^{a=14,y=2025} \lambda_a \left(\ln(\hat{C}_{a,y}) - \ln(C_{a,y}) \right)^2 +$$
$$\sum_{a=1,y=2001}^{a=14,y=2025} \lambda_{a,fleet1} \left(\ln(q_{a,fleet1} \cdot \hat{N}_{a,y}) - \ln(Fleet1_{a,y}) \right)^2 +$$
$$\sum_{a=1,y=2001}^{a=14,y=2025} \lambda_{a,fleet2} \left(\ln(q_{a,fleet2} \cdot \hat{N}_{a,y}) - \ln(Fleet2_{a,y}) \right)^2$$

XSA

XSA was applied without time-taper on the survey indices. This option was considered appropriated in order to detect the trends in catchability. Plots of CPUE residuals against stock numbers showed no evidence of a relationship for any fleet, so no power model was used for catchability. No shrinkage was used in order to be able to detect the signal in the surveys most clearly. Three different runs were carried out. Differences are related only to the tuning surveys used in the assessment.

In the first run, the two survey fleets both incorporated. Log catchability residuals for fleet 1 (with q trend) gave a slightly increasing trend in log-catchability residuals in the last years of the series. For all ages, residuals were high and with some year effects in these last year. A slight decreasing trend could be also identified during the period 2016 to 2018. For the second survey residuals are quite small and no apparent trend was observed although a year effect occurred in the last year of the series. A retrospective analysis was back-calculated over all the time period and shows a considerable underestimation of mean fishing mortalities (Figure 8.3.6a), in contrast to the results of the retrospective analysis produced by AMCI (Figure 8.3.6b).

For the single fleet runs, no general trend in log-catchability residuals were apparent for the second survey (without q trend) and for all ages the residuals were low. The SSB estimates at the beginning of the time-series showed relatively large disagreement from the true values. Also for the first survey, no apparent trend in log-catchability residuals were observed. In increase in catchability was interpreted by the model as an increase in stock size and therefore the SSB estimates in the final years departed from the true values.

CADAPT

CADAPT was run with a proportional catchability model assuming a plateu for ages 5 and older. Two runs were made, one for each data set. Results are shown in Figures 8.3.3-8.3.5 and results from retrospective analysis for recruitment, SSB and mean $F_{8\text{-}12}$ shown in Fig. 8.3.9 for key management parameters.

For simulated data:

$$o_{cadapt\ for\ simulated\ data} = \sum_{y=2001}^{2025} \sum_{a=1}^{15} \left(\frac{(\ln(I_{ya}) - \ln(q_a) - \ln(N_{ya}))^2}{2 * \sigma_a^2} + \ln(\sigma_a) \right)$$

where I_{ya} and N_{ya} are numbers- and index-at-age, respectively, q_a factor of proportionality between numbers- and index-at-age, constant for a>=5, and, σ_a^2 is the mean square of log age group residuals, used for inverse variance weighting residuals in optimization, given by:

$$\sigma_a^2 = \frac{1}{25} \sum_{y=2001}^{2025} \left(\ln(I_{ya}) - \ln(q_a) - \ln(N_{ay}) \right)^2$$

In this Section we compare the various assessment methods with respect to how well they are able to estimate the true state of the stock. Because the stock is artificially generated, the true state of the stock is known precisely. A detailed description of the generated data set is given in Section 3.2 and the input data are shown in Figure 8.3.1 and 8.3.2. Here we refer to the 'noisy' data (random error added to the survey and the catch-at-age data) with low fishing mortality. Data of two survey fleets are available, one of which has an increase in catchability of 4% per year over the last 10 years. Several types of assessment methods were applied to the simulated stock, namely ICA, XSA, ISVPA, AMCI, and CADAPT, with various data sources. The model settings that were used are described in Section 8.3.1. The assessments were run using the catch-at-age data and data from one or both surveys. In the case of the ISVPA, assessments have

been run with catch-at-age data only, the fleet data without trend only, the fleet data with the catchability trend only, and all three data sources with equal weights respectively. Two versions of ISVPA are run, one in which natural mortality is a fixed (given) value (M=0.2), and one in which natural mortality is estimated. Because the true state of the stock is known it is possible to evaluate the performance of each of the methods and the use of a particular (combination of) data source(s).

Figure 8.3.3 displays the time-series of SSB. ICA appears to over-estimate SSB most strongly, in particular in the first half of the time-series, where an over-estimation of 75% is seen. In the second half of the time-series the assessment with the fleet without trend performs quite well, whereas the use of the fleet with the q trend leads to overestimation of SSB. XSA with the fleet without the trend performs very well too, although in the first half of the time-series the use of this fleet leads to overestimation. ISVPA performs better when M is estimated than when a specified value is given. In the case of M given the method performs equally well whether only catch-at-age data are used or only the fleet without the trend. In the case of estimated M the method performs best when the fleet without trend is used. Clearly, AMCI performs very well, regardless which (combination of) data source(s) is used. CADAPT performs very well too, especially when the fleet without trend is used.

In Figure 8.3.4 the time-series of F_{8-12} are displayed. All methods detect a dip in the plateau of high fishing mortality in the middle of the time-series, which in truth does not exist. Apparently, this particular realisation of 'noise' given to the data, results in this pattern. In this respect AMCI deals with the data best. Excepting the dip in the middle, XSA and CADAPT replicate the true fishing mortality very well, although XSA with the fleet without the trend underestimates F a bit and CADAPT with the fleet without trend overestimates F a bit, especially in the later part of the time-series. ICA consistently underestimates F to a large extent, although at the end of the time-series the assessment with the fleet without the trend approximates the true F. The same is true for the ISVPA with given M. ISVPA with estimated M performs quite well if all data sources or the fleet with the trend is used. ISVPA estimating M using the fleet without trend overestimates F in the later part of the time-series. ISVPA estimating M with catch-at-age data only overestimates F to a large extent over the whole period.

Figure 8.3.5 displays the time-series of recruits (numbers of age 1). ICA with the fleet without the trend estimates recruitment very well over the whole time-series, whereas if the fleet with the trend is used the estimates diverge upward. The same is true, but to a lesser extent, for XSA and CADAPT. ISVPA with given M performs well as long as the fleet with the trend is not used. ISVPA with estimated M performs well if the fleet without the trend or if all data sources together are used. AMCI over-estimates recruitment in all cases.

8.4 Results of stock assessments on blue whiting data

8.4.1 Model settings

AMCI

The same options were applied as used by the NPBW working group in last years assessment. Also, the same data were used. The assessment was run with data from 1981 – 2001, plus the data from the Norwegian spawning acoustic survey in 2002.

Selection at age was allowed to vary with a gain factor of 0.5 for age 0, 0.2 for age 1, and 0.1 for ages 2-9. The fishing mortality of the plus group was set to be the average of F at ages 8 and 9. In the first 4 years, the selection was assumed to be fixed. Fishing mortality and recruitment were estimated parameters for all years 1981 - 2001, except for the recruitment in 2001, which was fixed at approximately the historical geometric mean. The fishing mortality and selection in 2002 were assumed to be equal to those in 2001. Stock numbers-at-age in the first year were estimated parameters.

Survey catchabilities were estimated for each age, but were assumed to be constant over years. For the Norwegian and Russian acoustic surveys on the spawning grounds, separate catchabilities were estimated for the early and late periods.

The objective function was the weighted sum of squared log residuals, with equal weighting given to each source of data. This is the same as used by the NPBW WG. Implicitly, AMCI will give the same weight to each year of catch data as to each whole survey. Thus, most weight is given to the catches. Catches-at-age 0 were given a weight of 0.1 and catches-at-age 1 of 0.5.

The objective function to be minimised was:

$S = w_0 \Sigma_{a=0,10} \Sigma_{y=1981,2001} w(a)*\ln(Cobs(a,y)/Cmod(a.y))$

$+ \Sigma_{fl=1,4} * w_{fl} \Sigma_{a=0,10} \Sigma_{y=1981,2002} \ln(Iobs(y,a,fl)/Imod(y,a,fl))$

where w_0=1/11 and each wfl was 1/the number of valid observations

and w(0) = 0.1, w(1)=0.5, w(a) =1 for a=2-10

ISVPA

ISVPA was run in the version most traditional to separable models: residuals in cohort part of the model were attributed to errors in catch-at-age data. Restriction of zero age- and year- sums in residuals (so-called "unbiasedness") was applied to residuals in the separable representation of the fishing mortality coefficients.

The objective function was the following:

$S = w_0 \Sigma_{a=1,10} \Sigma_{y=1981,2001} *\ln(Cobs(a,y)/Cmod(a.y))$

$+ \Sigma_{fl=1,4} * w_{fl} \Sigma_{a=1,10} \Sigma_{y=1981,2002} \ln(Iobs(y,a,fl)/Imod(y,a,fl))$

$\Sigma_{SSB=1,2} * w_{SSB} \Sigma_{y=1981,2002} \ln(SSBobs(y,a,fl)/SSBmod(y,a,fl))$

where $w_0 = w_{fl} = w_{SSB}$.

ICA

The ICA model was set up in a spreadsheet in order to be able to explore the behaviour of the model given different terminal fishing mortalities. The model was set up slightly differently from the WGNPBW 2002 settings. Notably the separable period was extended to 10 years (WG used 4 years) and the acoustic surveys were taken without a break in the time-series (WG used a split in 1989). The reasons for these deviations were give by the limitations of the spreadsheet implementation of ICA and the number of parameters which were estimatable in Excel.

The objective function used for the full model is given below.

$$\sum_{a=1,y=1992}^{a=9,y=2001} \lambda_a \left(\ln(\hat{C}_{a,y}) - \ln(C_{a,y}) \right)^2 +$$
$$\sum_{a=2,y=1983}^{a=8,y=2001} \lambda_{a,NorSpa} \left(\ln(q_{a,NorSpa} \cdot \hat{N}_{a,y}) - \ln(NorSpa_{a,y}) \right)^2 +$$
$$\sum_{a=3,y=1982}^{a=8,y=1996} \lambda_{a,RusSpa} \left(\ln(q_{a,RusSpa} \cdot \hat{N}_{a,y}) - \ln(RusSpa_{a,y}) \right)^2 +$$
$$\sum_{a=1,y=1983}^{a=6,y=2001} \lambda_{a,SpaTra} \left(\ln(q_{a,SpaTra} \cdot \hat{N}_{a,y}) - \ln(SpaTra_{a,y}) \right)^2 +$$
$$\sum_{a=1,y=1982}^{a=7,y=2001} \lambda_{a,NorSea} \left(\ln(q_{a,NorSea} \cdot \hat{N}_{a,y}) - \ln(NorSea_{a,y}) \right)^2$$

The spreadsheet models are available on the WGMG 2003 folder (W:\rmc\wgmg\2003\personal\blue whiting\ica_spreadsheet)

XSA

The Extended Survivors Analysis (XSA Shepherd 1999, Darby and Flatman 1994) was fitted to the catch-at-age and natural mortality data sets for blue whiting datasets described previously. The model was calibrated using four Norwegian research survey data series (Norway Spawning Area acoustic 1981-90 and 1991-2002, Norwegian Sea acoustic 1981-90, 1991-2001), two Russian surveys (Russian Spawning Area Acoustic 1982-91 and 1992-1996), and a Spanish pair trawlers cpue index 1983 – 2001. Tuning data was only available for ages 1 – 8.

After examination of the diagnstics from exploratory runs of the XSA model the following specification for the model structure was used for the comparative runs.

- Catch data for the years 1981 - 2001 at ages 0 to 8+.
- Tapered time weighting not applied.
- Catchability independent of stock size for all ages.
- Catchability independent of age for ages greater than 5.
- Survivor estimates in the final year not shrunk towards the mean fishing mortality.
- Survivor estimates at the oldest age shrunk towards the mean fishing mortality of the 3 oldest ages; s.e. of the mean to which the estimates are shrunk 0.5.
- Minimum standard error for population estimates derived from each fleet 0.3.

In order to compare the trends in the estimated stock dynamics XSA assessment models were fitted fitted to each cpue series individually. Residual plots were examined for departures from the assumption of constant log catchability, that is strong trends, age or cohort effects. The survey information at all ages in all fleets is noisy with minimum c.v of 40%. All of the acoustic surveys have year effects which illustrate inconsistencies with the population dynamics reconstructed from the catch-at-age data (Figure 8.4.25). Whilst the recent acoustic series have strong year effects they do not have strong trends and were therefore retained within the fitted XSA model. The Spanish cpue index has strong trends in catchability with opposite patterns in the young and oldest ages indicating a change in selection towards the older ages. The catchability is clearly not constant in time and the series was removed from the calibration data used to fit the assessment model.

A review of the estimates of fishing mortality-at-age established that the fishing mortality-at-age 7 is very high relative to the preceding ages. This outlier results in the average F at that age being raised in the last two years of the assessment by a values that appears to be an outlier. An examination of the catch-at-age data shows that there is a marked decrease in the catch-at-ages 7 – 9 in 2001. This change in selection may have resulted from a random year effect or a change in fishery practices. In order to examine the trends in fishing mortality without the potential bias from this potential outlier the average fishing mortality in the final year was calculated over the age range 3 – 6. The fitted time-series of fishing mortality (3-7, apart from year 2001 which is 3 - 6), recruitment and spawning stock biomass estimates are presented in Figure 8.4.27.

CADAPT

CADAPT was run with proportional catchability on all age groups, tuned with age group indices for the age range common for the Norwegian spawners survey and Norwegian Sea summer survey. The plusgroup in the catch-at-age data was treated as age group 10, and alternative runs were made using only ages 1-9. These explorations should only be considered as indicative as the method is still under development.

Choice of year and age range in the surveys considered may not have been the ideal for a 'fair' comparison of the two acoustic fleets and the performance of CADAPT. It would have been better to have started both tuning series in the same year and have an index after the last catch for both survey series. Furthermore, the blue whiting data may not be the best test data for a new method like CADAPT.

For blue whiting data:

$$o_{cadapt\ for\ blue\ whiting\ data} = \sum_{y=surveyStart}^{surveyEnd} \sum_{a=2}^{7} \left(\frac{(\ln(I_{ya}) - \ln(q_a) - \ln(N_{ya}))^2}{2 * \sigma_a^2} + \ln(\sigma_a) \right)$$

with same notation as in objective function for simulated data, q_a: estimated for age group indices 2-7, for the summer acoustic survey in the Norwegian Sea (*norSea*): *surveyStart*=1981, *surveyEnd*=2001, for the Norwegian acoustic

survey on spawners (*norSpa*): *surveyStart*=1986, *surveyEnd*=2002, and as there were years missing from the survey series the mean square of log age group residuals was calculated accordingly:

$$\sigma_a^2 = \frac{1}{nObsSurvey} \sum_{y=surveyStart}^{surveyEnd} \left(\ln(I_{ya}) - \ln(q_a) - \ln(N_{ya}) \right)^2$$

where *nObsSurvey* was 18 and 16 for the *norSea* and *norSpa* surveys, respectively.

8.4.2 General diagnostics

The blue whiting data were explored as outlined in Section 4.5.

Catches-at-age by year class are shown on a logarithmic scale in Figure. 8.4.1. Most year classes decline in the catches from ages 2-3 onwards with a slightly increasing slope, indicating an increasing fishing mortality towards older ages. The pattern at ages 1 and 2 is more irregular. Some year classes have very low catches at young age, like the 1992 year class. The strong 1996 year class had maximum catch in numbers-at-age 3. The 2000 year class was caught in very large numbers-at-age 1. If the exploitation at age is similar to previous year classes, this year class must be much larger than the 1996 year class. However, if the fishery has been directed towards this year class, the estimate of this year class by a separable model will not be valid. There are indications in the fishery that this might be the case.

Log catch ratios are shown for selected ages in Figure 8.4.2. The lines are largely parallel, except for year to year noise. There is some declining tendency for the curve for ages 2-3, however, which may suggest either a lower selection at age 2, a higher selection at age 3 or both. The curve for ages 8-9 is mostly above the curve for ages 5-6, confirming that F at the oldest age may be higher than at intermediate ages. Because of this, the total mortality has been higher than indicated by the level of the curves.

The Survey indices are shown by year class and survey in Figure 8.4.3, and the log indices by year class in Figure 8.4.4. The most striking finding is that in both the recent acoustic surveys, the abundance indices are far higher than in previous years, but the patterns are highly divergent, and inconsistent from year to year. Thus, the signals in the surveys about mortality in recent years is extremely noisy, and the signal about year class strengths are conflicting. The Spanish CPUE series is not consistent with the other surveys. Here, the recent presumably large year classes are not apparent. This series covers a very limited part of the distribution area, and the data may not be representative of the stock as a whole.

8.4.3 SSQ surfaces

As noted in Section 4.5, most of the time trajectories of mortality and stock abundance are determined by the catches once the terminal fishing mortality has been fixed. Figure 8.4.x.5 confirms this for blue whiting. Here, the estimates of F3-7 by the WGNPBW with ICA, AMCI and ISVPA are shown together with results from a simple VPA with similar fishing mortalities. The VPA was constrained by requiring that the selection in the last year was similar to that the year before, and that the fishing mortality at the oldest age was 1.5 times the fishing mortality-at-age 5.

Profiles of the loss function for blue whiting are presented for AMCI, ISVPA and ICA. The ICA program does not allow screening over different terminal F values. Therefore, a spreadsheet version of ICA was constructed where the backward VPA part was replaced by a cohort analysis. The settings for ICA were taken from the WGNPBW meeting in 2002 except for the length of the separable period and the absence of splits in the time-series of acoustic indices (see Section 8.4.1.

AMCI was run using all survey data and using only one survey at a time together with the catches. Attempting to use the Russian acoustic survey only or the Spanish CPUE only, gave unstable results and are not presented.

ISVPA was used in its effort-controlled version (i.e. strict separability) with minimization of SSE terms for each source of data. For testing the impact of possible outliers in the data on the solution, the median of the distribution of squared residuals was also used. In order to outline the role of the separable constraint in the model, the catch-controlled (VPA-type) version of the model, was also tested. Since analysis of ISVPA residuals have shown that the 1992 year class appears to be problematic (see Section 8.4.4.) runs were also made for catch-at-age data with this year class excluded. A similar run was also made with AMCI

Figure 8.4.7 shows that each of indexes, except the data on SSB from Russian surveys, being used for ISVPA fit one by one, produce minimum of the SSE. These minima correspond to similar values of effort factor in the terminal year, except for CPUE of Spanish trawlers. To diminish the influence of outliers on the ISVPA solution, in its objective function for age-structured data instead of SSE the median of squared residuals was used. The reason was that in the case of minimization of median the residuals witch take place in marginal positions of the distribution, with not contribute to signal. Fig. 8.4.8 shows that only Norwegian sea acoustic surveys and catch-at-age data still produce signals about state of the stock in recent years and, perhaps, might be considered as more reliable in comparison to others. In case of application of VPA-like mode of the ISVPA with SSE-minimization (Fig.8.4.9) signals from most of data were dissipated or were shifted to extremely high F, which is common for situations with conflicting data.

The impact of excluding the 1992 year class, which produces high residuals because of contrast between its low representation in catches in young ages and high in older ones, was tested. With ISVPA, exclusion of this generation from the catch-at-age data resulted in a shift towards higher terminal F for signals obtained from most of the indexes and from the catch-at-age. Application of a median minimization instead of the SSQ minimization, makes the situation more clear and leaves the position of minimum unchanged for one age-structured index (Norwegian Sea acoustic survey for 1992-2001). With AMCI, exclusion of the 1992 year class from catch and survey data gave a slight increase in F (from 0.78 to 0.85) and a corresponding reduction in SSB.

Similar experiments with AMCI also have shown that position of the minimum of the model objective function strongly depends on the source of data is used and minima in most cases are very flat (Figure 8.4.6). For ICA the minima are almost not detected (Figure 8.4.10), but there also could be some problems with minimization routine when it is started rather far from global minimum.

8.4.4 Sensitivity analysis and selection patterns

In order to explore which data are most important for the final estimate of the terminal fishing mortality, the individual squared residuals at the optimum terminal F were compared with the squared residuals at some different terminal F. Outside the optimum, the sum of the differences between the squared residuals should be positive, since the sum of squared residuals will be larger. However, when changing the terminal F, the model will fit better to some data and worse to others, and the optimum will be determined by the balance between these 'winners' and 'losers'. The purpose of this analysis was to identify data that were particularly sensitive to changes in terminal F, and thus identify the data which are the most important for deciding on the optimum terminal fishing mortality.

Figure 8.4.12 shows the difference between squared log residuals of catches by increasing terminal F above the optimum value with the effort controlled version of ISVPA, using the log sum of squares as the objective function. The negative differences represent data where an increase in the fishing mortality would give an improved model fit. This is the case with the catches-at-ages 1 and 2 in the recent years, and also the catch-at-age 1 in 1983. The data to which the model fit would be substantially poorer are the catches at most ages from the 1992 year class, and to a lesser extent, some of the catches in 1985 – 1990. The large catches at young age in recent years would require either a high fishing mortality or large stock numbers, which may explain why the model fits better to these data with a higher fishing mortality. It is more surprising that these are counter-balanced by the catches from the 1992 year class. The catches from this year class were unduly small at young age and unduly large at old age, compared to what one should expect according to the selection pattern (see Figure 8.4.14). The model fits quite well to the older ages, by assuming a year class strength in accordance with the relatively low terminal fishing mortality, but gets larger negative residuals at younger age (see Figure 8.4.22). As the terminal fishing mortality increases, the fishing mortality in the neares years before will also have to increase, and this would lead to even larger negative residuals.

Figure 8.4.11 shows the difference in squared log residuals with AMCI, but with a decrease in fishing mortality. In the catches, almost all difference is in the last year, and balanced between the ages. Most of the signal is in the surveys, where again the differences are great in the last two years. In the Norwegian spawning acoustic survey in 2001, reducing terminal F worsens the fit to the oldest ages and improves the fit to age 2. In the Norwegian sea acoustic survey, the fit to most ages improves in 2001 when the terminal fishing mortality is reduced, but becomes worse for age 5 in 2001. It also improves for some young ages in 1992-1993. The fit to the Russian survey remains virtually unchanged. The fit to the Spanish CPUE series is again mostly changed in the last years, where it gets poorer by decreasing terminal F for most ages, but it also worsens slightly for most ages and years throughout the whole time-series.

The change in squared residuals by reducing the terminal F in ICA is shown in Figure 8.4.13. Here, the changes are very large and unsystematic and prominent over all years, suggesting problem with the estimates of the stock even in the period covered by the VPA. It also gives the impression that the final estimate of terminal F to a large extent is driven by the noise in the survey data.

The most striking difference between ISVPA and AMCI is that in ISVPA, the terminal F seems to be fixed by the balance between the catches at young age in the most recent years and the catches from the 1992 year class, whereas in AMCI, it apparently is a compromise between conflicting evidence from the various surveys. The strong impact of the 1992 year class in ISVPA may be related to the strong year class effects in the catch residuals with this method (Figure 4.4.22(?)). This pattern is probably caused by the constraint on row sums and column sums. The analyses so far suggest this as the most likely cause of the discrepancy between AMCI and ISVPA with respect to estimates of recent fishing mortality and stock abundance.

The selection patterns in the last years for the optimum terminal F and the alternative terminal F in the analyses described above. Changing the terminal F in general leads to small changes in the selection, although the selection at the old age is sensitive to the shift in terminal F in both AMCI and ISVPA, but to a different extent (Figure 8.4.14). This probably reflects the problem of fitting to the relatively large catches-at-age 9, as discussed above.

8.4.5 Residual patterns

Residual patterns of catches (for separable models) and of catchability were analyzed for all models applied. For AMCI, the residuals were calculated for the blue whiting catch and tuning fleets data for different periods of the time-series and using all fleets together and in individual fleet runs.

Figure 8.4.15 shows the results of the first run carried out with the catch and all tuning fleets. In general, catchability residuals of the catch were rather low except for the age 0. No apparent trend are observed. This situation was observed in all runs carried out just with the catch-at-age data and also in runs deployed with the individual fleets (Figure 8.4.16, 8.4.17, 8.4.18, 8.4.19, 8.4.20 and 8.4.21).

Catchability residuals of the catch and all fleets deployed without the year 2002 are presented in Figure 8.4.16 no essential differences were found in relation to the previous run.

Norwegian spawning acoustic Survey showed in all runs deployed a positive year effect from 1988 to 1991 and in the last year of the series. Also, for the same survey a negative year effect is detected in 1986 (Figure 8.4.15, 16 and 18). Catchability residuals of the Russian spawning acoustic survey presented in all runs deployed with the rest of the fleets and individually evenly distributed residuals and no much noticeable year and age effects (Fig. 8.4.15, 16,19). Opposite to the Norwegian spawning acoustic Survey, the Spanish CPUE series presented negative trends in the last years of almost all ages and positive ones also in all ages at the beginning of the series. The year effect is very noticeable in this fleet (Figure 8.4.15, 8.4.16 and 8.4.20). The last of the surveys studied, the Norwegian Sea Survey, showed larger residuals than the others and a detectable year effect in 1989 and 1992 (Figure 8.4.15, 8.4.16 and 8.4.21).

ISVPA catchability residuals

ISVPA was used to estimate the catchability residuals of the catch and the tuning fleets for their global minimum and a second run was deployed for each fleet separately and for each minimums (Figure 8.4.22 and 8.4.23).In both runs an important age effect was detected in the residuals of the catch along all the period. Residuals of the fleets behaved in a very similar way even when in the individual fleet runs, Norwegian spawning acoustic Survey, Russian spawning acoustic Survey and Norwegian Sea Survey were separated in different periods and these considered as different fleets. Norwegian spawning acoustic Survey showed in all runs and for the different period used, a negative year effect from 1988 to 1991 opposite to what it was observed in the AMCI runs. Catchability residuals of the Russian spawning acoustic survey presented in all ISVPA runs positive trends along the first years of the series and also a clear age effect, however in the middle part of the period larger and negative residuals are detected (Fig. 8.4.22 and 23). The Spanish CPUE series presented positive trends in the last years of almost all ages, opposite to what AMCI showed, and age effect is detected at the beginning and the end of the period. As in the model described above, the Norwegian Sea Survey, showed largest residuals of the fleets used and important detectable year (1989) and age effect (Figure 8.4.22 and 23).

In ICA, four different runs for the diagnostics were carried out (Figure 8.4.24). The first one was just over the catch (a) . In the second run the catch was also analysed but with constrains on the sum of columns and rows of the residual matrix (b). The other two runs were carried out with the catch and: c) the Norwegian spawning acoustic Suvey and d) the Norwegian Sea Survey.

In all cases, residuals are much larger in the latest period (from 1992 onwards) than in the earliest. Also these residuals were bigger when estimated by ICA than by AMCI or ISVPA. For the catch, in all cases, there is a considerable negative trend and a marked year and age effect in the early years, for the rest of the period some age effect might be perceived.

Norwegian spawning acoustic Survey showed positive year effect from 1987 to 1991, as occurred in all runs in AMCI. Also, for this survey a negative year effect is detected in 1986. As detected in AMCI, an age effect is noticeable in the last three year and for the older three ages (Figure 8.4.24 c). The last of the surveys studied, the Norwegian Sea Survey, showed larger residuals than the other and a detectable year effect in 1989 and an oldest ages effect in the early years (Figure 8.4.24 d).

XSA catchability residuals

Individual fleet runs were deployed in XSA to check for catchability residuals. As it was done in ISVPA, the different fleets, except for the Spanish CPUE series were separated in different periods and these were considered as different fleets. The year 2002 and the plus group 8 were not considered in any of the fleets.

As in AMCI, the Norwegian spawning acoustic Survey showed a positive year effect from 1988 to 1990 and a negative year effect is detected in 1986. Some age effect is detected. Catchability residuals of the Russian spawning acoustic survey presented evenly distributed residuals and a slight year effect in 1989 and 1992. Catchability residuals appeared to be slightly larger than those estimated by AMCI. In general, residuals of the Norwegian Sea Survey were larger than the ones estimated for the other fleets. As it happened in AMCI, an important year effect is detected in 1989 when also the largest residuals are found. Some age effect are noticeable in the younger ages at the beginning of the series (1980-1985). Opposite to the Norwegian spawning acoustic Survey, the Spanish CPUE series presented negative trends in the last years for all ages and positive ones also for almost all ages at the beginning of the series. Year and age effect are very evident in this fleet (Figure 8.4.25).

CADAPT catchability residuals

Individual catch and fleet runs were deployed in CADAPT to check for catchability residuals. In this case only the Norwegian Surveys were analysed. Two different runs were carried out for each fleet one with the age 10 included and the other without.

No much differences appeared to be between the two runs of the Norwegian spawning acoustic Survey. Relatively smaller residuals appeared to occur when the age 10 is not considered in the analysis. In both cases, this survey showed, as in AMCI, ICA and XSA, a positive year effect from 1988 to 1991 and a negative year effect in 1986. In the Norwegian Sea Survey, also relatively bigger residuals occurred when the age 10 is not considered in the analysis. However, in general, this fleet showed smaller residuals when compared to the results from the previous models. Year effects are detected at the beginning of the period and in 1989, 1992 and 2000 (Figure 8.4.26).

8.4.6 Investigative exploration with CADAPT

Exploratory runs of CADAPT were carried out using the blue whiting assessment data.

A word of caution – CADAPT was not developed for assessing blue whiting and the software and model are still under development. Results are presented here for the purpose of comparison with more established methods and perhaps to highlight how cadapt might be modified to accommodate the needs of a blue whiting assessment.

Four runs are presented:

- run5: Catch-at-age tuned with Norwegian Sea acoustic survey age-group indices. Plus group (10+) in catch-at-age treated as age 10.
- run6: Catch-at-age tuned with Norwegian Sea acoustic survey age-group indices. Plus group (10+) in catch-at-age treated as age 10.
- run7: Catch-at-age tuned with Norwegian spawning acoustic survey age-group indices. Plus group omitted, age groups 1-9 in catch-at-age used as input to CADAPT.
- run8: Catch-at-age tuned with Norwegian spawning acoustic survey age-group indices. Plus group omitted, age groups 1-9 in catch-at-age used as input to CADAPT.

In the 4 runs, CADAPT was run with proportional catchability on all age groups, tuned with indices of age groups 2-7, the age range given in both the Norwegian spawners survey and Norwegian Sea summer survey data sets. Index data sets used were from 1986-2001 for the summer survey in the Norwegian Sea, 1986-2002 for the Norwegian spring survey on the spawning grounds.

Retrospective analyses were carried out on runs 7 and 8 (norsea0-9 and norspa0-9) for numbers-at-age 0 and age 1, reference fishing mortality (ag 3-7) and SSB.

8.4.7 Comparisons and conclusions

Results of the different assessment methods applied to blue whiting are summarized in Figure 8.4.27. For each of the assessment methods the plots contain the full calibrated estimates of recruitment, SSB and mean fishing mortality (ages 3-7). In addition a number of single fleet runs have been included to indicate the level of uncertainty that emerges from the combination of the auxiliary information with the catch-at-age data.

A number of general observations can be made from the comparisons.

- AMCI and ICA are reasonably comparable in their explanations of the catches given the auxilliary information. Both tend to give a high fishing mortality in the recent years and a decrease in the spawning stock. However, both models also give a wide range in the possible fishing mortalities in the final year. The SSQ surfaces shown in Figure 8.4.11 and 8.4.13 indicate that the minimum is not well defined and therefore the solutions tend to be very unstable.

- ISVPA estimates a much larger stock and a lower fishing mortality in the recent years when compared to AMCI and ICA. It is also noted that the assessment of ISVPA is somewhat out of phase with the other two methods. For example, the high fishing mortality in the late 1980's is estimated to have been in 1990 by AMCI and in 1989 by ISVPA. The fact that ISVPA generates very high fishing mortalities in the late 1980's can probably be explained by the strong cohort effect in the residual patterns. It is thought that the constraint to minimize the row- and column-sums of the catch residual matrix (or the F-residual matrix) may cause high year class effect which can cause the stock numbers and fishing mortalities on entire cohorts to be shifted up or down. From an alternative perspective, the 1992 year class has a strong anomaly in its catch dynamics for young ages (Figure 8.4.1) and it may not be unreasonable to retain high residuals for this year class, which means that the estimates of the model parameters' are less based on this cohort.

- The XSA assessment tends to produce relatively lower fishing mortalities and higher stock abundances when compared with ICA and AMCI. However, the mean fishing mortalities are expressed in terms of mean F over age 2-5 instead of ages 3-7. The estimates are similar to ISVPA. The estimates in the final year are, as for the other models, very uncertain. Depending on the source of information included, the estimate of fishing mortality (at ages 2-5!) can vary between 0.3 and 0.6 (the Spanish CPUE series gives and outlier of F=0.9).

- The CADAPT analysis has been included for comparison only. The results are only exploratory in the sense that the model is still under development.

- The Spanish CPUE series appears to give a very high fishing mortality in all the models that were tested. From the log catchability residuals and from the raw survey data itself, it was already clear that this survey does not give reliable information on the stock development for blue whiting. The WGMG considers that the assessment would improve by leaving out this index from the calibration process.

- There is general agreement between the methods about the size of the strong 1996 year class.

- Recruitment of the recent year classes is considered to be very uncertain.

Overall, the findings can be summarized as follows. The different assessment methods find very different estimates of stock size and exploitation rate in the most recent years. The auxiliary information is contradictory and does not lend itself to unique characterization of the stock development. Also, model mis-specification may contribute to the difficulty in assessing the state of the stock. Two notable problems appear to stand out:

- Conflicting sources of information appear to present the main problem in the blue whiting assessment. The conflict in the **data sources** is handled differently by the different methods that have been applied to this stock (e.g. AMCI and ISVPA)

- There are indications of changes in selection of the most recent (strong) year classes which appear to have a higher exploitation on the younger ages compared to the older ages. Although this may be a relative change only, it could seriously affect models that assume a fixed selection pattern over a longer period of time.

- The constraint of zero row- and column-sums of the residual matrix in ISVPA seems to be a contributing factor to the difference between ISVPA and other separable models. Further work is necessary in order to fully understand the causes and implications of these constraints.

WGMG recommends that the Northern Pelagic and Blue Whiting Working Group [WGNPBW] explore the following issues when assessing the stock of blue whiting in 2003 (and beyond):

- The choice of appropriate model to assess the stock is not clear-cut and the approach used at this meeting of WGMG of exploring a number of competing models is to be commended as an aid to disentangle the apparent conflicting sources of data.

- To exclude the Spanish CPUE series from the calibration process as this survey has clearly shown to have no relevant information to the overall stock.

- To exclude the acoustic series which end in the 1980s because these surveys only provide information on the cohorts that have been fished out by now. Alternatively it could be considered whether the two survey periods could be combined if a gear effect can be estimated for the introduction of the new EK500 equipment.

- It was found that the survey data was very noisy and often contradictory. This may be caused by sampling problems in the sampling for age in the acoustic surveys. The WG could explore the possibility of including acoustic survey data as SSB estimates rather than age-disaggregated indices. This would perhaps get rid of the noise in the age signal although the mortality signal would be lost.

Furthermore, WGMG recommends that its members:

- **further explore the historical behaviour of ICA, given the observations of the simulated data with ICA.**

- **further investigate the use of VPA type models (e.g. XSA) that are independent of separable assumptions and may give alternative interpretations of the data next to the family of separable models as ICA, AMCI and ISVPA.**

- **explore the convergence behaviour of AMCI in the light of very shallow SSQ surfaces.**

8.5 Answer to the special request

During the coastal state meeting on blue whiting in Oslo on 7-8 November 2002, the assessment and management advice given by ICES was presented and discussed. Assessments and predictions for blue whiting and Norwegian Spring Spawning herring based on the ISVPA model (Vasilyev WF3) was presented in the Oslo meeting and was compared to the output from the models presented by ACFM in 2002 (AMCI and SeaStar respectively). The results with ISVPA gave substantially different estimates of the stock size of both stocks. It is known that both these models are used by ICES in *The Northern Pelagic and Blue Whiting Working Group* along with other available tools.

The parties have requested ICES to "evaluate the two assessment models with respect to Norwegian Spring Spawning herring" and to "extend these evaluations to also include assessment of blue whiting".

Unfortunately, the evaluation of the SeaStar model could not be carried out during the working group meeting, because the main author of the model could not be present, and therefore the group was not able to evaluate the Norwegian Spring Spawning herring assessment. However, the approach taken for blue whiting could equally be applied to the Norwegian Spring Spawning herring.

WGMG has looked extensively at the methods that have been applied for the assessment of blue whiting and also to other methods that could be applied (i.e. AMCI, ISVPA, ICA, XSA, CADAPT). The WG has attempted to characterize the main features of the models (Section 4.2). In Section 8.3 the group has investigated the behaviour of the models on simulated data with known properties. Unfortunately only one simulated dataset could be analysed and the properties of the dataset did not compare to the type of problems that are encountered in the blue whiting stock. Nevertheless, the exploration of the assessment methods on simulated data was considered to be useful.

The group noted that in general the models were able to pick up the signals from the simulated data. The trend in catchability in one of the tuning series caused the assessment to predict levels of SSB which were higher than the true

values. The estimation of natural mortality within the ISVPA appeared to remedy the increase in catchability, however it is clear from this example that the estimation of natural mortality without additional information may seriously compromise our understanding of the causes of the observed dynamics of natural populations.

A second notable feature from the explorations, was that the ICA model tended to over-estimate the historical stock size substantially. The ICA model consists of a mix of two models: a separable model for the most recent years and a traditional VPA over the historical part of the time-series. The VPA over the historic part is initiated from the population numbers and fishing mortalities that were estimated for the first year of the separable period. Furthermore, the fishing mortality on the oldest age is fixed to be the average of the fishing mortality of the immediately preceding ages. A possible explanation for the low correspondence between the ICA estimates and the true populations may be that the initiation of the cohorts is too high due to the average process, which then propagates back over time. However, this needs closer examination and could not be fully addressed during the meeting.

In the second part of the special request to evaluate the blue whiting assessment, WGMG has investigated several diagnostics of the input data to the assessment and the behaviour of the different assessment models given that input data. The conclusions to the analysis can be summarized as follows: **The different assessment methods find very different estimates of stock size and exploitation rate in the most recent years. The auxiliary information is contradictory and does not lend itself to unique characterization of the stock development. Also, model mis-specification may contribute to the difficulty in assessing the state of the stock**. Three notable problems appear to stand out:

- Conflicting sources of information appear to present the main problem in the blue whiting assessment. The conflict in the data sources is handled differently by the different methods that have been applied to this stock (e.g. AMCI and ISVPA).

- There are indications of changes in selection of the most recent (strong) year classes which appear to have a higher exploitation on the younger ages compared to the older ages. Although this may be a relative change only, it could seriously affect models that assume a fixed selection pattern over a longer period of time.

- The minimization of row- and column-sums of the residual matrix in ISVPA may be connected with year class effects in the catch residuals found for blue whiting but further work is necessary in order to fully understand the causes and implications.

The ICES methods WG (WGMG) has a general remit to consider the methodological issues that are pertinent to the process of stock assessment and the provision of biological advice. The group considers that it's findings with regards to the blue whiting assessment should therefore be considered as explorations into the underlying assumptions of the models that have been or could have been applied to blue whiting. The group does not have all the biological and fishery related knowledge to formulate the final solution to the question of how large the stock of blue whiting is at present. WGMG has outlined several approaches by which *The Northern Pelagic and Blue Whiting Working Group* should be able to improve the assessment in the forthcoming stock review (April 2003).

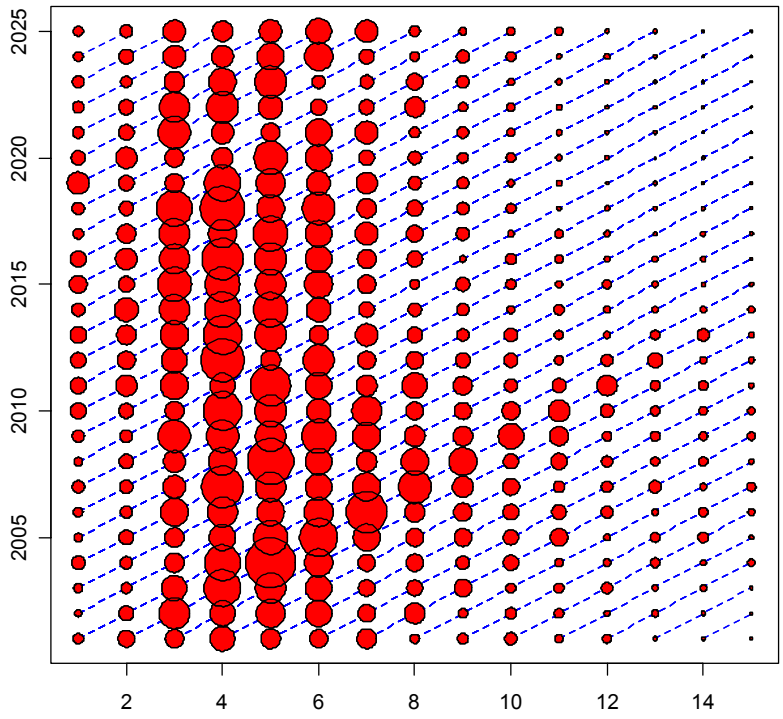


Figure 8.3.1 Simulated catch-age-age data as 'bubbles' proportional to numbers (the simulated data set is referred to as the noisy data set and is described in Section 3.2.

a) b)

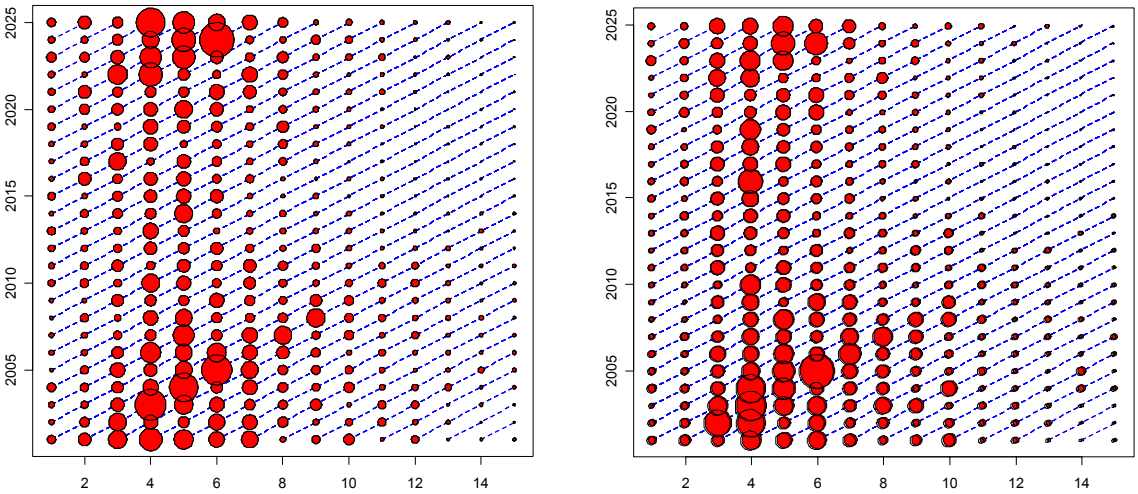


Figure 8.3.2 Simulated survey data. a) sim_I with a trend in catchability of 4% per year over the last ten years, b) sim_I2 without q trend. Bubbles are proportional to numbers. This is the noisy data set described in Section 3.2.

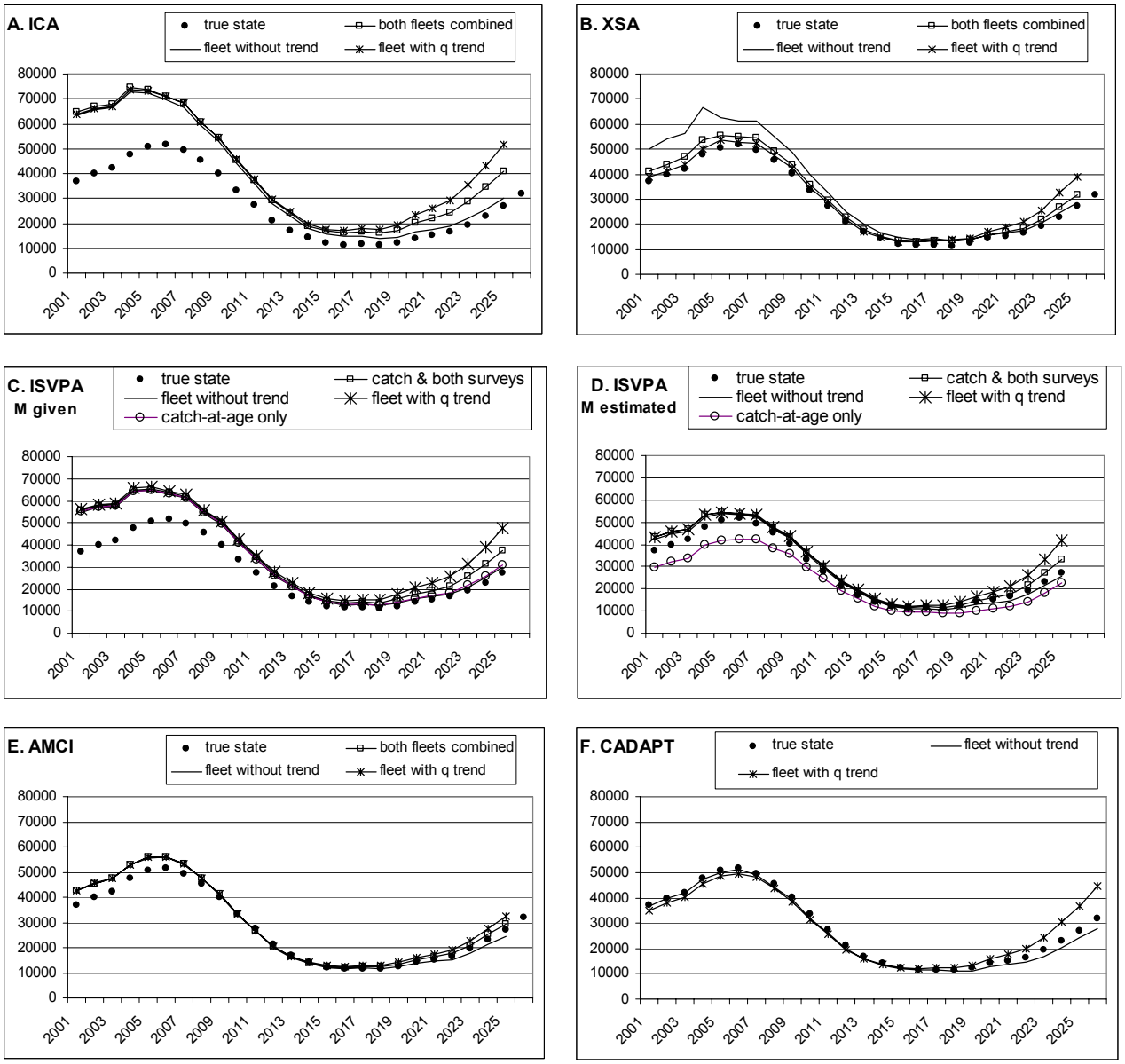


Figure 8.3.3 Comparison of SSB as estimated by the various assessment methods from the generated data set with noise. A. ICA; B. XSA; C. ISVPA with "given" M; D. ISVPA with estimated M; E. AMCI; F. CADAPT. Data from two fleets were available for use in the assessment, either singly or both combined. One of the fleets was simulated to have an increase in q of 4% per year over the last 10 years. Details of the settings are in Section 8.3.1.

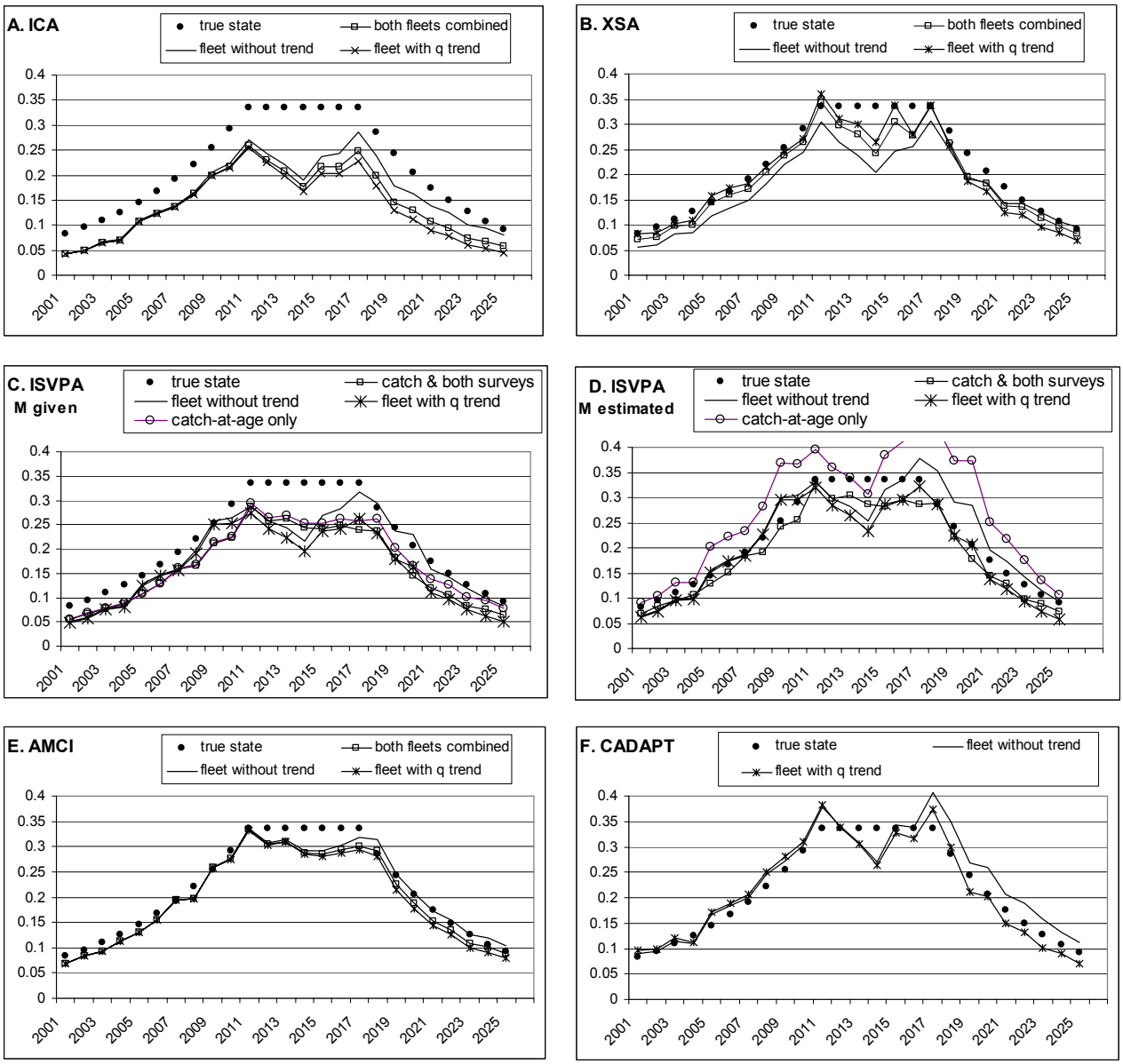


Figure 8.3.4 Comparison of mean $F_{8\text{-}12}$ as estimated by the various assessment methods from the generated data set with noise. A. ICA; B. XSA; C. ISVPA with "given" M; D. ISVPA with estimated M; E. AMCI; F. CADAPT. Data from two fleets were available for use in the assessment, either singly or both combined. One of the fleets was simulated to have an increase in q of 4% per year over the last 10 years. Details of the settings are in Section 8.3.1.

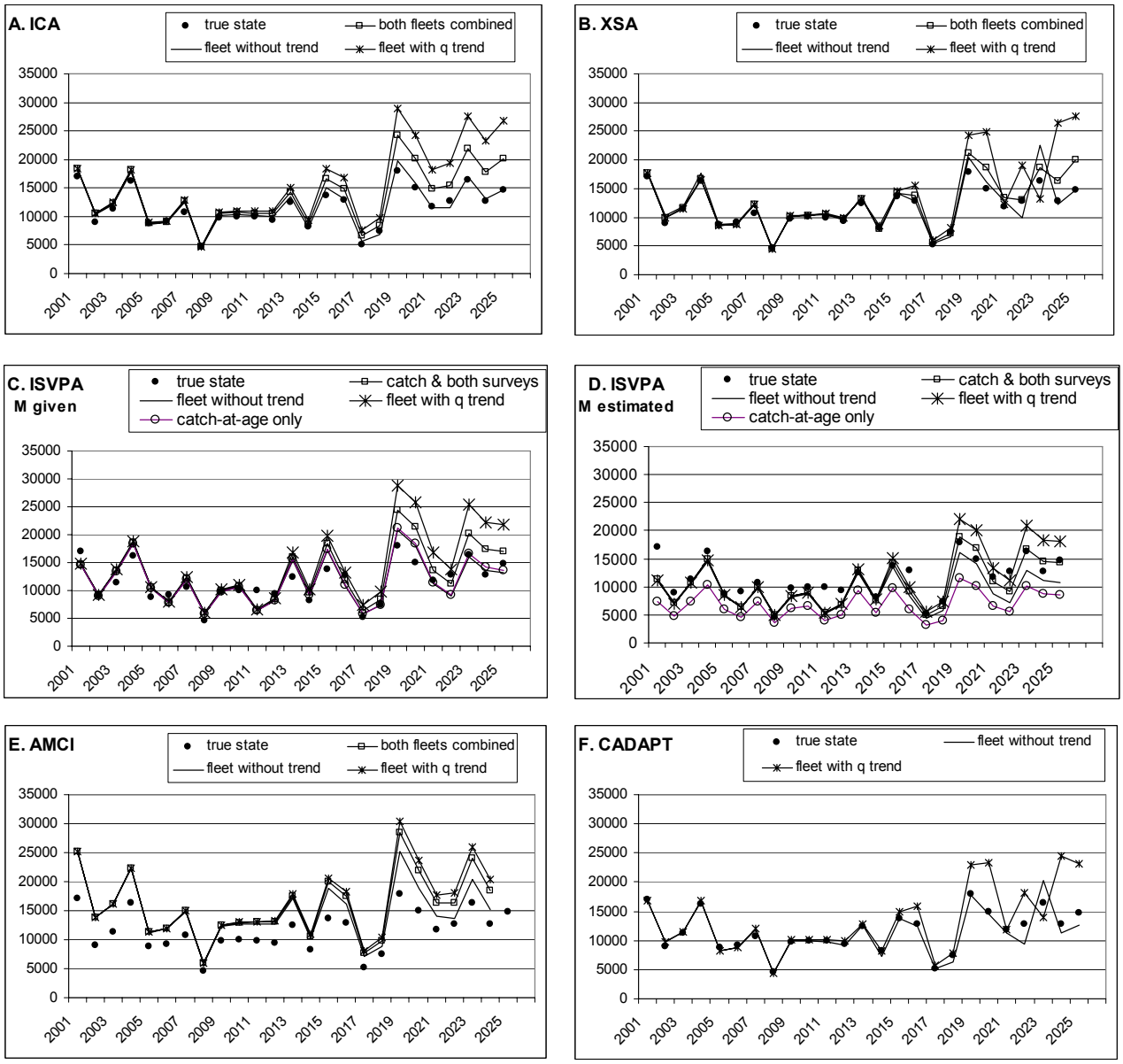


Figure 8.3.5 Comparison of number of recruits (age 1) as estimated by the various assessment methods from the generated data set with noise. A. ICA; B. XSA; C. ISVPA with "given" M; D. ISVPA with estimated M; E. AMCI; F. CADAPT. Data from two fleets were available for use in the assessment, either singly or both combined. One of the fleets was simulated to have an increase in q of 4% per year over the last 10 years. Details of the settings are in Section 8.3.1.

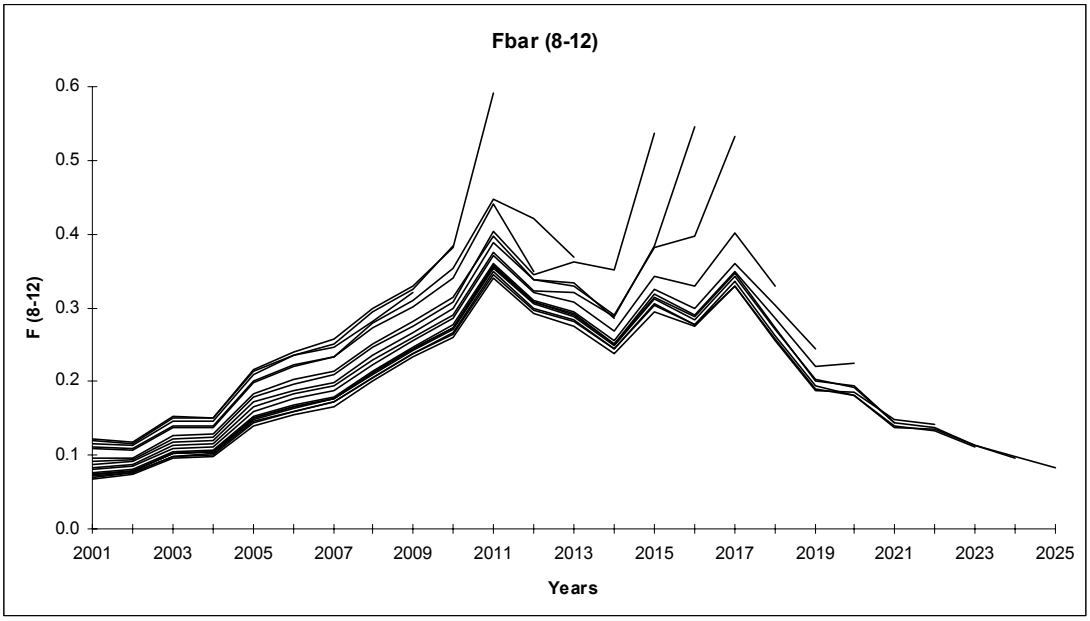
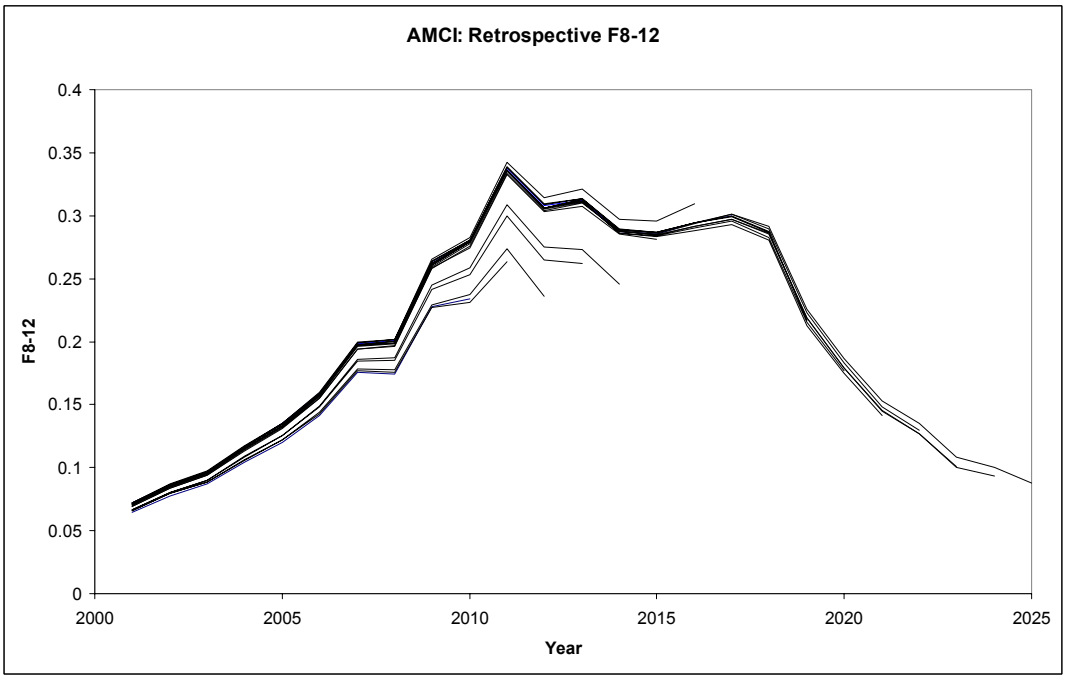


Figure 8.3.6. a) XSA retrospective analysis using simulated data (the noisy data described in Section 3.2).



b) AMCI retrospective analysis using simulated data (the noisy data described in Section 3.2).

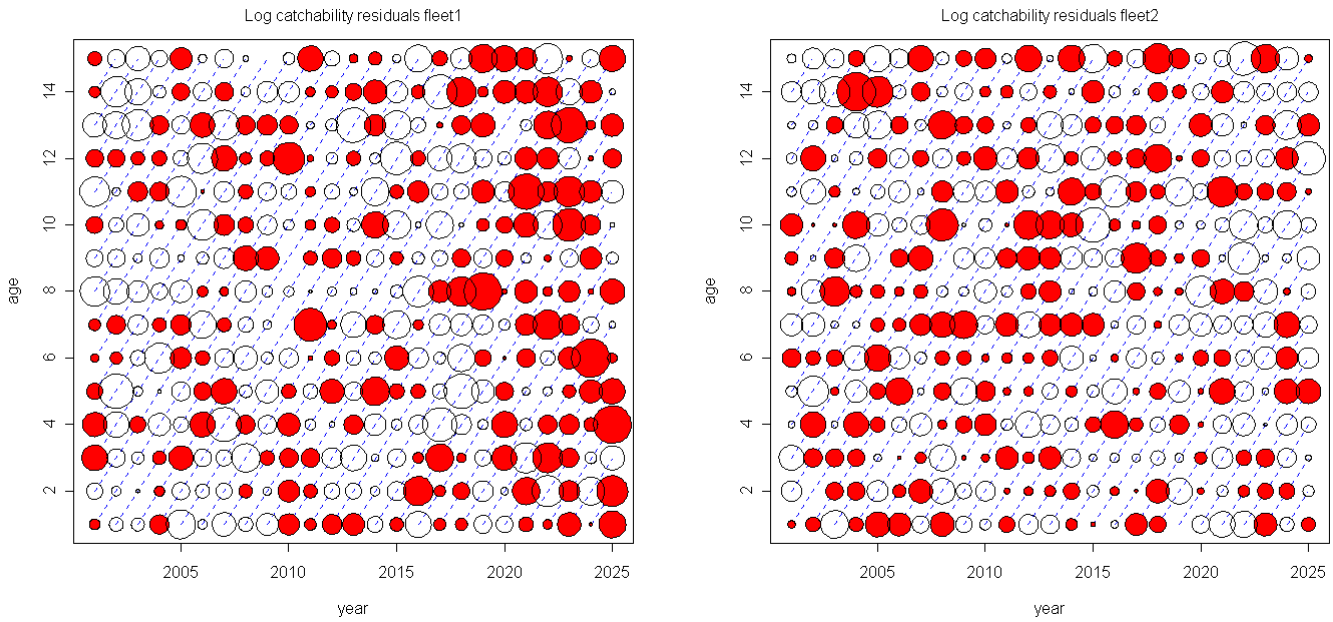


Figure 8.3.7 AMCI log catchability residuals for the two simulated survey fleets of which fleet 1 has a q-trend of 4% per year over the last ten years.

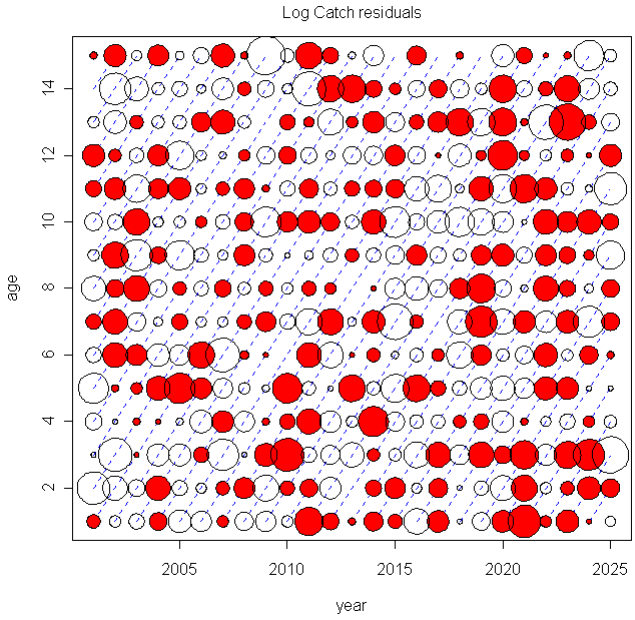
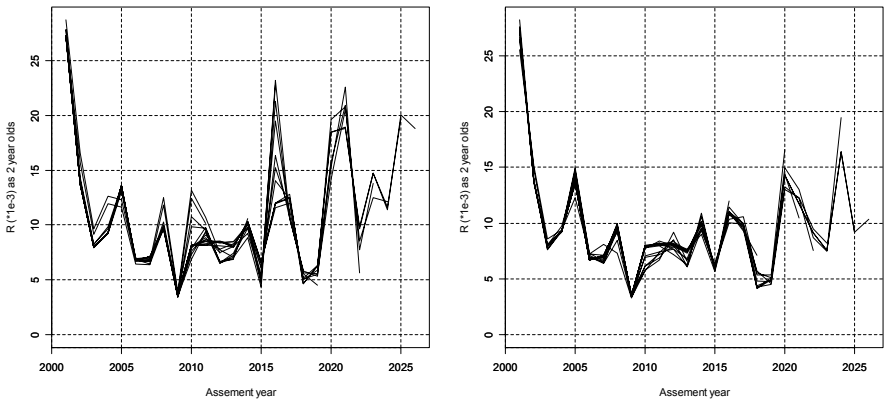
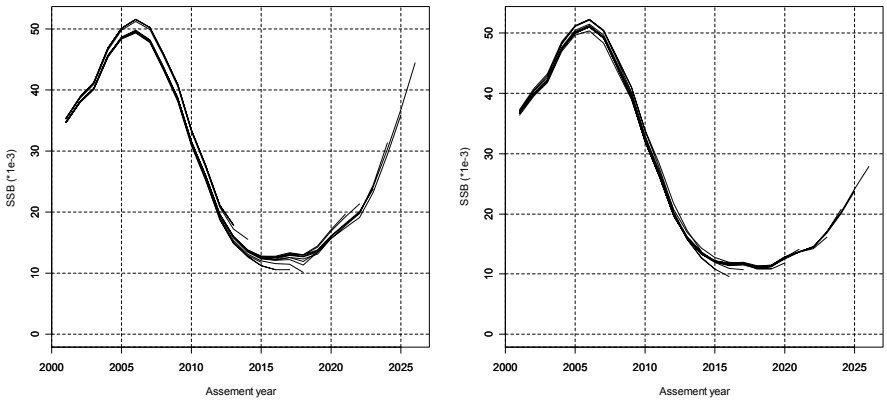


Figure 8.3.8 AMCI log catch residuals for an assessment of the two simulated survey fleets of which fleet 1 has a q-trend of 4% per year over the last ten years.

a) b)



c) d)



e) f)

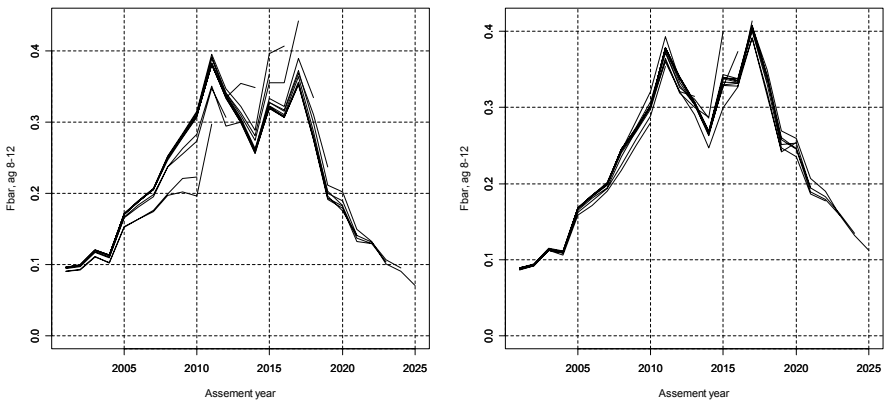
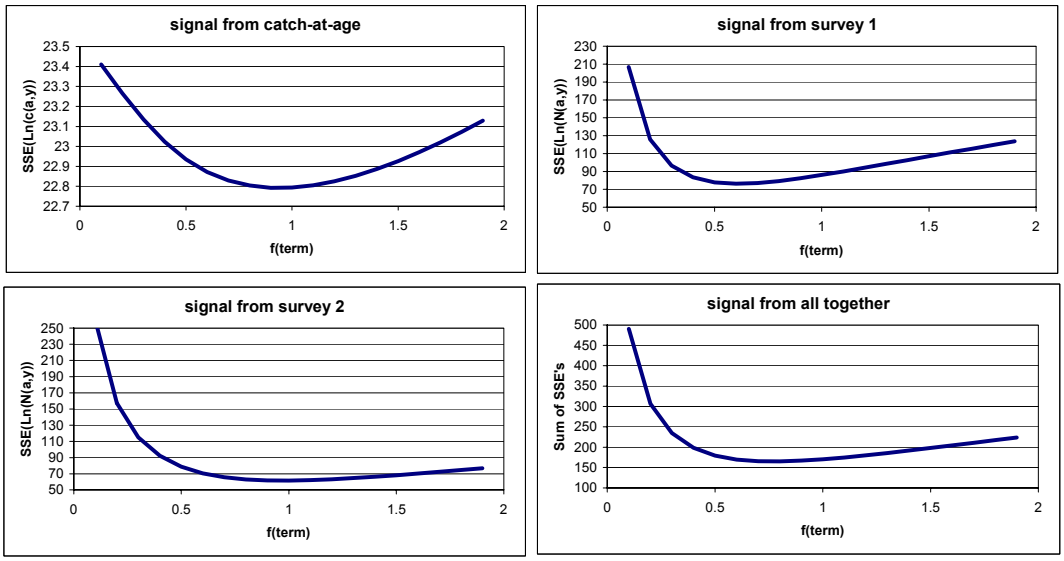


Figure 8.3.9 Retrospecive analysis of estimates of recruitment as 2 year-olds, $F_{8\text{-}12}$ and SSB of simulated stock from CADAPT tuned with: a,c,e) fleet with trend (sim_I); and b,d,f) fleet with no trend (sim_I2).



Profiles of the ISVPA loss function for different sorces of information

Figure 8.3.10 ISVPA SSE profile plots of runs with catch only, signal from the two surveys (survey 1 is with q trend) and all information combined.

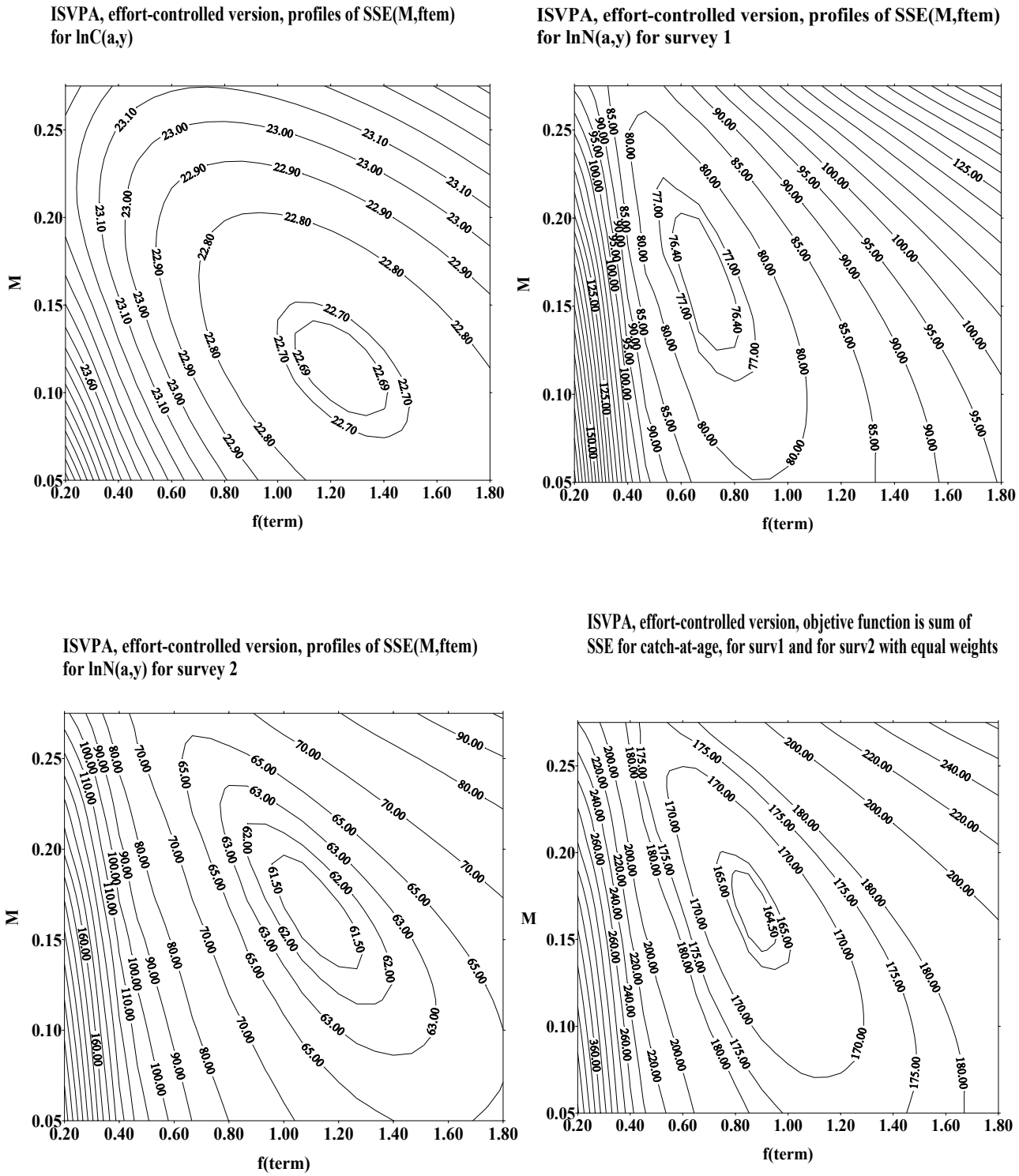


Figure 8.3.11 Isolines of the ISVPA objective function for the case where M is estimated within the model.

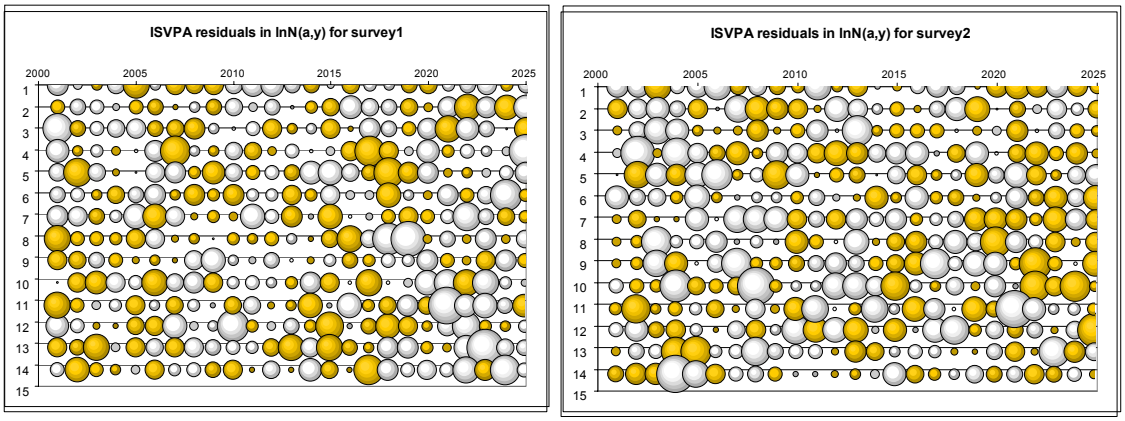


Figure 8.3.12 ISVPA log catchability residuals for two simulated survey fleets of which fleet 1 has a q-trend of 4% per year over the last ten years.

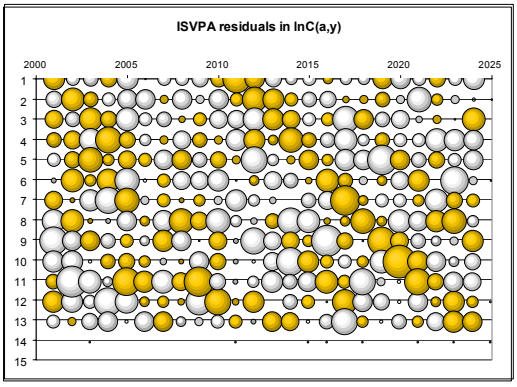


Figure 8.3.13 ISVPA log catch residuals for an assessment of two simulated survey fleets of which fleet 1 has a q-trend of 4% per year over the last ten years, when only catch-at-age data were used

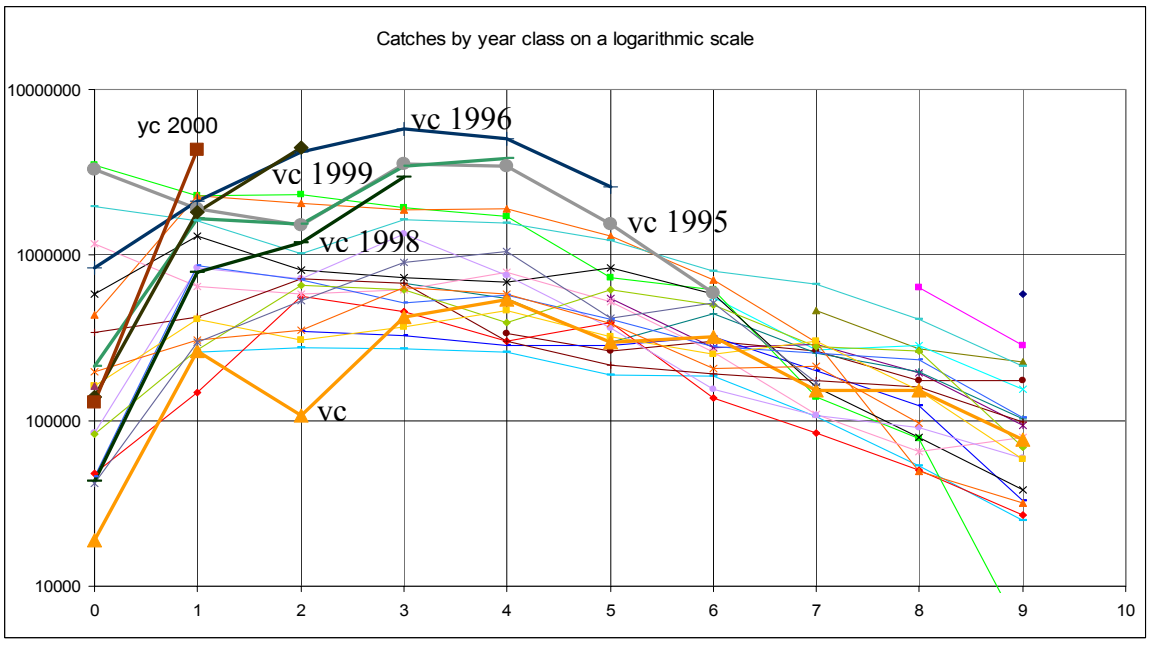


Figure 8.4.1 Blue whiting catches by year class on a logarithmic scale. We focus on the highlighted year-classes (thick lines with labels).

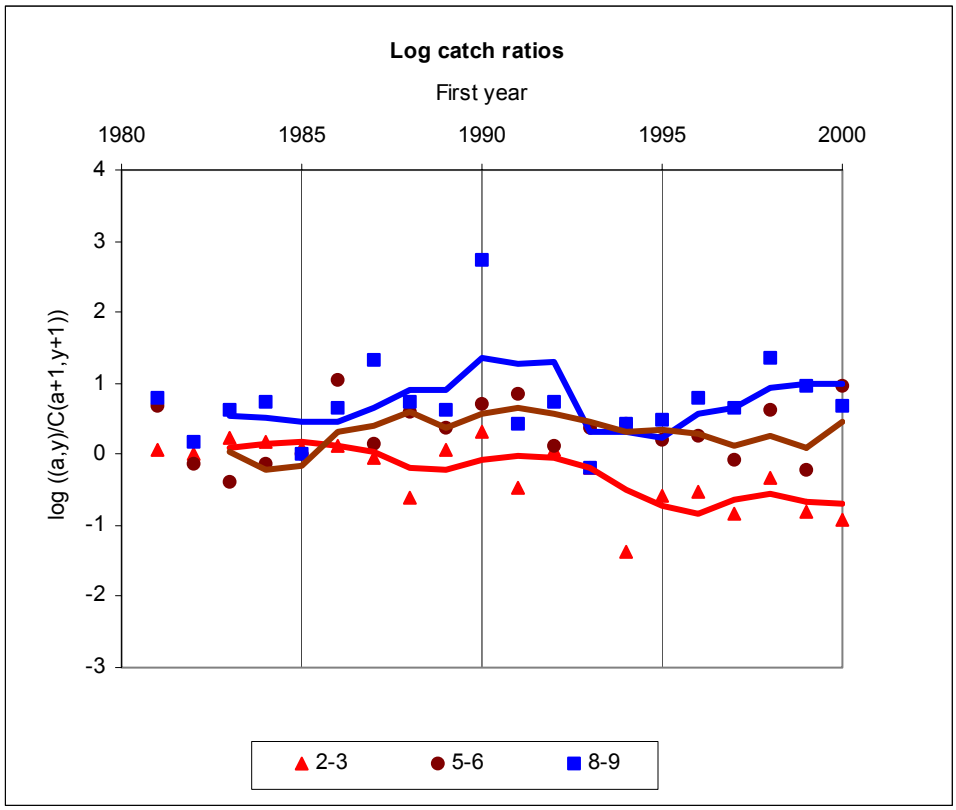


Figure 8.4.2 Blue whiting, log catch ratio. Points: Observed values. Lines: three-year averages.

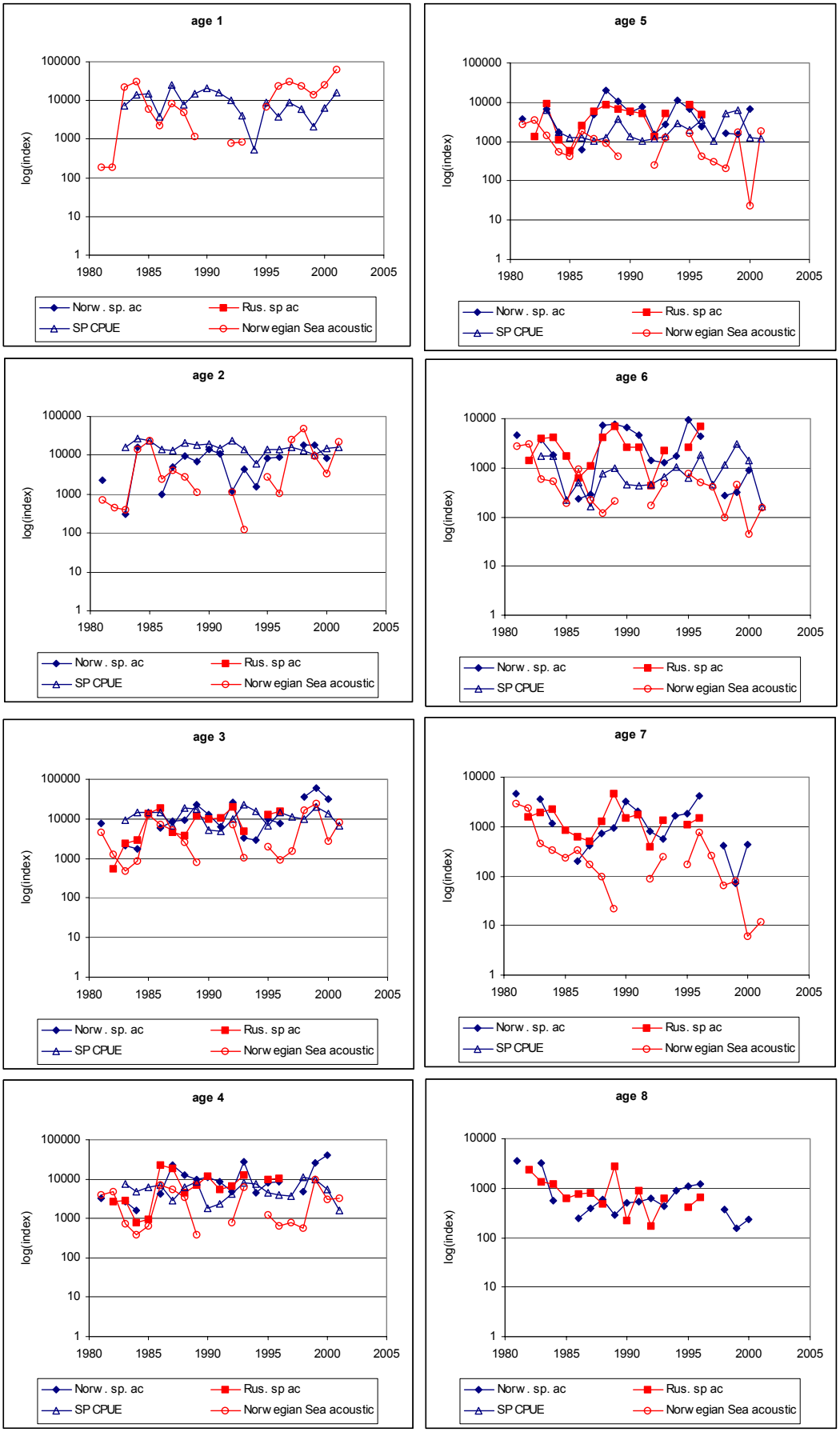


Figure 8.4.3 Blue whiting, log index value by age, by year and by survey.

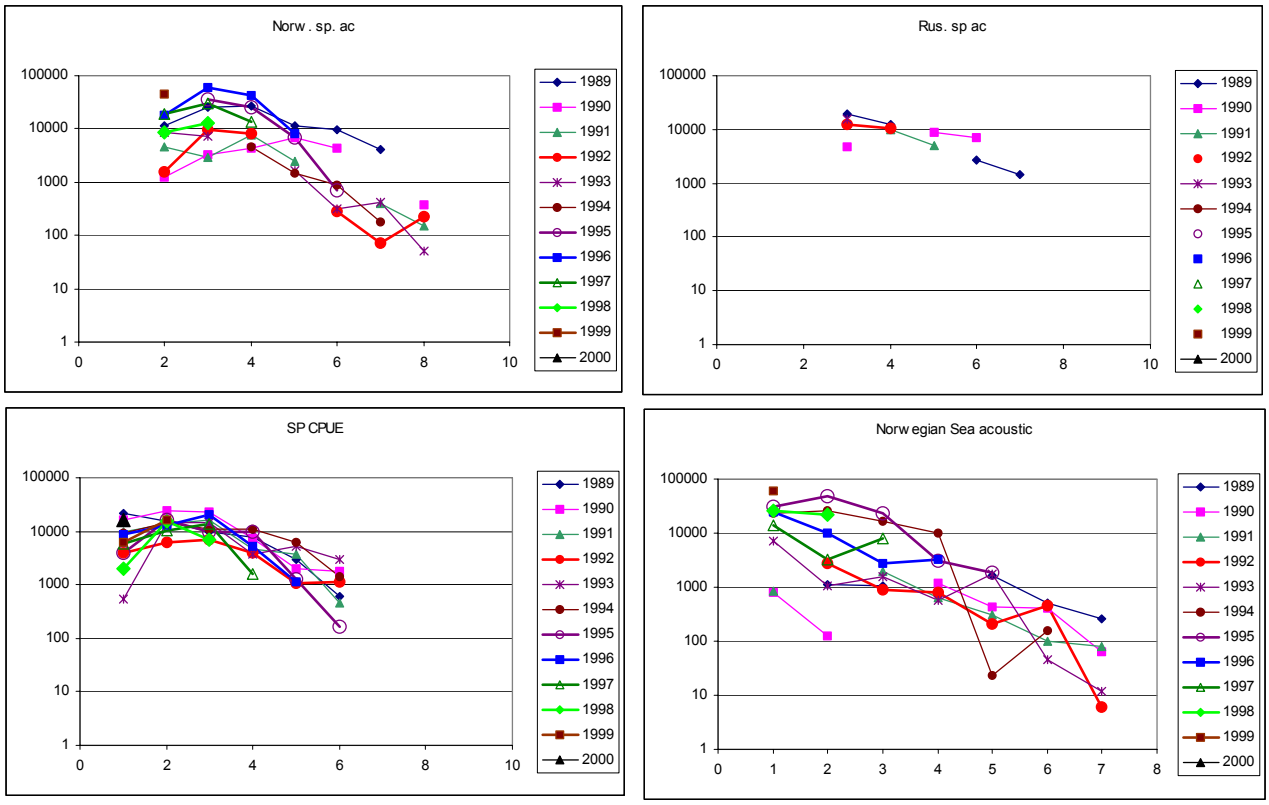


Figure 8.4.4 Blue whiting log index values by age, by year class and by survey.

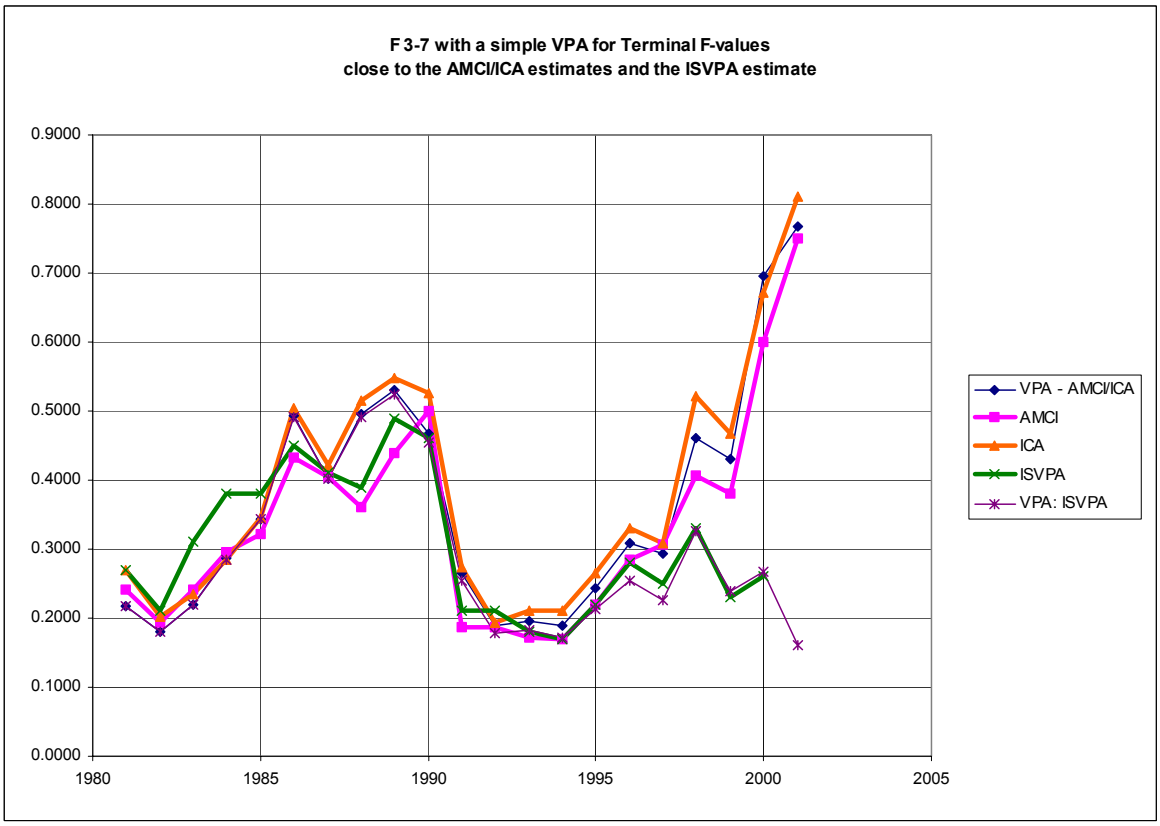
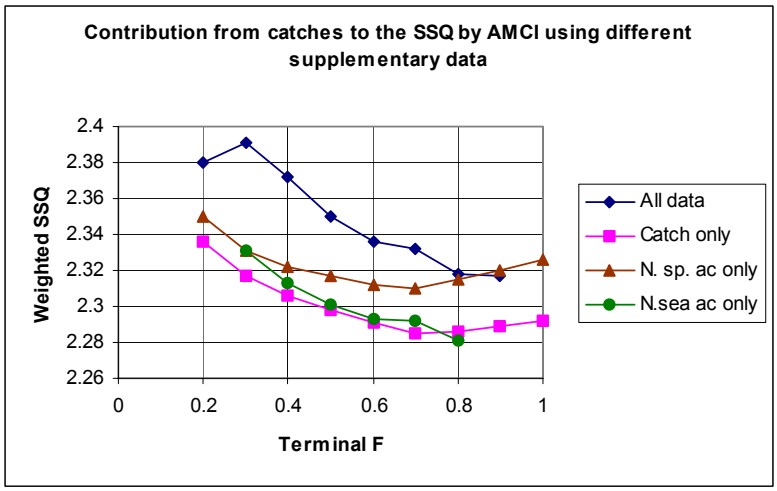
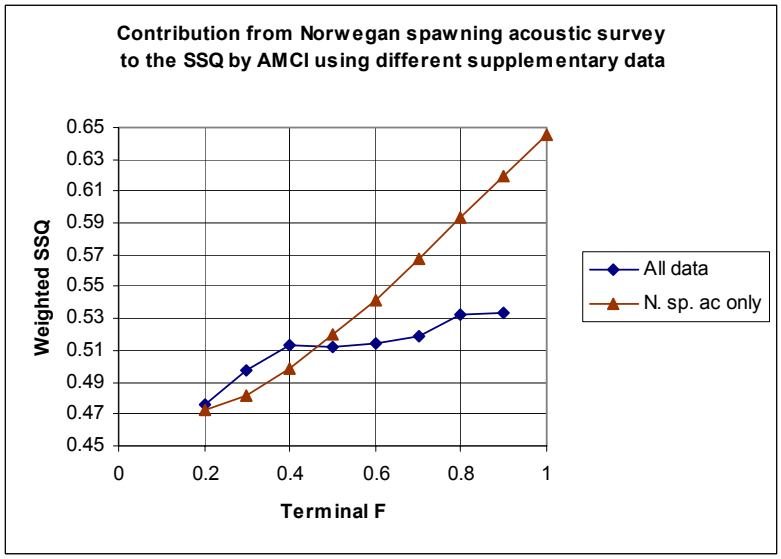


Figure 8.4.5 F_{3-7} with a simple VPA for terminal F-values close to the AMCI/ICA estimates and the ISVPA estimate.





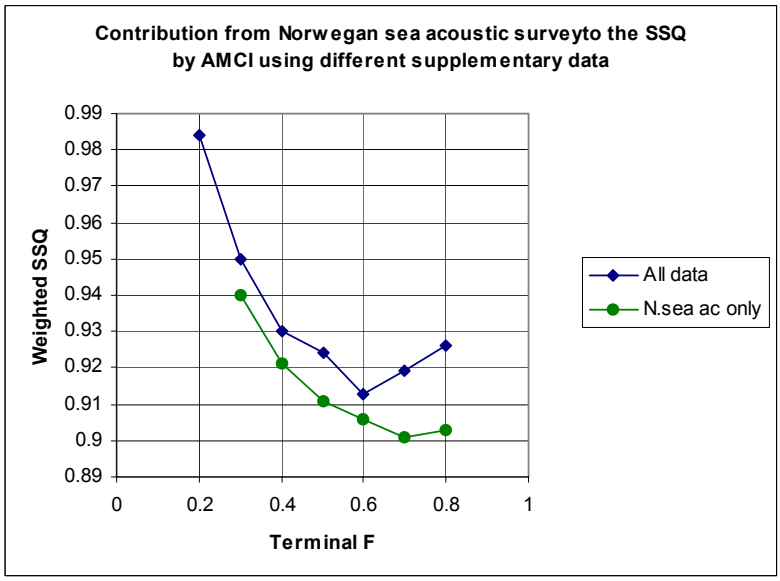


Figure 8.4.6 AMCI : Contribution of different sources to the total SSQ of the AMCI mimimization process using scanning over terminal F values.

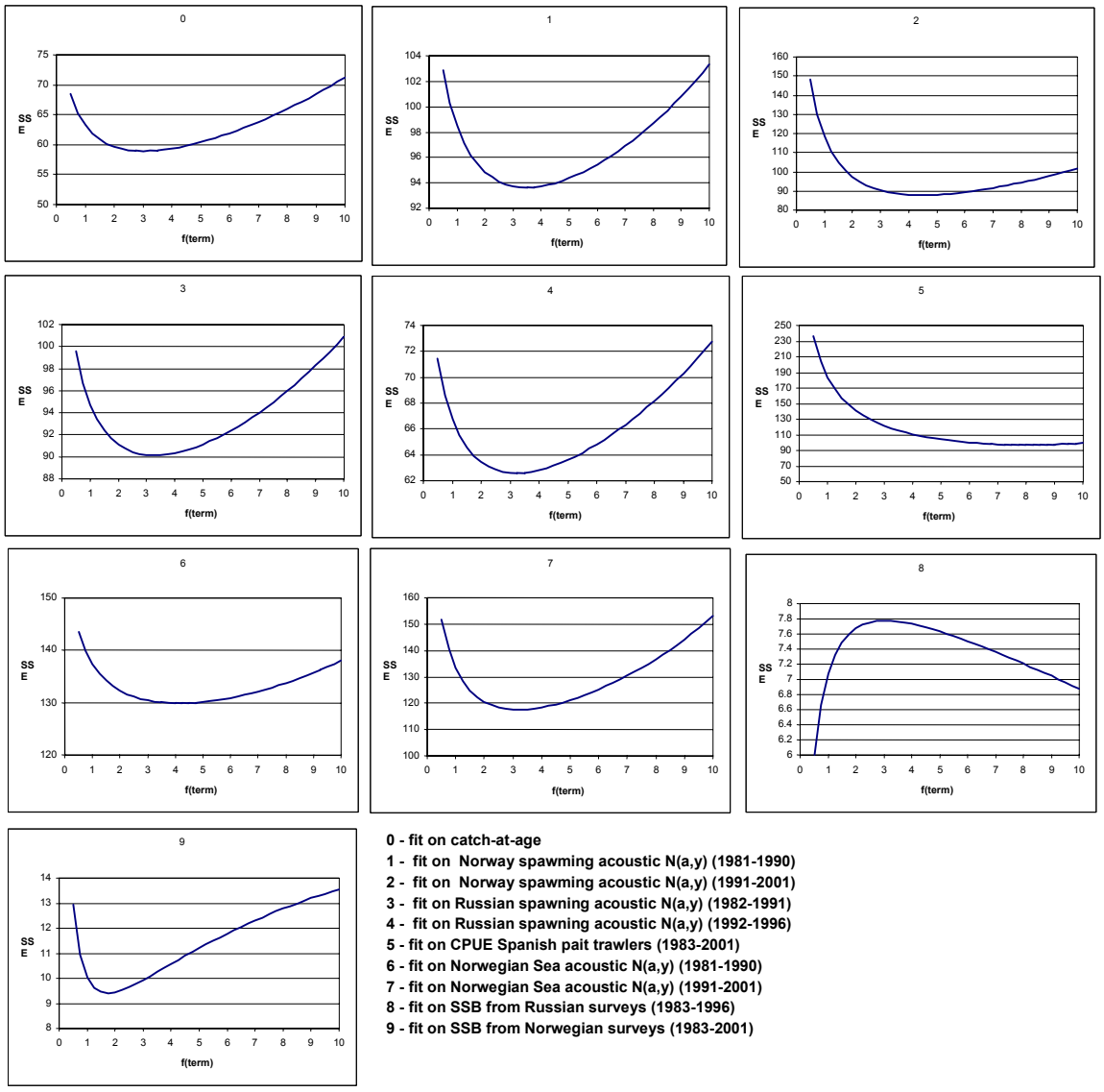


Figure 8.4.7 ISVPA SSE surfaces in effort-controlled (strictly separable) model with restriction of zero row and column sums for residuals in the separable representation of F compared to the VPA representation of F.

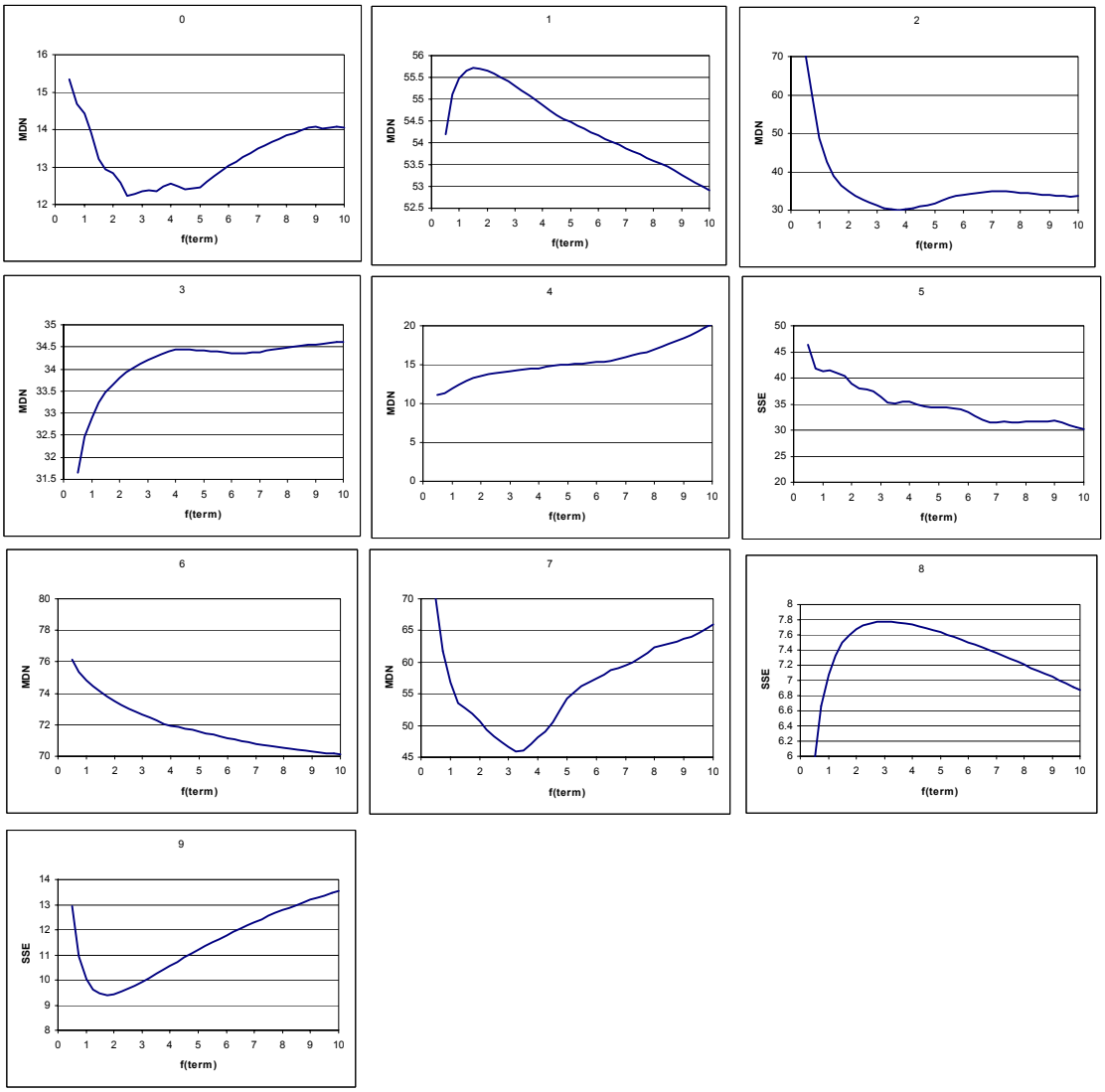


Figure 8.4.8 ISVPA Median Deviation (MDN) surfaces in effort-controlled (strictly separable) model with restriction of zero row- and column sums for residuals in the separable representation of F compared to the VPA representation of F. For explanation of index codes see Figure 8.4.7.

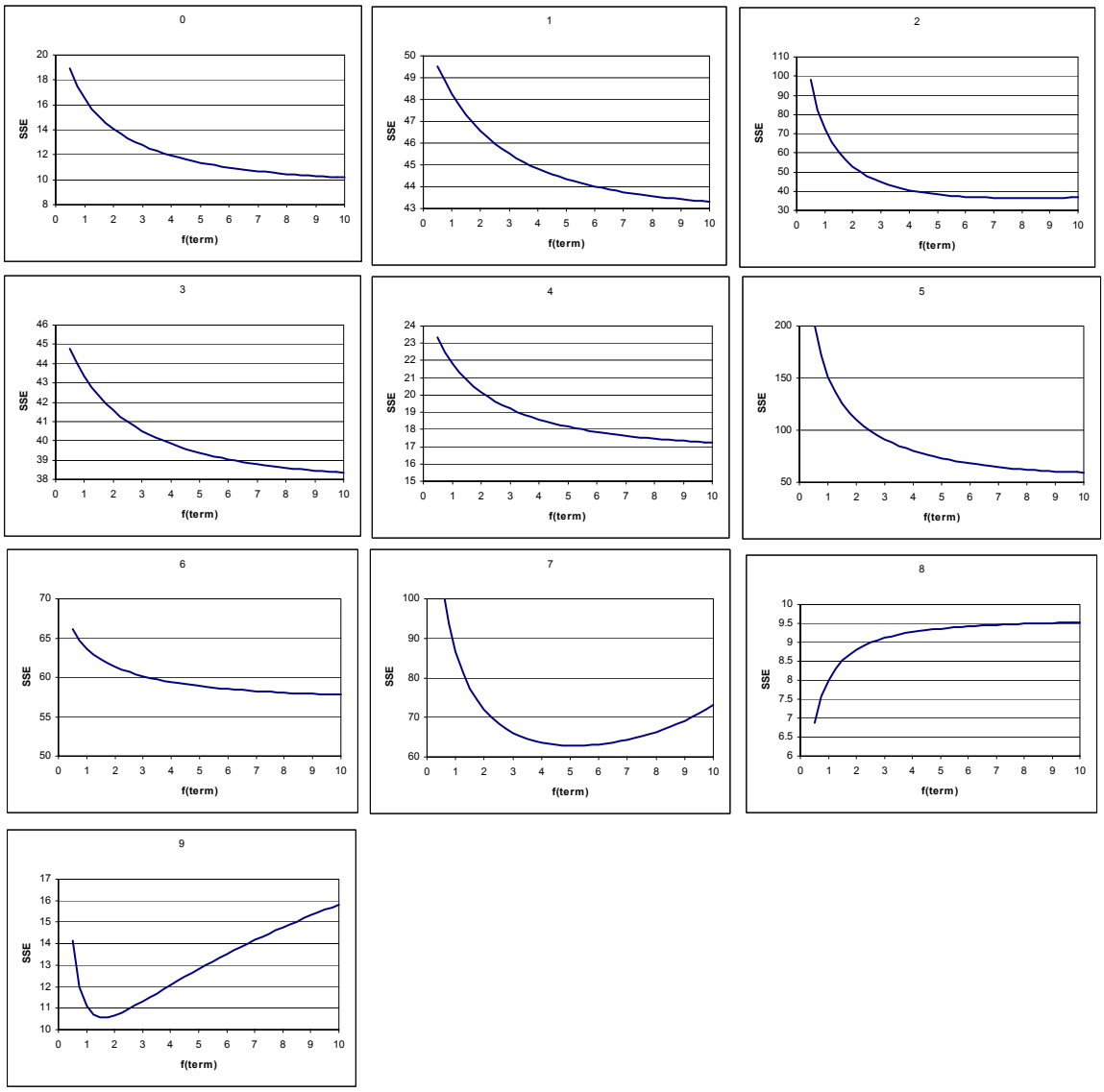


Figure 8.4.9 ISVPA SSE surfaces in catch-controlled (VPA type) model with restriction of zero row- and column- sums for residuals in the catch-at-age. For explanation of index codes see Figure 8.4.7.

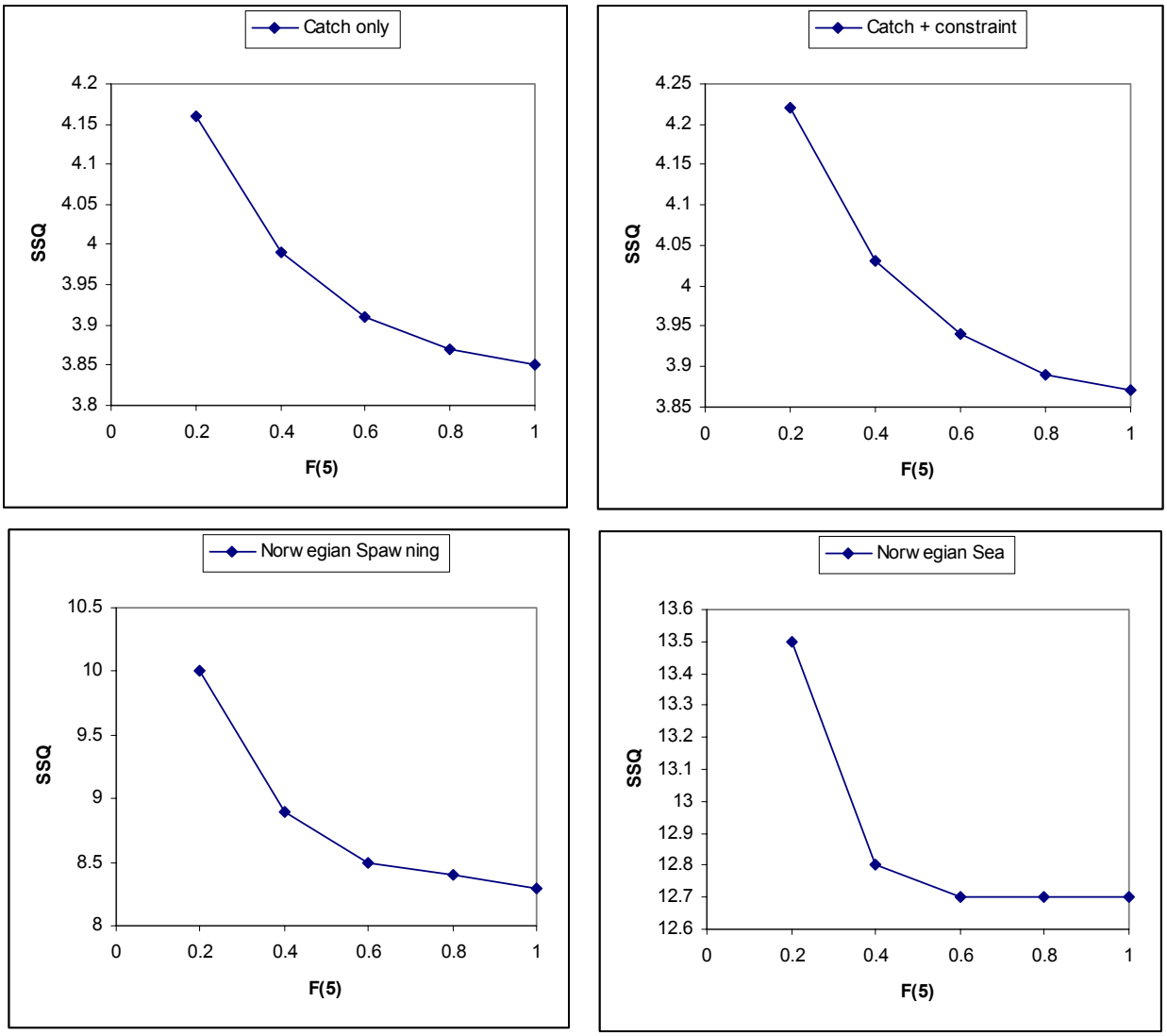


Figure 8.4.10 SSQ surfaces for ICA-like runs tuned to catch data only (top left-hand panel), catch data only but with additional constraints on the row-sums and column-sums of the log residual catch matrix (top right-hand panel), the Norwegian spawning acoustic survey and the catch data (bottom left-hand panel) and the Norwegian Sea acoustic survey and the catch data (bottom right-hand panel).

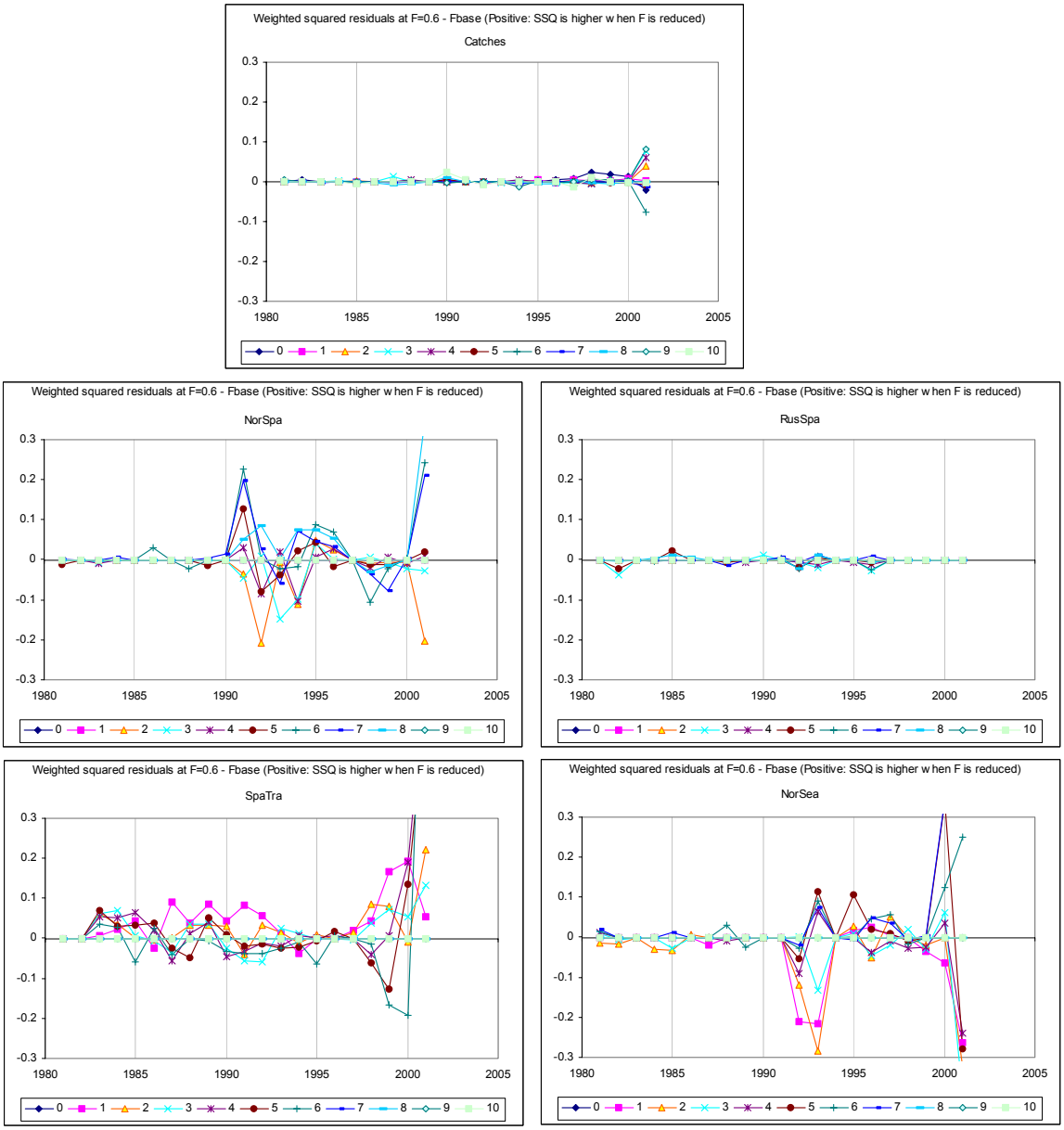


Figure 8.4.11 AMCI : sensitivity analysis of reducing terminal F to F=0.6 compared to the base line AMCI run. Changes due are expressed in terms of weighted squared residuals. A positive weighted squared residual indicates that the SSQ would increase if the terminal F were to be reduced.

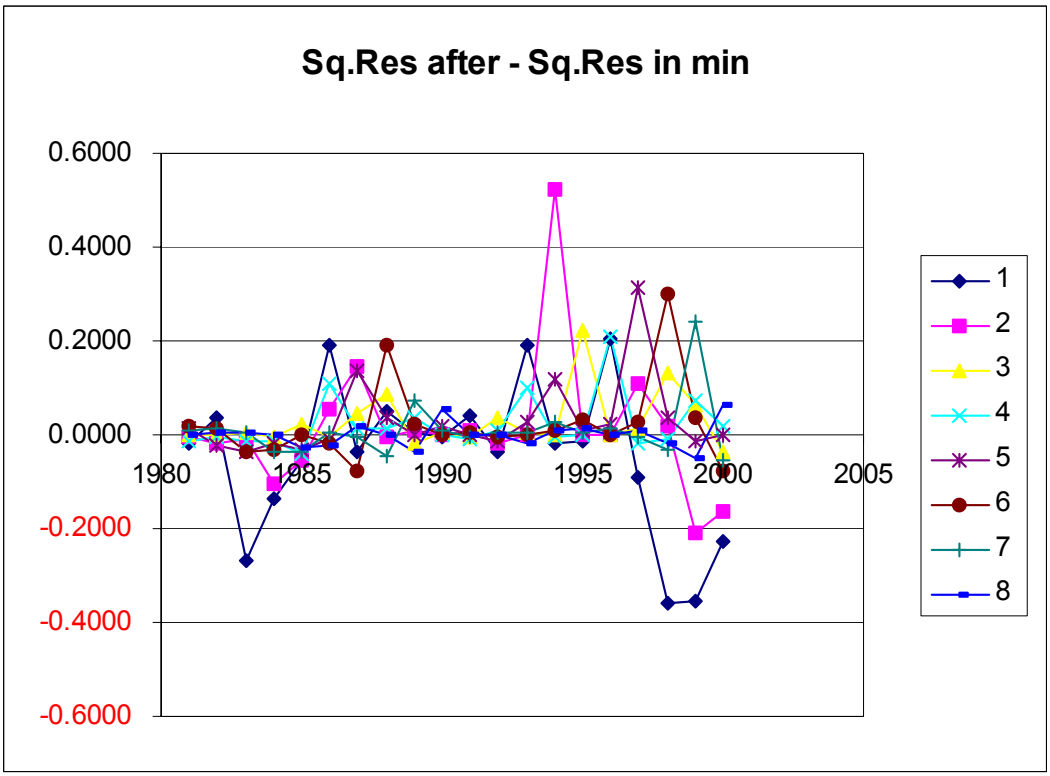


Figure 8.4.12 ISVPA: Sensitivity analysis. Effort-controlled with minimization of SSE. The graph shows the effect of increasing the terminal F compared to optimum F in ISVPA on the squared residuals by age and year. The 1992 year class is shown as positive values at subsequent ages which indicates that this year class 'would like' the fishing mortality to be lower (their contribution to the SSQ increases with higher F)

Figure 8.4.13 ICA : sensitivity analysis of reducing terminal F to F=0.6 compared to the base line AMCI run. Changes due are expressed in terms of weighted squared residuals. A positive weighted squared residual indicates that the SSQ would increase if the terminal F were to be reduced.

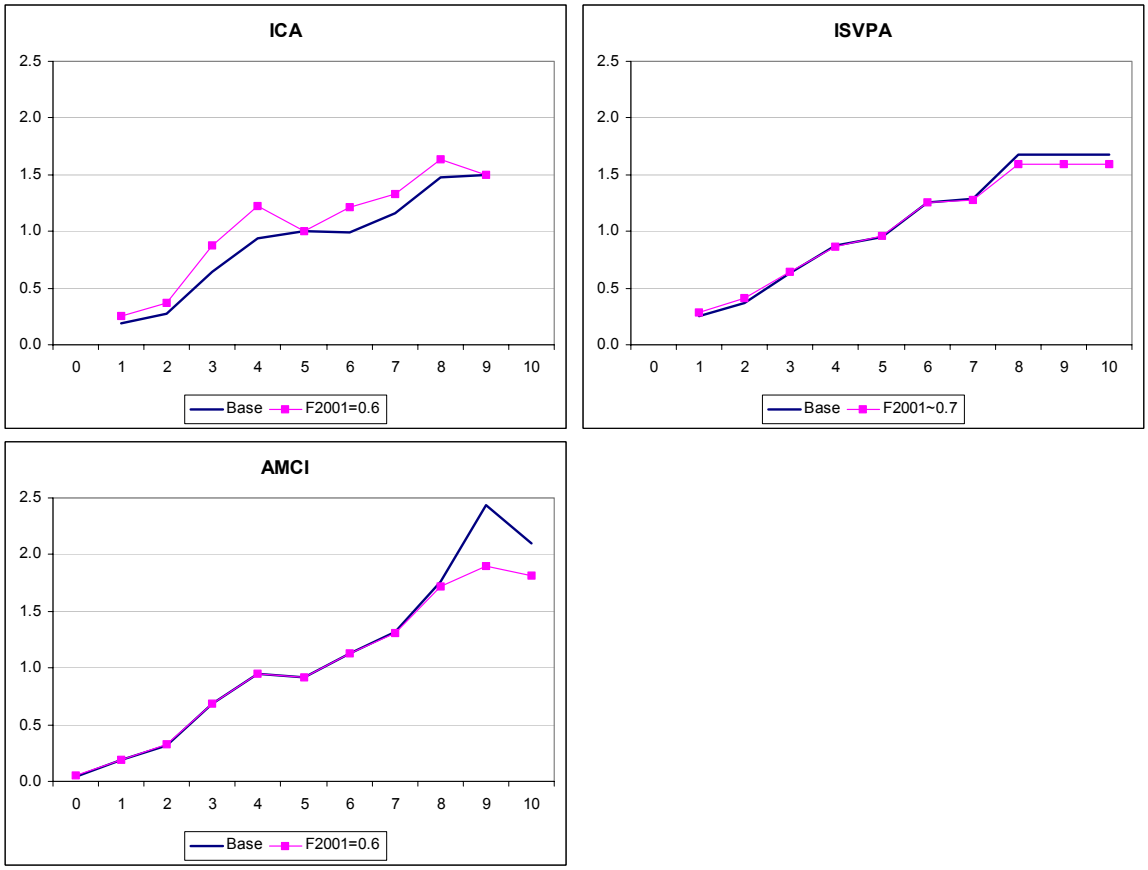


Figure 8.4.14 Selection patterns estimated for the final year by ICA, ISVPA, and AMCI, respectively.

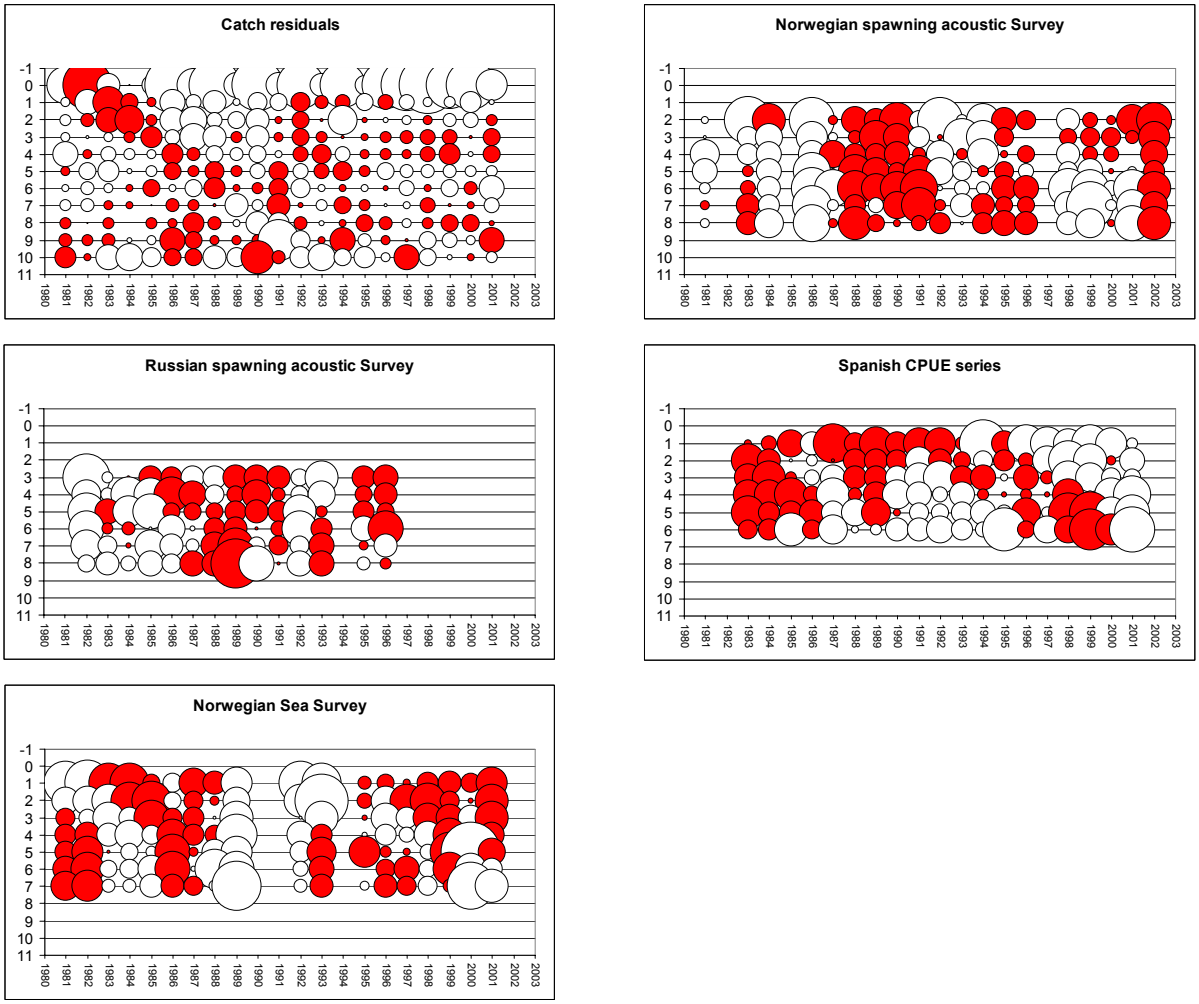


Figure 8.4.15 Catchability residuals estimated in AMCI for the catch and the tuning fleets (all together).

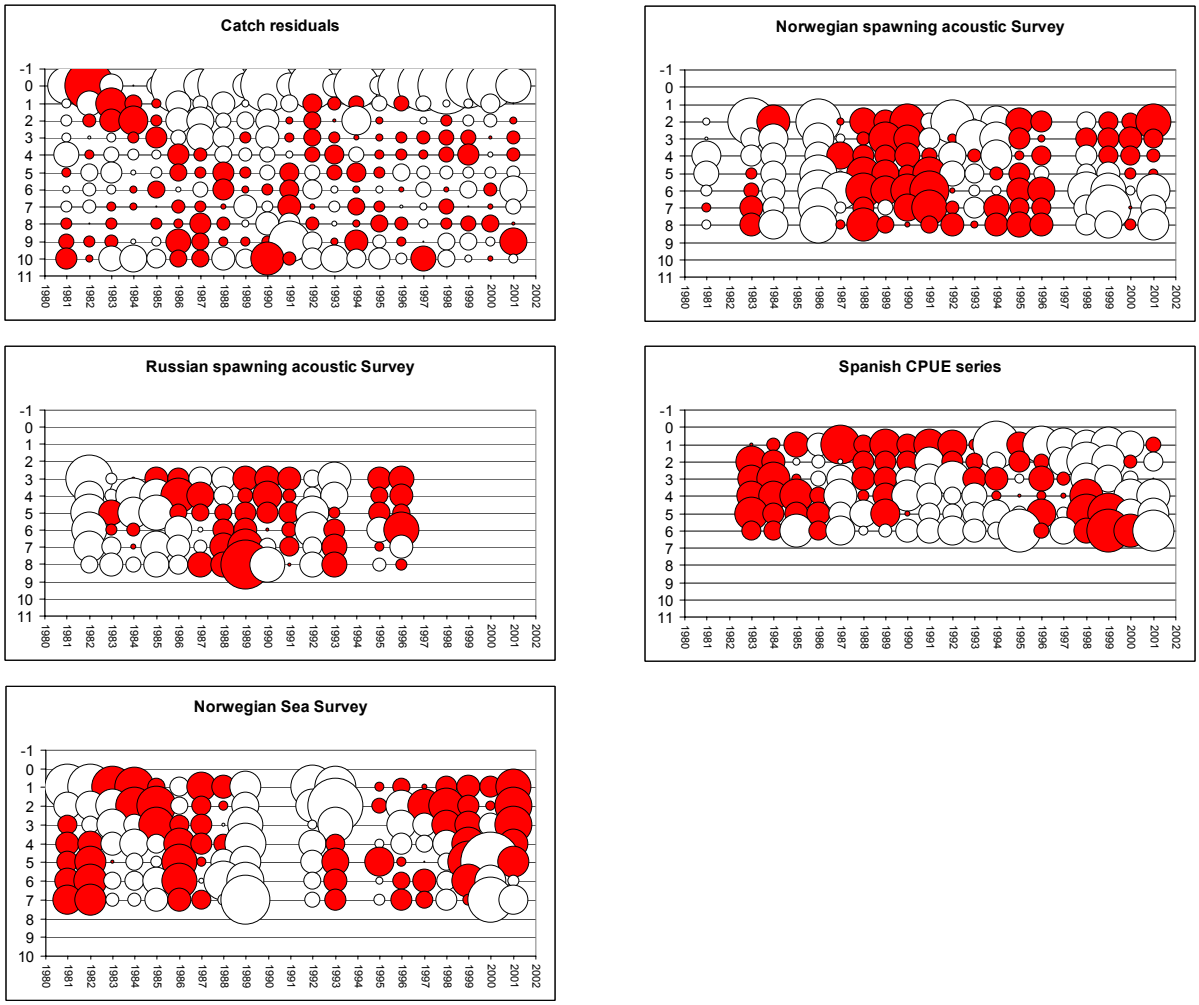


Figure 8.4.16 Catchability residuals estimated in AMCI for the catch and the tuning fleets (the year 2002 is not considered).

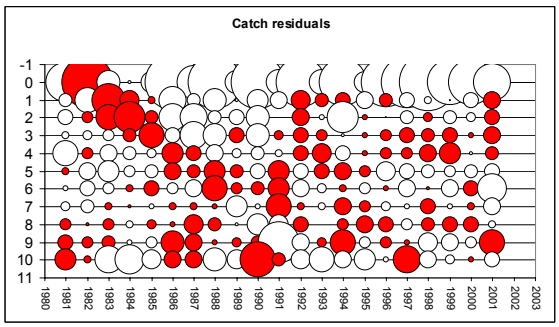


Figure 8.4.17 Catchability residuals estimated in AMCI just for the catch.

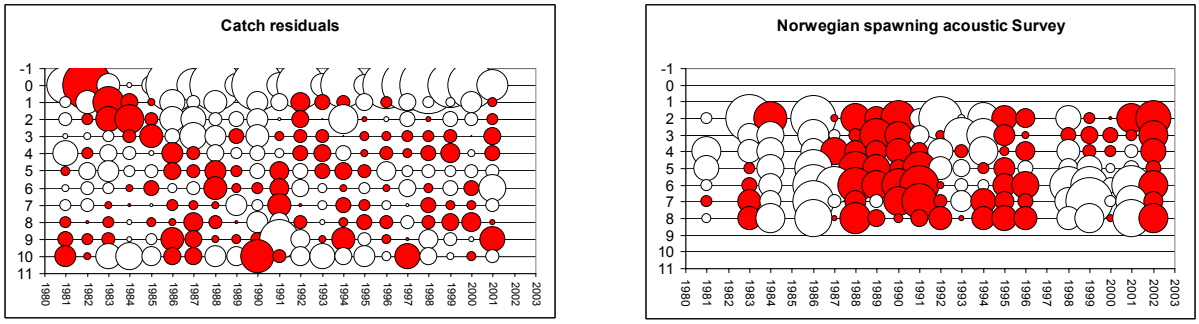


Figure 8.4.18 Catchability residuals estimated in AMCI for the catch and the Norwegian spawning acoustic Survey.

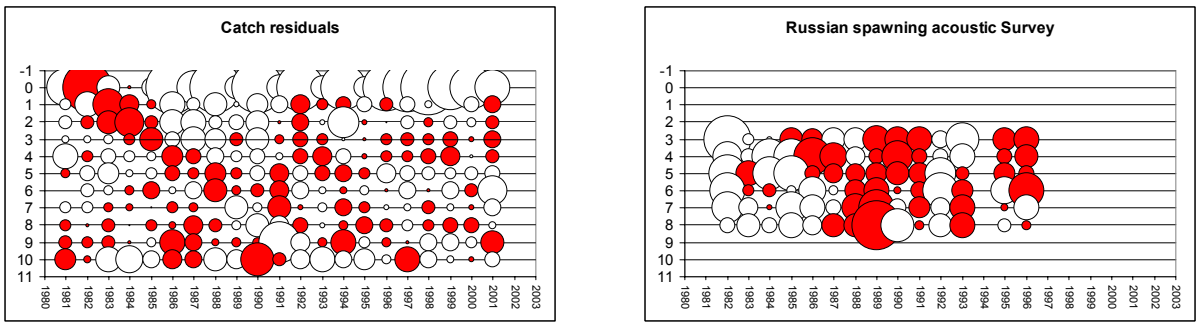


Figure 8.4.19 Catchability residuals estimated in AMCI for the catch and the Russian spawning acoustic Survey.

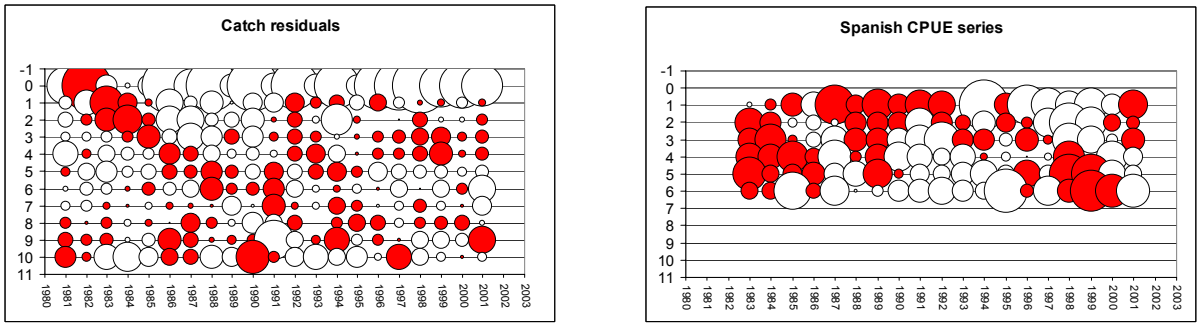


Figure 8.4.20 Catchability residuals estimated in AMCI for the catch and the Spanish CPUE series.

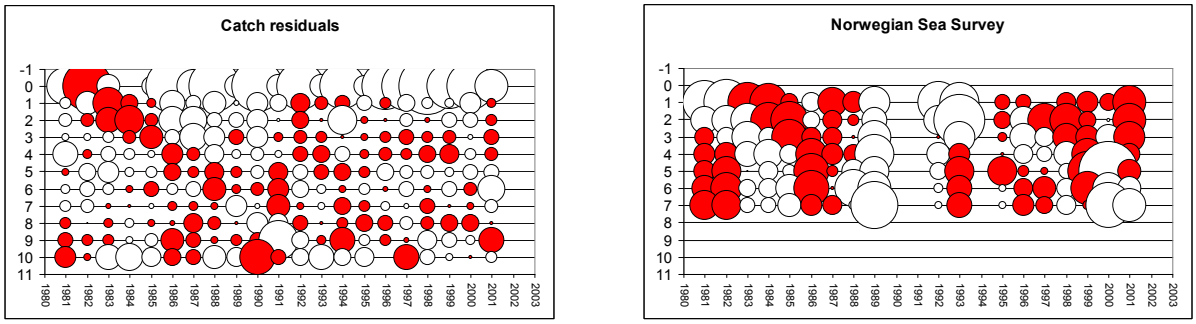


Figure 8.4.21 Catchability residuals estimated in AMCI for the catch and the Norwegian Sea Survey.

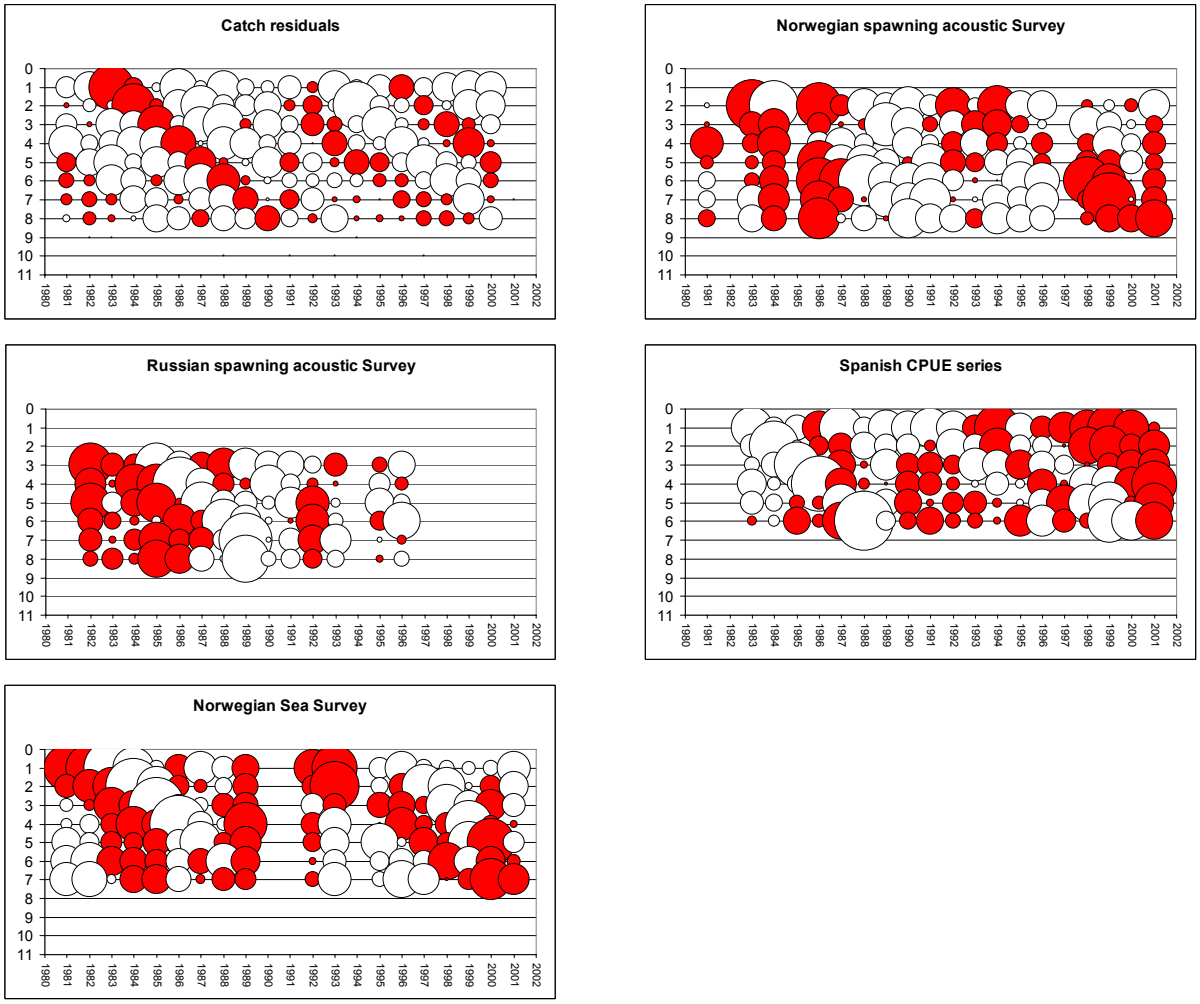


Figure 8.4.22 Catchability residuals estimated in ISVPA global minimum for catch-at-age.

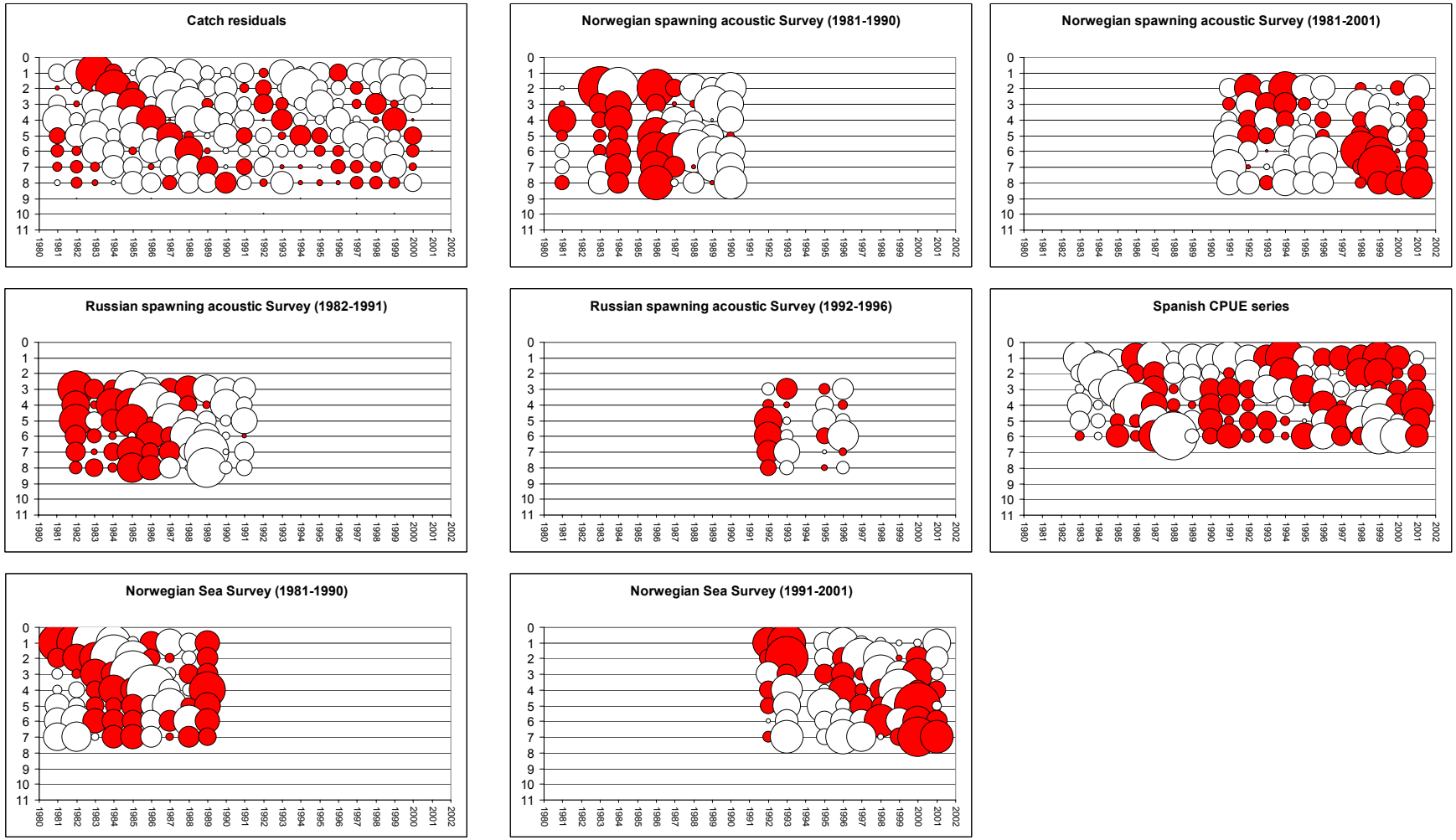
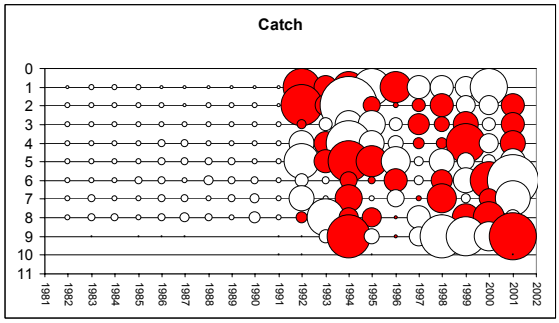
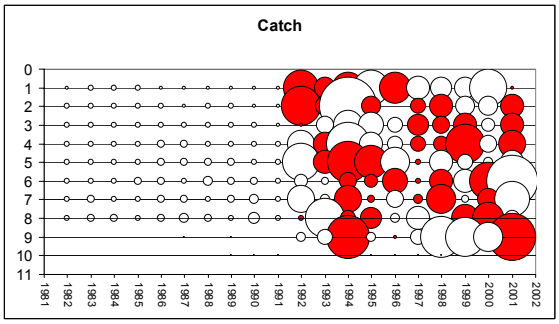


Figure 8.4.23 Catchability residuals estimated by ISVPA for each fleet and for each minimum when used with catch-at-age.

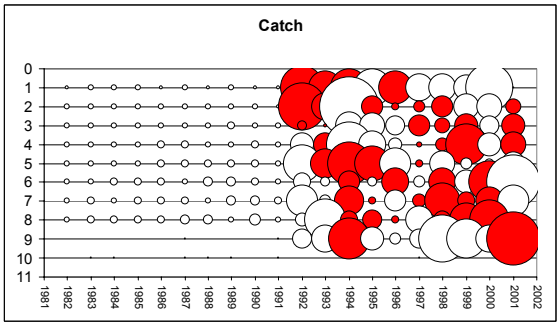
a) catch only.

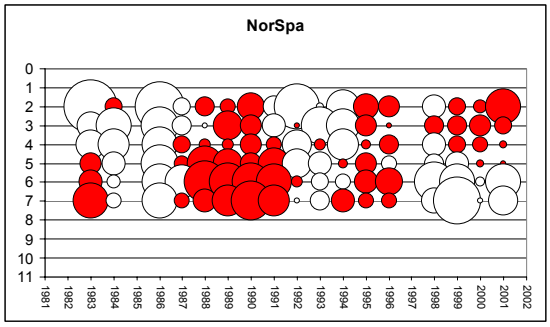


b) catch only with constraints on column-sums and row-sums of the residual matrix.

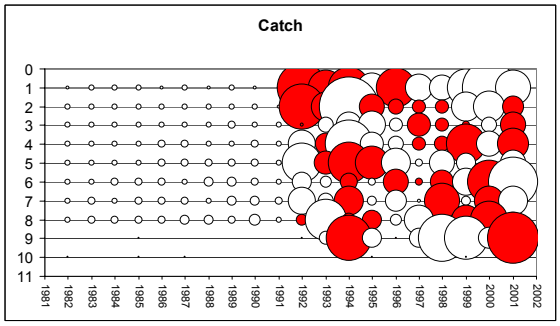


c) catch and Norwegian Spawning Acoustic survey only.





d) catch and Norwegian Sea Acoustic survey only.



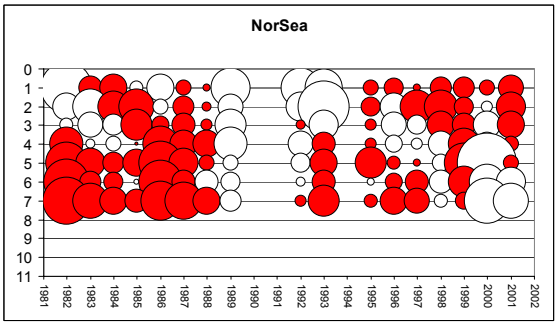


Figure 8.4.24 ICA residuals.

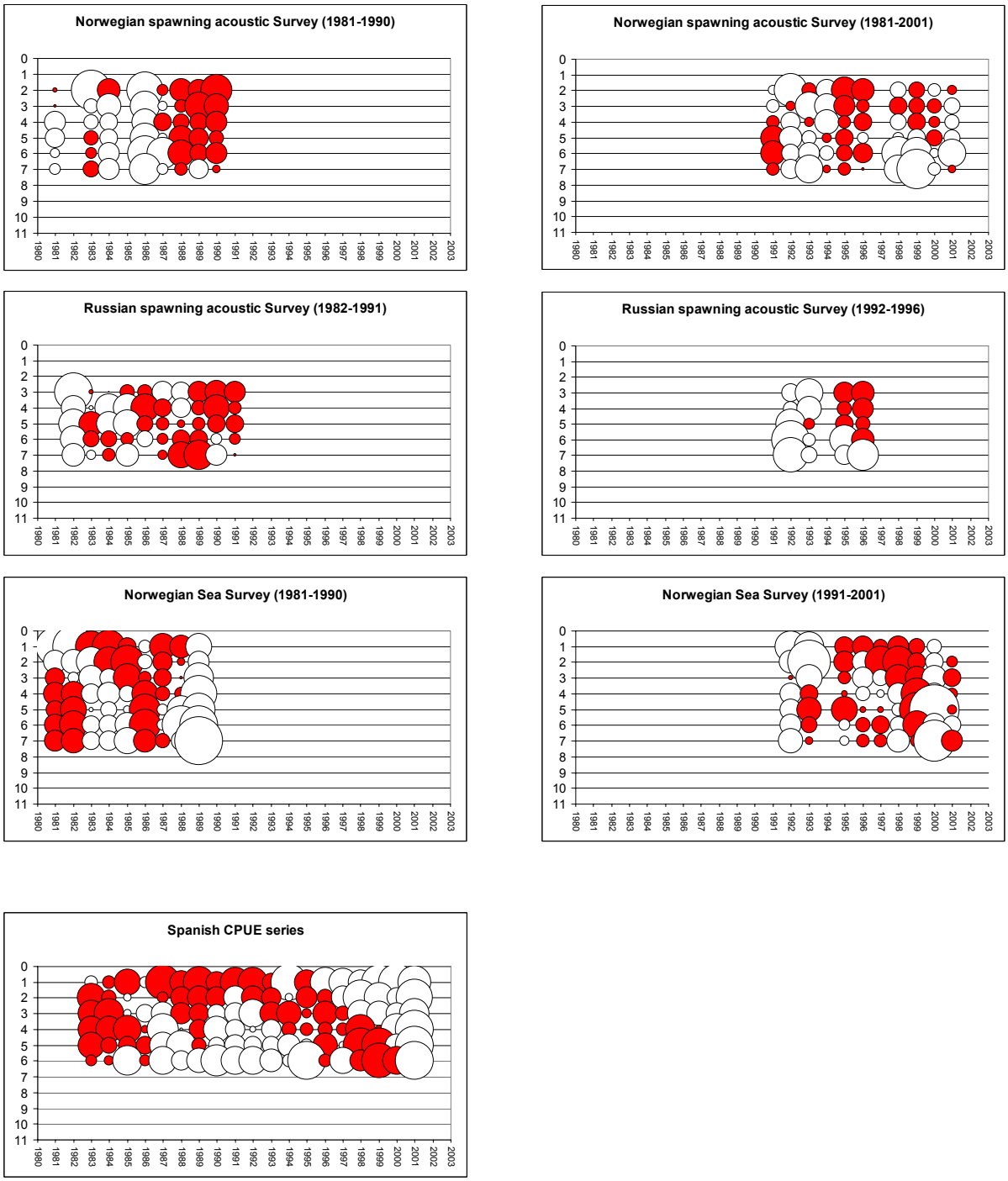


Figure 8.4.25 Catchability residuals estimated by XSA for all fleets.

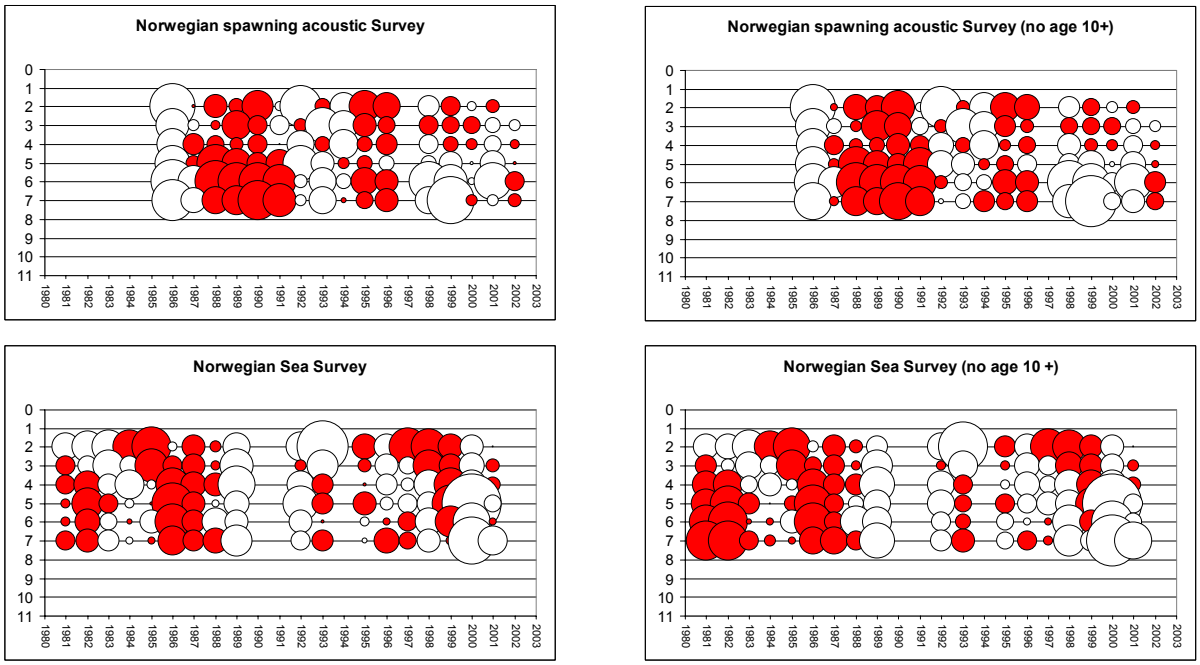


Figure 8.4.26 Catchability residuals estimated by CADAPT for each fleet with and without age 10-plus

.

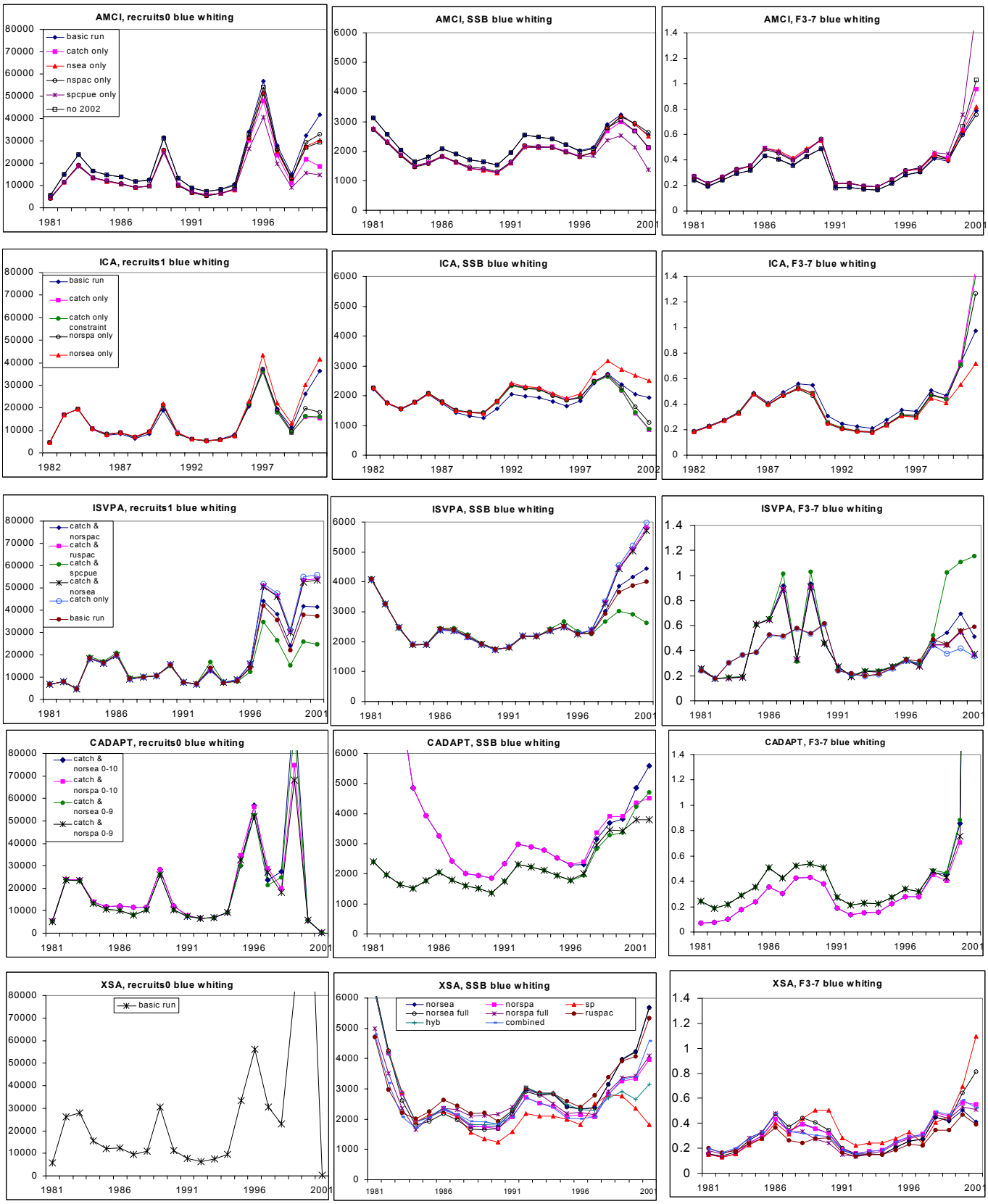


Figure 8.4.27 Comparison of assessments of the blue whiting stock by different methods

Abbreviations used in the figure legends: nsea = norsea = Norwegian Sea Acoustic survey; nspac = norspa = norspac = Norway Spawning Acoustic survey; spcpue = sp = Spanish Pair Trawl CPUE; no 2002 = without 2002 survey data; ruspac = Russian Spawning Acoustic survey; hyb = Laurec-Shephard hybrid model. The ISVPA F_{3-7} values for the basic run and for the catch only run are derived from the selection pattern.

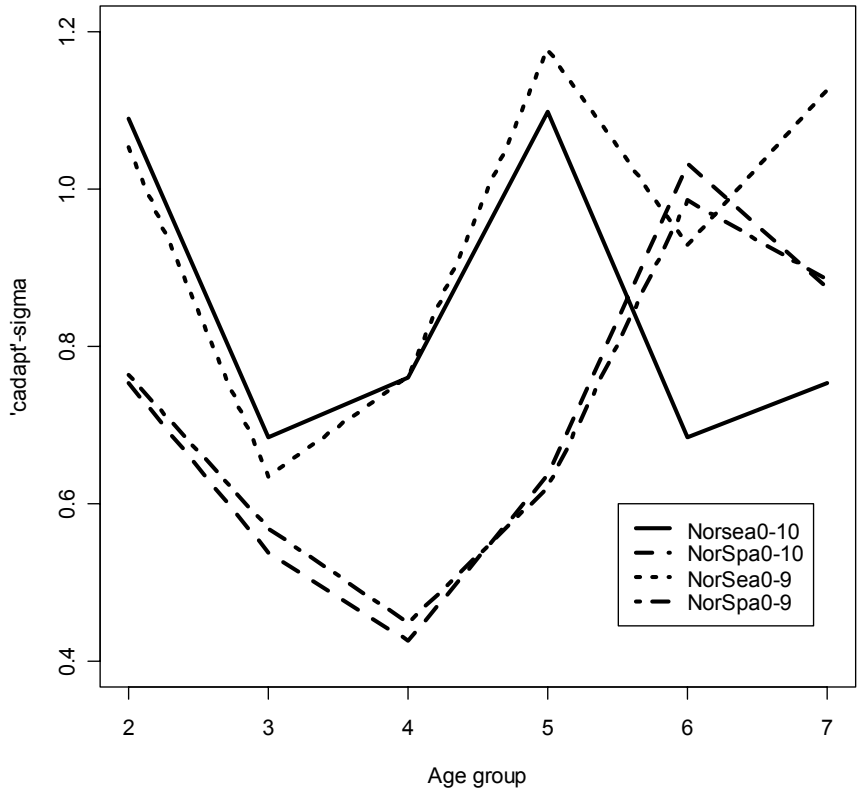


Figure 8.4.28 CADAPT Mean square log age group catchability residuals. Used in optimizations as inverse weights.

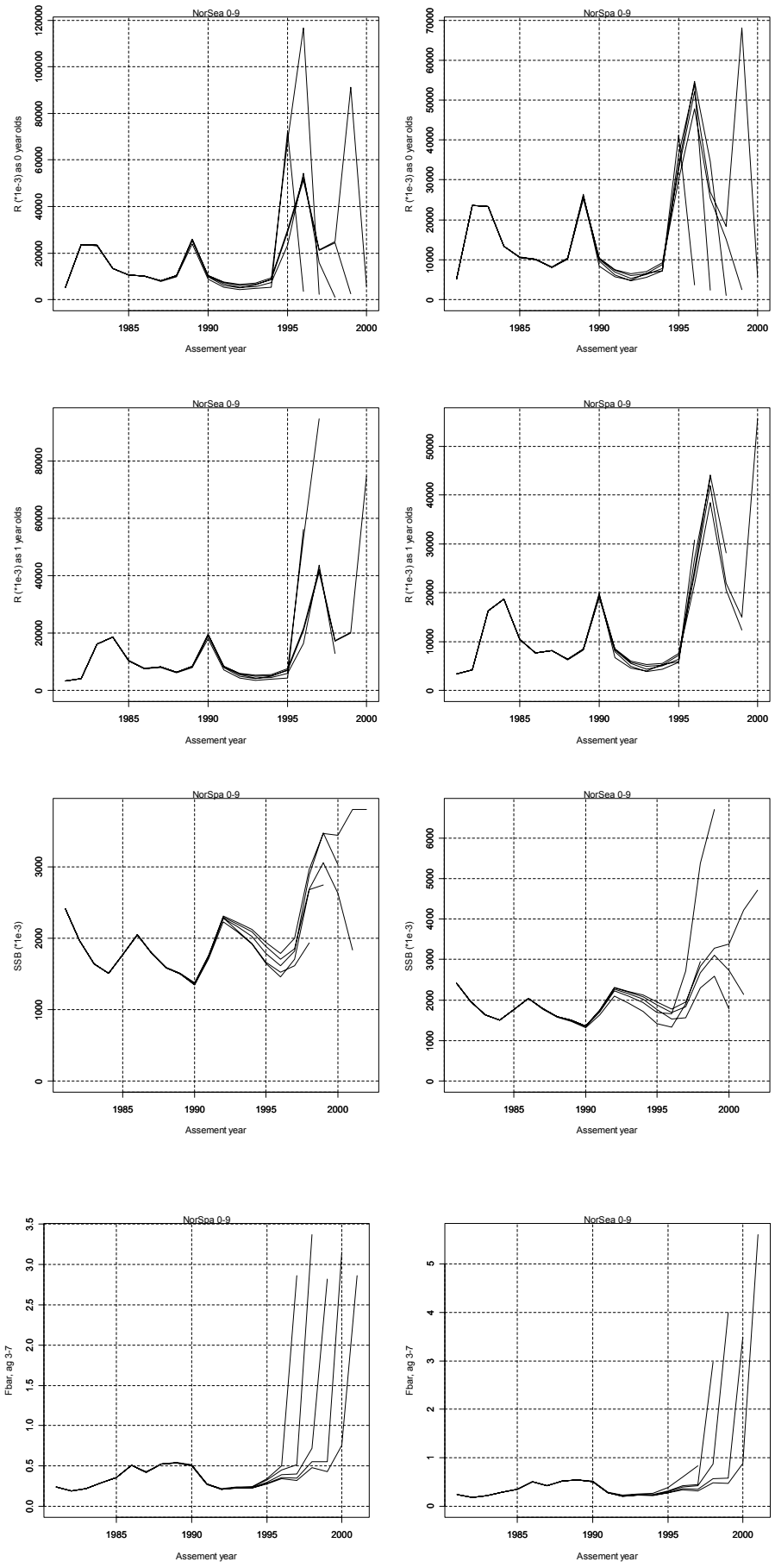
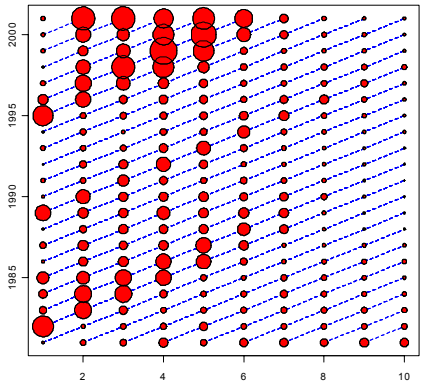


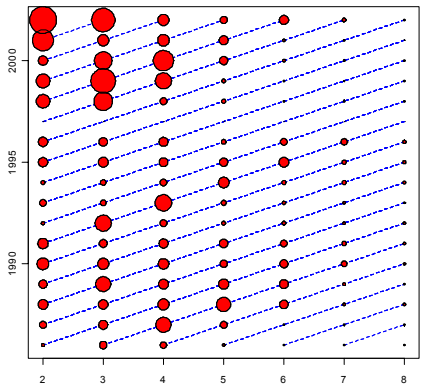
Figure 8.4.29 CADAPT Retrospective analysis of R as 0, R as 1, SSB and mean F_{3-7} for blue whiting runs 5-8.

a) blue whiting catch-at-age bubbles with area proportional to numbers

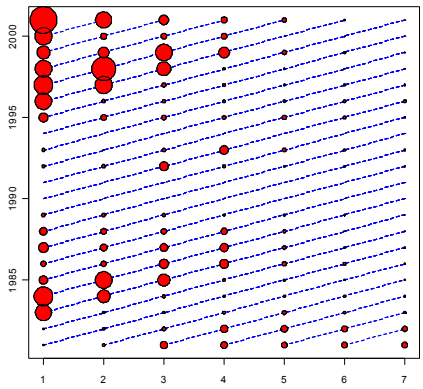


b) Index at age as bubbles proportional to index value:

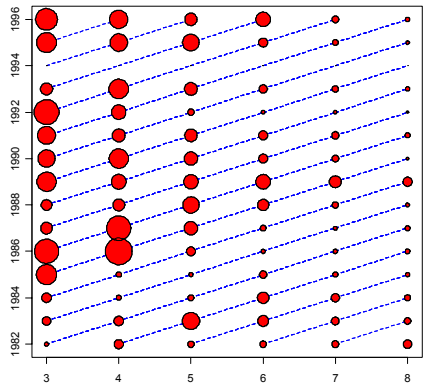
b1) Norwegian spawning acoustic survey



b2) Norwegian Sea survey



b3) Russian spawning acoustic survey



b4) Spanish CPUE series

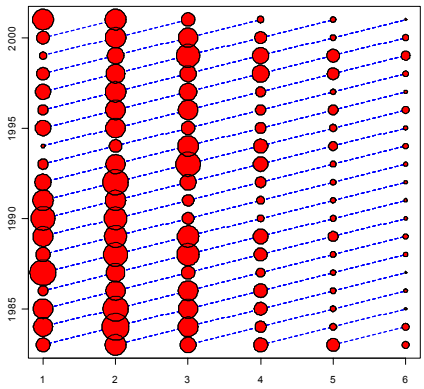


Figure 8.4.30 Additional diagnostics of the blue whiting catch and survey data.

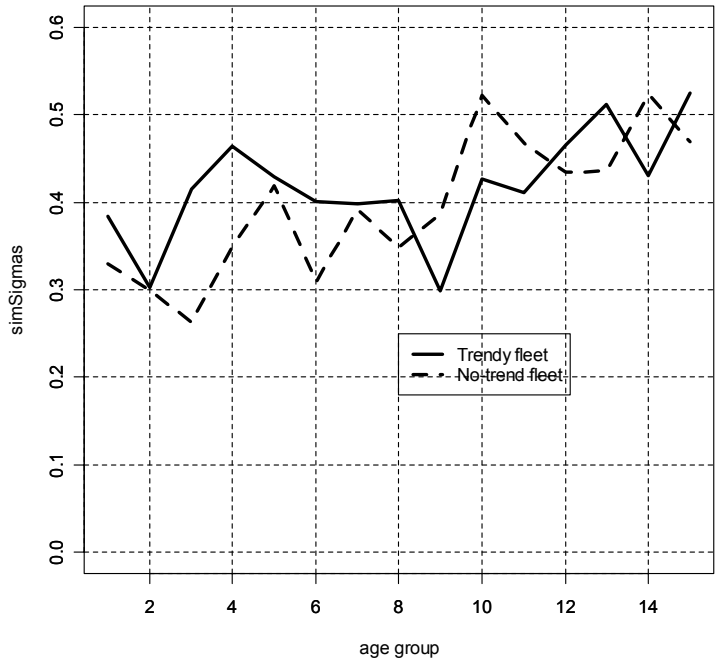


Figure 8.4.31 CADAPT Mean square log age group catchability residual on **simulated data**. Used in optimizations as inverse weights, unbroken line tuning with fleet with trend (sim_I), broken line fleet with no trend (sim_I2).

9 RECOMMENDATIONS AND FURTHER WORK

The Group has made a number of suggestions and recommendations 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 9.1.

Amongst other topics, the Group had attempted to address the urgent issue of the retrospective problem in stock assessments but it could be anticipated, in advance of this second meeting, that the problems of ICES' assessments would not be fixed at short notice. However, a way to proceed in the development of a solution has been proposed and this is reflected in the proposed ToRs for the next meeting in February 2004 (Section 9.2).

9.1 Suggestions and recommendations

Section 2. Model structure and data simulation

Sub-section 2.3. Data simulator

It would also be helpful to have a standardized data set which can be used for initial testing of models, so that any new model designs or formulations can be run against a single common data set.

It is therefore suggested that any final program and sample data sets, with accompanying documentation, should be placed on a freely accessible website, which could be hosted at ICES Headquarters.

Data sets should be constructed that violate the common assumptions in assessment models, and which replicate real-world situations.

Sub-section 2.5. Implementation

WGMG proposes that funding be sought to support the development of this system as stand-alone software influenced by knowledge gained from existing simulation models. In the shorter term it may be possible to begin by writing simple pre-processor and output filters to use an existing simulation model, such as Gadget, as a population generator. The modular nature of these parts of the system would allow such a beginning to be expanded incrementally as specific problems need to be addressed.

It would also be useful to take a number of the currently available simulated data sets, and the corresponding data simulators, and place them on a web or ftp site. Along with the data sets themselves there should be a description of their characteristics and an indication of which data sets are appropriate for which sorts of problem. This repository could then be expanded to include the flexible data simulator system described here when this has been written.

To this end, WGMG proposes the following ToR for their next meeting:

> **To examine software capable of generating simulated data, and agree an initial suite of standard data sets for use in model-testing and evaluation that will be made generally available from the ICES website.**

Section 3. Specification of data sources

Sub-section 3.1. Simulated data without noise

It was found, however, that this was **insufficient to create a significant retrospective pattern in estimates** obtained with the various methods tried during this meeting (see Section 5).

Section 4. Software tools for stock assessment purposes

Sub-section 4.1. Testing, validation and certification of software

To alleviate a number of these concerns, it is proposed that WGMG work inter-sessionally before its next meeting to **draft guidelines on the formal procedures to be adopted by WGMG for the testing, evaluation and validation of software for use by ICES stock assessment Working Groups**.

Sub-section 4.3.1. Current and future developments to TSA

It would be beneficial to evaluate more precisely the response of TSA to rapid fishery changes, using the kind of simulated datasets described elsewhere in this report.

Section 5. Influence diagnostics for detecting deviations from model assumptions

Sub-section 5.4.2. Local influence diagnostics on exact simulation data with model mis-specification

A useful direction for future research is to investigate whether other diagnostics exist that could be incorporated into the local influence approach to detect mis-specifications.

Fortunately this is simple to do with LID's, and **is useful to investigate for the next meeting of the working group**.

Sub-section 5.5. Influence diagnostics to diagnose the cause of retrospective patterns

This finding supports the conclusion made in ICES (2002a) that the lack of a retrospective pattern cannot be taken as *proof*' that the model is valid.

Sub-section 5.6.6. Suggestion for further work

The WGMG propose the following ToR for their next meeting:

To investigate and implement statistical approaches that identify and quantify uncertainty due to conditioning choices in fish stock assessments.

Section 6. Alternative stock assessment methods

Sub-section 6.1.3. Discussion and further work

To this end, WGMG propose the following ToR for their next meeting:

To develop fishery-independent assessment methods, measures of uncertainty, and appropriate diagnostics, with particular attention to data-poor situations and the estimation of relative catchability.

In this context, we intend that "data-poor situations" should encapsulate cases where data are limited in either extent (such as elasmobranch and deep-sea fisheries) or reliability (such as catch data of low quality) or both.

Sub-section 6.2.2. Application to data with q trend

The lack of pattern is problematic since it indicates that, contrary to earlier expectations, **retrospective analyses are inadequate to detect the kind of model mis-specification assumed in this data set.**

Sub-section 6.2.4. Conclusions regarding CSA

In view of the robustness of the relative trends in stock abundance indicated by CSA, **this method offers ACFM a chance to give management advice when data are insufficient for VPA**; e.g. when catches in number are known but no time-series of reliable age data is available.

Not only can CSA be suggested as a substitute for VPA in such circumstances, but it is also a good candidate for analysing data from a different perspective and verifying VPA-based assessments.

Consistent with its recommendations in 1995, this meeting of **WGMG encourages its members to explore this method further**.

Sub-section 6.3. Detection of inconsistencies in different sources of information – applying Benford's law to fisheries stock assessment

Overall, the evidence from these papers for the law to be applied to model-derived quantities (e.g. catch-, survivors- and fishing mortality-at-age) is not as compelling as for fisheries data but it could, nonetheless, still be a useful component of quality assurance checks which screen large numbers of data sets.

Section 8. Special request on blue whiting and Norwegian spring spawning herring

Sub-section 8.4.6. Investigative exploration with CADAPT

Exploratory runs of CADAPT were carried out using the blue whiting assessment data.

Sub-section 8.4.7. Comparisons and conclusions

- Conflicting sources of information appear to present the main problem in the blue whiting assessment. The conflict in the **data sources** is handled differently by the different methods that have been applied to this stock (e.g. AMCI and ISVPA)

Furthermore, WGMG recommends that its members:

- **further explore the historical behaviour of ICA, given the observations of the simulated data with ICA.**
- **further investigate the use of VPA type models (e.g. XSA) that are independent of separable assumptions and may give alternative interpretations of the data next to the family of separable models as ICA, AMCI and ISVPA.**
- **explore the convergence behaviour of AMCI in the light of very shallow SSQ surfaces.**

Sub-section 8.5. Answer to the special request

The different assessment methods find very different estimates of stock size and exploitation rate in the most recent years. The auxiliary information is contradictory and does not lend itself to unique characterization of the stock development. Also, model mis-specification may contribute to the difficulty in assessing the state of the stock.

9.2 Future terms of reference

The Working Group on Methods on Fish Stock Assessments [WGMG] (Chair: C. O'Brien, UK) meet in Lisbon, Portugal from 11-18 February 2004 to:

a) examine software capable of generating simulated data, and agree an initial suite of standard data sets for use in model-testing and evaluation that will be made generally available from the ICES website;

b) investigate appropriate diagnostics that detect model mis-specification in fish stock assessment;

c) investigate and implement statistical approaches that identify and quantify uncertainty due to conditioning choices in fish stock assessment;

d) review, revise and adopt draft guidelines on the formal procedures to be adopted by WGMG for the testing, evaluation and validation of software for use by ICES stock assessment Working Groups; and

e) develop fishery-independent assessment methods, measures of uncertainty, and appropriate diagnostics, with particular attention to data-poor situations and the estimation of relative catchability;

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

10 WORKING DOCUMENTS AND BACKGROUND MATERIAL PRESENTED TO THE WORKING GROUP

10.1 Working papers and documents (W)

ToRs (A and B)

WAB1
Cadigan, N.G. and Farrell, P.J.. Local influence diagnostics for the retrospective problem in sequential population analysis.

ToR (B)

WB1
Jónsson, S.T., Hjörleifsson, E. and Björnsson, H.. Description of elementary resource assessment and prognosis tools ('cadapt' and 'camera').

WB2
Mesnil, B.. Catch-Survey Analysis (CSA): a very promising method for stock assessment, particularly when age data are missing or uncertain.

ToR (C)

WC1
Needle, C.L.. Survey-based assessments with SURBA.

WC2
Maxwell, D.L. and Dunn, M.R.. Is Benford's law applicable to fisheries landings data?

WC3
Azevedo, M.. Looking out for No.1 in fisheries data: catch, survivors and fishing mortality-at-age.

ToR (D)

WD1
Darby, C.. Inconsistencies in the North Sea cod short-term and medium-term forecasts.

ToR (E)

WE1
Tretyak, V.L.. A model of recruitment of the commercial stock of the Northeast Arctic cod.

ToR (F)

WF1
Fryer, R.. Proposed developments to TSA.

WF2
De Oliveira, J., Roel, B., Darby, C. and O'Brien, C.. Comments on AMCI Version 2.2.

WF3
Vasilyev, D.A.. Description of the ISVPA.

10.2 Background material (B)

ToR (A)

BA1
Cadigan, N.G. and Farrell, P.J. (2002). Generalized local influence with applications to fish stock cohort analysis. *Appl. Statist*. **51**:469-483.

BA2
Patterson, K.R. (2002). Exploring and quantifying structural uncertainty in age-structured fish stock assessment: an approach based on a "kernel" survivors analysis. ICES CM 2002/V:10.

ToR (B)

BB1
Eero, M. (2003). Comparison of different stock assessment models using data on the Icelandic summer-spawning herring (*Clupea harengus*) stock. Draft M.Sc. dissertation, The United Nations University, Iceland.

ToR (D)

BD1
Skagen, D.W. and Aglen, A.. Evaluating precautionary values of fishing mortalities using long-term stochastic equlibrium distributions. Working document to ICES SGPA, December 2002.

ToRs (E and H)

BEH1
Punt, A.E., Smith, A.D.M. and Cui, G. (2002). Evaluation of management tools for Australia's South East Fishery 1. Modelling the South East Fishery taking account of technical interactions. *Marine and Freshwater Research* **53**:615-629.

BEH2
Punt, A.E., Smith, A.D.M. and Cui, G. (2002). Evaluation of management tools for Australia's South East Fishery 2. How well can management quantities be estimated? *Marine and Freshwater Research* **53**:631-644.

BEH3
Punt, A.E., Smith, A.D.M. and Cui, G. (2002). Evaluation of management tools for Australia's South East Fishery 3. Towards selecting appropriate harvest strategies. *Marine and Freshwater Research* **53**:645-660.

11 REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *In* Proceedings of the 2nd international symposium on information theory, pp. 267-281. Ed. by B.N. Petrov and F. Csàki. Akadémiai Kiadó, Budapest.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society* **78**:551-572.

Cadigan, N.G. & Farrell, P.J. (2002). Generalized local influence with applications to fish stock cohort analysis. *Applied Statistics* **51**: 469-483.

Cadrin S.X. (2000). Evaluating two assessment methods for Gulf of Maine northern shrimp based on simulations. *J. Northw. Atl. Fish. Sci*. **27**:119-132.

Cadrin S.X., S.H. Clark, D.F. Schick, M.P. Armstrong, D. McCarron & B. Smith (1999). Application of catch-survey models to the Northern shrimp fishery in the Gulf of Maine. *N. Am. J. Fish. Manag*. **19**:551-568.

Collie J.S. & G.H. Kruse (1998). Estimating king crab (*Paralithodes camtschaticus*) abundance from commercial catch and research survey data. *In* Proceedings of the North Pacific Symposium on Invertebrate Stock Assessment and Management. *Edited by* G.S. Jamieson and A. Campbell. *Can. Spec. Publ. Fish. Aquat. Sci.* **125**:73-83.

Cook, R.M. (1993). The use of sensitivity analysis to quantify uncertainties in stock projections. ICES CM 1993/D:66.

Cukier, R.I., Levine, H.B. & Schuler, K.E. (1978). Nonlinear sensitivity analysis of multiparameter model systems. *Journal of Computational Physics* **26**: 1-42.

Darby, C. and Flatman, S. (1994). Lowestoft VPA Suite Version 3.1: User Guide. MAFF: Lowestoft.

Fryer, R.J. (2002). TSA: is it the way? In Appendix D of the Report of the Working Group on Methods on Fish Stock Assessments. ICES CM 2002/D:01.

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

Hill, T.P. (1995). A Statistical Derivation of the Significant-Digit Law. *Statistical Science* **10(4)**:354-363.

ICES (1988). Report of the Workshop on Methods of Fish Stock Assessment, Reykjavik, Iceland, 6-12 July 1988. ICES CM 1988/Assess:26.

ICES (1993). 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 (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 (2002a). Report of the Working Group on Methods on Fish Stock Assessments, ICES Headquarters, Copenhagen, Denmark, 3-7 December 2001. ICES CM 2002/D:01.

ICES (2002b). Report of the Northern Pelagic and Blue Whiting Fisheries Working Group, Vigo, Spain, 29 April – 8 May 2002. ICES CM 2002/ACFM:19.

ICES (2002c). Report of the Working Group on the Assessment of Northern Shelf Demersal Stocks, ICES Headquarters, 27 August – 5 September 2002. ICES CM 2003/ACFM:04.

ICES (2002d). Appendix document to the Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak, ICES Headquarters, 11-20 June 2002. ICES CM 2003/ACFM:02.

ICES (2002e). Report of the Study Group on the Incorporation of Process Information into Stock-Recruitment Models, Lowestoft, UK, 14-18 January 2002. ICES CM 2002/C:01.

ICES (2003a). Report of the Study Group on Growth, Maturity and Condition in Stock Projections, ICES Headquarters, Copenhagen, Denmark, 5-10 December 2002. ICES CM 2003/D:01.

ICES (2003b). Report of the Study Group on Biological Reference Points for Northeast Arctic cod, Svanhovd, Norway, 13-17 January 2003. ICES CM 2003/ACFM:11.

ICES (2003c). Report of the Study Group on the Further Development of the Precautionary Approach to Fishery Management, ICES Headquarters, Copenhagen, Denmark, 2-6 December 2002. ICES CM 2003/ACFM:09.

Matthews, R. (1999). The power of one. New Scientist, 10 July 1999, pp.27-30.

McCullagh, P. and Nelder, J.A. (1983). Generalized linear models. Chapman & Hall, London.

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.

Needle, C.L. (2001). Some alternative assessments for 3Ps cod. Working Paper to the Newfoundland Regional Groundfish Assessment Meeting, St. John's, October 2001.

Needle, C.L. and Fryer, R.J. (2002). A modified TSA for cod in Division VIa: separate landings and discards. Working document to the ICES Advisory Committee on Fisheries Management, October 2002.

Newcomb , S. (1881). Note on the frequency of use of the different digits in natural numbers. *Amer. J. Math.* **4**:39-40.

Nigrini, M.J.(2000). Digital Analysis Using Benford's Law: Tests & Statistics for Auditors. Global Audit Publications.

O'Brien, C.M., Adlerstein S., and Ehrich S. (2000). Accounting for spatial-scale in research surveys: analyses of 2-year old cod from English, German and International groundfish surveys in the North Sea. ICES C.M. 2000/K:19.

Patterson, K.R. (2002). Exploring and quantifying structural uncertainty in age-structured fish stock assessment: an approach based on a "kernel" survivors analysis. ICES C.M. 2002/V:10.

Pope, J.G. and Shepherd, J.G. (1982). A simple method for the consistent interpretation of catch-at-age data. *Journal du Conseil International pour l'Exploration de la Mer* **40**: 176-184.

Prager, M.H. & MacCall, A.D. (1988). Sensitivities and variances of virtual population analysis as applied to the mackerel, *Scomber Japonicus*. *Canadian Journal of Fisheries and Aquatic Science* **45**: 539-547.

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., Stefánsson, G. (2000). Do different methods provide accurate probability statements in the short-term? ICES CM 2000/V:08, 18pp.

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.

Skagen, D.W. and Hauge, K.H. (2002). Recent development of methods for analytical fish stock assessment within ICES. *ICES Marine Science Symposia* **215**:523–531.

Stefánsson, G. and Palsson, O. K. (1998). A framework for multispecies modeling of Boreal systems. *Reviews in Fish Biology and Fisheries* **8**: 101-104.

Stratoudakis, Y., Fryer, R.J., Pierce, G.J. and Cook, R.M. (1997). Differences in life history features of long rough dab *Hippoglossoides platessoides* within Scottish waters. *Mar. Ecol. Prog. Ser*. **158**:303-306.

Vasilyev, D.A. (2001). Cohort models and analysis of commercial bioresources at information supply deficit. VNIRO Publishing, Moscow.

APPENDIX A - CATCH-SURVEY ANALYSIS (CSA) IN BRIEF

In essence, Catch-Survey Analysis (CSA) is an assessment method that aims at estimating absolute stock abundance given a time-series of relative abundance indices, typically from research surveys, by filtering measurement error in the latter through a simple two-stage population dynamics model known in the literature as the Collie-Sissenwine (1983) model.

CSA is one attempt to fill the gap between age-based methods, involving sequential population analyses (in the generic sense, including VPA) for estimation, and biomass dynamic (or surplus production) models. The former require relatively long time-series of catch-at-age data which are:

(i) generally expensive to process and more so as they need to be provided on a routine basis;

(ii) must be reliable over the whole age range (errors in some cells affect estimates for younger ages in each cohort); and

(iii) are simply unavailable for a number of species for which age determination is still an open question.

Surplus production approaches may not provide reliable estimates when variations in stock abundance are more influenced by changes in recruitment than by response to fishing intensity, but if information about recruitment strength is available, they cannot make use of it.

The starting point for CSA is that one is able to partition the survey data into two components (or stages). The younger stage comprises the recruits which should be from a single year class. This group can be identified based on actual age readings when feasible, or alternatively by counting animals below a distinct cut-off length class in the length composition of the survey catch; there may be exceptions but, in general, spotting the younger age group(s) out of length composition data is possible, notably when these are analysed in short time steps. Animals of all larger lengths or ages are accumulated into the second stage, called *fully-recruited*, which is akin to a super plus-group. In addition, one must provide an estimate of the total catch (i.e. fishery's removal) in number in each year. The main advantage of CSA is that the catch data need not be provided by age; however, if the reported catches are inexact for any reason, CSA is no better off than VPA, say: **CSA is not a fishery-independent method**.

Conser (1995a,b) presented the method to the 1995 meeting of the ICES Working Group on Methods on Fish Stock Assessments (ICES 1995) under the name *modified DeLury model*. The tool is regularly used, notably for shellfish assessments, in the Northwest Atlantic (e.g. Conser and Idoine 1992 ; Cadrin *et al*. 1999) and in the North Pacific (e.g. Collie and Kruse 1998; Zheng *et al*. 1997, 1998). Although the 1995 meeting of WGMG was very supportive of its broader use, it does not seem to have been used in ICES except, on a preliminary basis, by the *Nephrops* Working Group (ICES 2002b).

A1. The two-stage model

The essentials of the methods were already presented in Section 3.6 of ICES (1995), and are briefly recalled here.

Following Collie and Sissenwine (1983), it is assumed that the population consists of two distinct stages: the recruits (preferably a single year class), and the fully recruited animals. The time step is annual, with years defined either on a calendar basis or as the interval between regular surveys.

• Dynamics:

The population dynamics is described by a discrete difference model:

$$N_{t+1} = (N_t + R_t)e^{-M} - C_t e^{-(1-\tau)M} \qquad (1)$$

where

N_t : population size, in number, of fully recruited animals at start of year t,

R_t : population size, in number, of recruits at start of year t,

C_t : catches in number during year t (known),

M : instantaneous rate of natural mortality (assumed, same for both stages),

τ : fraction of the year when the catch is taken, typically 0 when fishing season is close to start of year, or 0.5 (*Pope-like*) when the catch is taken at mid-year or is evenly distributed throughout the year.

• Observations:

Estimating the time-series of N_t and R_t given the catches is the basic task of any assessment but, as with other methods, this requires additional information in the form of relative indices n_t and r_t of abundance for each stage, typically from surveys, which are assumed to be proportional to absolute population sizes N_t and R_t. The indices are deemed to be measured with some (log-normal) observation error:

$$n_t = q_n N_t \exp(\eta_t) \text{ ; t= 1, …, T} \qquad (2)$$

$$r_t = q_r R_t \exp(\delta_t) \text{ ; t= 1, …, T-1} \qquad (3)$$

where q_n and q_r : catchability coefficients of fully recruited and recruits, respectively, supposed to be constant over time; η and δ normally distributed random variables.

Whilst Collie and Sissenwine considered a unique catchability coefficient for both stages, Conser (1991) has added the specification that the catchability of the recruits is a fraction s of that of the fully recruited:

$$s = q_r / q_n \qquad (4)$$

In principle, s is estimable together with the other parameters but, in fact, it is strongly (negatively) correlated with q_n; in practice, it has to be fixed using external information and/or the sensitivity of results should be studied for a range of s values. If the information allows, year-specific values can be set for s.

A2. Parameter estimation

• Mixed-error:

All early works about this method considered fitting the model in the context of a mixed error structure, where process error in the dynamics in equation (1) is considered on top of observation (measurement) error on the indices. The process equation is arrived at by substituting equations (2) and (3) into equation (1) to give:

$$n_{t+1} = \left[\left(n_t + \frac{r_t}{s} \right) e^{-M} - q_n C_t e^{-(1-\tau)M} \right] . e^{\varepsilon_t} \qquad (5)$$

where ε : normally distributed process error, which here enters multiplicatively (for an additive process error, replace the exponential after the brackets by $+\varepsilon_t$).

Given time-series of catches ***C*** and of abundance indices ***r*** and ***n*** for T years, 2.T parameters have to be estimated: q_n + $\{n_1 \ldots n_T\}$ + $\{r_1 \ldots r_{T-1}\}$, using a set of T equations (2), T-1 equations (3) and T-1 equations (5) (i.e. 3T-2 residual terms, leaving T-2 degrees of freedom). Note that the catch and the recruitment index in the terminal year are not used.

In a least-squares approach, the objective function to be minimised is:

$$SS(\theta) = \lambda_\varepsilon \sum_{t=2}^{T} \varepsilon_t^2 + \sum_{t=1}^{T} \eta_t^2 + \lambda_\delta \sum_{t=1}^{T-1} \delta_t^2 \qquad (6)$$

where λ_ε and λ_δ are the (user defined) relative weights of the process error and of the observation error on recruits, relative to the observation error on the fully recruited, and $\boldsymbol{\theta}$ is the set of parameters. Increasing the process error weight λ_ε implies that there is less error in the description of the dynamics given by equation (1) than in the perception from the indices, and tends to smooth these indices so that they conform to equation (1) (in effect, it reduces the share of the process error in the overall SS). It is often difficult to find objective justifications for the choice of those weights and it is appropriate to evaluate the effects of changing them through a sensitivity analysis.

Minimising the non-linear function SS with respect to the parameters, using any suitable NLLS algorithms, yields the set of estimated parameters (noted with a hat), from which the time-series of population sizes can be reconstructed using the equations:

$$N_t = \frac{\hat{n}_t}{\hat{q}_n} \qquad \text{t= 1, …, T} \qquad (7)$$

$$R_t = \frac{\hat{r}_t}{s.\hat{q}_n} \qquad \text{t= 1, …, T-1} \qquad (8)$$

• Observation-error only:

There is a growing recognition that considering process error unduly complicates matters and that assuming observation-error only in fitting the model is preferable (Collie and Kruse 1998; see also Polacheck *et al*. 1993 in the context of surplus-production models). Experience so far indicates that estimates obtained with either route are fairly close anyway (e.g. Mesnil submitted).

This approach enables to estimate the absolute abundance R and N directly by minimising a variant of SS ((6)) where the first term is ignored. There are T+1 unknowns to estimate : q_n + $\{R_1 \ldots R_{T-1}\}$ + N_1, all other N's $\{N_2 \ldots N_T\}$ being derived by projection using (1). There are still T-2 degrees of freedom, but much less unknown parameters, which should make the estimation more robust. Furthermore, the abundance N of fully-recruited is *known* in each iteration and the corresponding indices are input data. Thus, the mean catchability can be computed:

$$q_n = \exp(mean(Log(n_t / N_t))) \qquad (9)$$

Here, we consider GM q consistent with the multiplicative error structure usually assumed in the former approaches, but note that trials with AM q show no difference. However, there are differences between NLLS estimates of q and computed values, and ensuing changes in stock size estimates. The advantage of computing q in this way is that it further reduces the number of parameters to search.

Contrary to the mixed-error structure, the all-observation-error approach does not require that abundance indices be available for all years. However, the effects of missing data have not been fully explored yet. In addition, it enables to make use of several sources of indices, but the issue of weighting these sources appropriately in the objective function still needs to be worked out.

• Stock trends:

Whichever error structure is considered, the procedure does not estimate recruitment or a recruitment index for the last year. The most recent recruitment is thus estimated with equation (8), but using the <u>observed</u> index (r_T).

Biomasses are derived in the usual way, by multiplying the population sizes by the observed mean weights in each stage and year, and summing. By analogy with an age-structured model, the fishing mortalities can be estimated by:

$$F_t = Z_t - M = Log\left(\frac{R_t + N_t}{N_{t+1}}\right) - M \qquad (10)$$

However, this quantity is not defined for the final year, which is often the year managers are most concerned with, and is at times not well behaved (negative values arise; these can serve as indicators of problematic estimates due to bad

choice of *s*). Perhaps, it is just as useful to consider the harvest rate (catch divided by stock abundance in each year, possibly differentiated by stage) as a measure of fishing intensity.

A3. Data needs

The basic data required by CSA are time-series of total catch in number, and of *survey* indices for the recruits and for the fully recruited. Data on mean weights for each stage are also required but are only used to translate stock numbers into biomass (they play no role in fitting the model). Natural mortality M must also be specified; the current implementation assumes equal M for both stages.

A4. Some implementation details

The method was coded in close accordance with the detailed description given by Conser (1995a), with some variants indicated below. There are essentially two versions, one allowing to choose between mixed process + observation or all-observation error structures, while the other one only considers the latter (note: this does not consist simply in setting the process error weight to zero; the parameters to estimate are radically different).

The user can choose between additive (as used by Cadrin) or multiplicative (used by Conser) process error (in equation (5)). A potential problem with the additive error is that its magnitude is out of scale with that of the observation errors in the contribution to the total SS.

Several options have been tried for the minimisation algorithm (Nelder-Mead Simplex, Marquardt, or simulated annealing). The Marquardt version[1] is by far the fastest and is suggested as default. However, it does not allow to set bounds on the parameters (which should all be positive) and does not provide the matrix of derivatives needed to compute approximate variances of the parameters. A simplex version is thus also offered in case this one converges to non-feasible solutions. No case of dramatic failure has been observed in trials so far, unless the values input for some parameters such as *s* are clearly inadequate.

A non-parametric model conditioned bootstrap is implemented as specified by Conser. However, confidence intervals are only provided for the annual biomasses and for the catchability coefficient (not for individual parameters). An estimate of bias is displayed for catchability only, and no bias correction is applied at this stage. The bootstrap summary table only takes into account runs in which convergence was achieved (not produced if less than 10 valid runs; the results of individual bootstrap runs can be saved to a file if desired). Two additional statistics based on the bootstrap are provided, giving the probability that the biomass in the last year is less than the average in the first 3 years in the time-series, and that it is less than the average in a recent period which is defined as T-4 to T-2.

Each run terminates with a retrospective analysis, with the restriction that the shortest time-series cannot be less than 10 years. The degree of retrospective discrepancy is measured with the summary statistic rho proposed by Mohn (1999).

All tabular layouts in the output file are formatted as comma separated fields (CSV) to facilitate import into a spreadsheet or, at the expense of some cutting/pasting, into S-PLUS or other packages. The output file starts with a reminder of the date, version, user's name and various settings as requested by ICES new rules.

Maximum dimensions are currently 25 for the number of years (i.e. 50 for parameters) and 500 for the number of bootstrap replicates.

The code is in FORTRAN 77 and is available either for a Unix/Sun platform or for a Watcom compiler on PC/Windows (differences only concern the specifics of file I/O). The all-observation-error version is also available as an S-PLUS 2000 script; this gives the same estimates as the FORTRAN program. An attempt was made to implement it in Excel with Solver, but it proved totally unreliable.

[1] It uses a set of subroutines developed by B.S. Garbow, K.E. Hillstrom and J.J. More in 1980 as part of the Minpack package obtained from the Netlib library at : http://www.netlib.org. The derivatives are computed by forward difference.

References

Cadrin S.X. (2000). Evaluating two assessment methods for Gulf of Maine northern shrimp based on simulations. *J. Northw. Atl. Fish. Sci*. **27**:119-132.

Cadrin S.X., S.H. Clark, D.F. Schick, M.P. Armstrong, D. McCarron & B. Smith (1999). Application of catch-survey models to the Northern shrimp fishery in the Gulf of Maine. *N. Am. J. Fish. Manag*. **19**:551-568.

Collie J.S. & G.H. Kruse (1998). Estimating king crab (*Paralithodes camtschaticus*) abundance from commercial catch and research survey data. *In* Proceedings of the North Pacific Symposium on Invertebrate Stock Assessment and Management. *Edited by* G.S. Jamieson and A. Campbell. *Can. Spec. Publ. Fish. Aquat. Sci.* **125**:73-83.

Collie J.S. & M.P. Sissenwine (1983). Estimating population size from relative abundance data measured with error. *Can. J. Fish. Aquat. Sci.* **40**:1871-1879.

Conser, R.J. (1991). A DeLury model for scallops incorporating length-based selectivity of the recruiting year class to the survey gear and partial recruitment to the commercial fishery. Research Document SAW 12/2. Appendix to CRD-91-03, Northeast Regional Stock Assessment Workshop Report, Woods Hole MA, 18 pp.

Conser, R.J. (1994). Stock assessment methods designed to support fishery management decisions in data-limited environments: development and application. PhD dissertation. School of Fisheries, University of Washington, Seattle, 292pp.

Conser R.J. (1995a). A modified DeLury modelling framework for data-limited assessments : bridging the gap between surplus production models and age-structured models. Work. Doc. to the ICES Working Group on Methods of Fish Stock Assessment, Copenhagen, February 1995, 85 p.

Conser R.J. (1995b). A Bayesian framework for the modified DeLury model with application to Atlantic surfclam. Work. Doc. to the ICES Working Group on Methods of Fish Stock Assessment, Copenhagen, February 1995, 58 p.

Conser R.J. & J. Idoine (1992). A modified DeLury model for estimating mortality rates and stocks sizes of American lobster populations. Research Document SAW 14/7, 28 p.

ICES (1995). Report of the Working Group on Methods of Fish Stock Assessments, Copenhagen, 6-14 February 1995. ICES CM 1995/Assess:11, 215 pp.

ICES (2002a). Report of the Working Group on Methods of Fish Stock Assessments, Copenhagen, 3-7 December 2001. ICES CM 2002/D:01, 102 pp.

ICES (2002b). Report of the Working Group on *Nephrops* Stocks, Lorient, 3-9 April 2002. ICES CM 2002/ACFM:15, 258 pp.

Mesnil, B. The Catch-Survey Analysis (CSA) method of fish stock assessment: an evaluation using simulated data. Submitted to *Fisheries Research*.

Mohn R. (1999). The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. *ICES J. mar. Sci*. **56**:473-488.

Polacheck T.R., R. Hilborn & A.E. Punt (1993). Fitting surplus production models: comparing methods and measuring uncertainty. *Can. J. Fish. Aquat. Sci.* **50**:2597-2607.

Zheng, J., Murphy, M.C., Kruse, G.H. (1997). Application of a catch-survey analysis to blue king crab stocks near Pribilof and St Matthew Islands. *Alaska Fish. Res. Bull.* **4(1)**:62-74.

Zheng, J., Murphy, M.C., Kruse, G.H. (1998). Abundance estimation of St. Matthew Island blue king crab using survey and commercial catch and effort data. In: Funk, F., Quinn II, T.J., Heifetz, J., Ianelli, J.N., Powers, J.E., Schweigert, J.F., Sullivan, P.J., Zhang, C.-I. (Eds), *Fishery Stock Assessment Models*. Alaska Sea Grant College Program Report No AK-SG-98-01, University of Alaska Fairbanks, pp. 575-589.

APPENDIX B - SURVEY-BASED ASSESSMENTS WITH SURBA 2.0

Cook (1997) described the results of applying a separable survey model called RCRV1A to survey data from six North Sea stocks, and showed (with some reservations) how survey data could be used to generate estimates of relative (not absolute) population trends. A version of this model was also described in the report of the 1995 WGMG meeting (ICES 1995), where it was called RCCPUE. Here we will describe briefly Cook's implementation, before summarising recent developments.

A separable model assumes that fishing mortality $\mathbf{F} = \lfloor F_{a,y} \rfloor$ is separable into an age effect $\mathbf{s} = \lfloor s_a \rfloor$ and a year effect $\mathbf{f} = \lfloor f_y \rfloor$, so that $\mathbf{F} = \mathbf{s} \times \mathbf{f}$. Here we will use the term *temporal trend* below to denote **f**, as this avoids confusion with "year-effects" in residual plots. Suppose that the abundance of a particular cohort declines exponentially from one year to the next, so that

$$N_{a+1,y+1} = N_{a,y} \exp(-Z_y), \tag{1}$$

and that the rate of that decline is given by

$$Z_{a,y} = F_{a,y} + M_{a,y} = s_a f_y + M_{a,y},$$

where $M_{a,y}$ is the natural mortality rate on age *a* during year *y*. Then if a cohort recruits to the stock in year *y* with recruiting abundance r_y, we can calculate its abundance at age *a* as

$$\begin{aligned} N_{a,a-1+y} &= r_y \exp\left(-\sum_{i=1}^{a-1} Z_{i,i-1+y} \right) \\ &= r_y \exp\left(-\sum_{i=1}^{a-1} s_i f_{i-1+y} + M_{i,i-1+y} \right). \end{aligned} \tag{2}$$

That is, the abundance at age *a* is given by the abundance at the recruiting age multiplied by the exponential of the sum of mortality rates in the intervening years. We will denote the vector of all recruiting abundances by $\mathbf{r} = \lfloor r_y \rfloor$.

In order to use relative abundance indices *a* to estimate relative stock size, we assume a time-invariant proportional relationship between stock size and the abundance index. This is given by

$$I_{a,y} = q_a N_{a,y},$$

where q_a is the catchability of the survey at age *a*. Thus, a survey for which the abundance index was a reliable indicator of stock size at age *a* would have $q_a = 1.0$, while it could be that $q_a = 0.0$ for a survey which will never catch fish of the age in question (a gillnet survey will never take very large fish, for example). Then we can rewrite Equation 2 as

$$I_{a,a-1+y} = \frac{q_r}{q_a} I_{r,y} \exp\left(-\sum_{i=1}^{a-1} s_i f_{i-1+y} + M_{i,i-1+y} \right), \tag{3}$$

where q_r and $I_{r,y}$ are respectively the catchability and the abundance index values for the recruiting age of the cohort.

This expression gives us a model for how the abundance index evolves through time for any given cohort. However, we must still estimate mortality rates, and to do this we use a variation of the standard catch equation:

$$I_{a,y} = \frac{F_{a,y} N_{a,y} \left(1 - \exp\left(-s_a f_y - M_{a,y}\right)\right)}{s_a f_y + M_{a,y}}. \qquad (4)$$

The RCRV1A model generates estimates for **s**, **f** and **r** by minimising the sum-of-squares difference between observed and fitted survey-derived abundance,

$$\mathrm{SSQ} = \sum_{a=1}^{A} \sum_{y=1}^{Y} w_a \left(\ln I_{a,y} - \ln \hat{I}_{a,y}\right)^2,$$

where A is the number of ages, Y is the number of years, and $\mathbf{w} = [w_a]$ are age-weighting factors. The minimisation is carried out assuming a lognormal error distribution. The progressive decline in cohort size is modelled using Equation 3, and mortality is estimated from the abundance-index catch equation (Equation 4).

However, the model as it stands is under-specified: since **s** and **f** are both estimated simultaneously, they can in theory take any number of nonsensical values. The RCRV1A solution to this is to fix the terminal value f_Y of the temporal trend in the terminal year, which is set so that the mean of all the temporal trends is 1.0: thus $f_Y = Y - \sum_{y=1}^{Y-1} f_y$. It is unfortunate that the terminal year is also the year of most interest to fisheries managers, and it may have been better to fix the first year instead. We must also provide a vector of catchabilities-at-age $\mathbf{q} = [q_a]$. There is currently no accepted method of determining empirically the catchability of a survey, so these values are usually chosen so as to ensure positive age effects. Summary statistics (total stock biomass, spawning stock biomass, yield) are calculated in the usual manner.

In the original analyses, Cook (1997) found that the model fit could be extremely sensitive to noise in the data. He therefore introduced a *smoother* λ, which constrains the minimisation by a penalty function:

$$\mathrm{SSQ} = \sum_{a=1}^{A} \sum_{y=1}^{Y} w_a \left(\ln I_{a,y} - \ln \hat{I}_{a,y}\right)^2 + \lambda \sum_{y=1}^{Y} \left(\frac{f_y}{f_{y-1}}\right)^2.$$

Finally, we note that estimates of fishing mortality rates F are obtained from Equation 1, which can be rewritten as

$$F_{a,y} = \ln\left(\frac{N_{a,y}}{N_{a+1,y+1}}\right) - M_{a,y}.$$

So our mortality rate estimates are derived by looking at the ratios of abundances. Since the number of ratios will always be one less than the number of abundances, we can only estimate A - 1 age effects and Y - 1 temporal trends.

SURBA 2.0 (SURvey-Based Assessment, version 2.0) is a recent development of RCRV1A. The basic method remains unchanged, but a Windows-based graphical user interface (GUI) with plotting capabilities has been added, and the following (more fundamental) modifications have been made:

1. Weights, proportion mature and natural mortality are read into the program as arrays, thus allowing variation through time as well as by age.

2. Estimation age weightings **w** may be entered manually. Alternatively, they can be calculated as the inverse of the variance of the survey index at that age, so that $w_a = \frac{n-1}{\sum_y \left(I_{a,y} - \bar{I}_a\right)^2}$ where n is the number of years in the survey time-series.

3. Problems arise if the model-fitting algorithm encounters zero index values. To avoid this, SURBA replaces such zero values with the lowest non-zero value at that age in the survey time-series, and reports that it has done so.

4. In RCRV1A, summary statistics (SSB, TSB, yield) were mean standardised before output. This mean-standardisation did not include the last year, although the last-year value was printed. In SURBA, mean standardisation is done over the full time-series. Furthermore, mean F is now calculated from the $F_{a,y}$ array, rather than from scaled selectivity vectors.

5. The new model now includes a simple, deterministic forecasting capability. This is done by rolling the survey-estimated population forward through time, assuming fixed geometric mean recruitment and the fitted temporal trends and age effects.

SURBA 2.0 has been used so far to produce supporting assessments at the Working Group on the Assessment of Northern Shelf Demersal Stocks (ICES 2002b), and at the forecast subgroup of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak (ICES 2002a).

References

Cook, R. M. (1997). Stock trends in six North Sea stocks as revealed by an analysis of research vessel surveys. *ICES Journal of Marine Science* **54**: 924–933.

ICES (2002a). Appendix to the Report of the Working Group on the Assessment of Demersal Stocks in the North Sea and Skagerrak. ICES CM 2003/ACFM:02.

ICES (2002b). Report of the Working Group on the Assessment of Northern Shelf Demersal Stocks. ICES CM 2003/ACFM:04.

APPENDIX C - WORKING DOCUMENT WAB1

Local Influence Diagnostics for the Retrospective Problem in Sequential Population Analysis

by

Noel Cadigan and Patrick Farrell

Abstract

The retrospective problem involves systematic differences in sequential population analysis (SPA) estimates of stock size or some other quantity in a reference year. The differences occur as successively more data are used for estimation. The differences appear to be structural biases that result from a mis-specification of the SPA. In some cases the retrospective problem is so severe that the SPA is considered to be too unreliable for stock assessment purposes. There are many possible causes of retrospective patterns, and it is usually difficult in practise to determine which causes are more likely. Local influence diagnostics are utilized to find small changes or perturbations to SPA input components such as catches or natural mortality that remove or reduce retrospective patterns. The plausibility of the perturbations can be used to assess the likelihood that the component is the source of the retrospective pattern. In this paper, the methods are applied to an example SPA for the eastern Scotian Shelf (ESS) cod stock, which has a severe retrospective pattern. We show that potentially reasonable errors in the assumptions about ESS cod catches, natural mortality, or SPA survey catchability could cause the retrospective pattern; however, additional information seems necessary to discriminate the more likely cause of the retrospective patterns. Our analyses suggest that reasonable changes in assumptions about model errors are not a likely source of the retrospective pattern.

1. Introduction

The retrospective problem in sequential population analysis (SPA) has received considerable attention in fish stock assessments. SPA is an analytical model of fishery catch data which can be used to estimate stock size. Based on a time series of annual catch numbers-at-age a in year y, $C_{a,y}$, SPA produces time series estimates of population numbers-at-age, $N_{a,y}$, and other derived quantities such as total biomass and spawner biomass (SSB). Once fishery catch statistics have been compiled in a given year to estimate $C_{a,y}$'s, an SPA can be performed on the updated catch time series to produce a new time series of $N_{a,y}$'s. The retrospective problem involves a systematic pattern in stock estimates as catches and other stock data are updated. More specifically, let $S_{y,t}$ denote an SPA estimate of stock size in year y based on data up to year $t \geq y$. A retrospective problem is said to exist if successive estimates $S_{y,t},\ S_{y,t+1},\ S_{y,t+2},\ ...$

deviate systematically in a decreasing or increasing trend. This problem can be so severe that the SPA is considered to be too unreliable for stock assessment.

The eastern Scotian Shelf (ESS) cod stock SPA provides a good example of the retrospective problem. This stock is found off the coast of Nova Scotia, Canada. Retrospective estimates of SSB are shown in Figure 1. The estimates cover the years $y = 1970, ..., t$, for $t = 1985, ..., 1994$, where t indicates the last year of catches and other fishery data used to estimate population size. Details about this SPA are provided in Section 3. Notice that the estimates of SSB for year y usually decrease as t increases and more data is used. For example, the estimate of SSB in 1985 based on data up to 1985 is 454 KT, whereas the 1985 estimate based on data up to 1994 is 172 KT, which is less than half of the 1985 estimate. This consistently decreasing trend indicates a structural bias in the SSB estimates caused by model mis-specification (Evans, 1996). The retrospective pattern shown in Figure 1 is severe, but not uniquely so. For example, a haddock stock considered by Sinclair *et al.* (1991) had a retrospective pattern of similar magnitude. Other stocks considered by Sinclair *et al.* (1991) also had retrospective patterns, although less severe than the pattern shown in Figure 1. Retrospective patterns are also present in recent stock assessments (e.g. see Figure 6 in Vaughan and Prager, 2002).

A common perception in stock assessments is that historic estimates of stock size are more accurate than recent estimates. This will tend to be true for stocks that are heavily exploited, and for which accurate catch data exists (Pope, 1972). Hence, when a retrospective problem exists like that in Figure 1, the common perception is that current stock size is over-estimated. This is important because current stock size and trends in recent stock size are required for fishery management decisions such as setting the total allowable catch (TAC) for next year. If current stock size is over-estimated then this may mean that the TAC will be set too high and not be sustainable. However, as pointed out by Sinclair *et al.* (1991), for certain types of SPA model mis-specifications the historic estimates may be less accurate than the current estimates. Mohn (1999) also demonstrated this using simulated data with model mis-specifications. He showed that certain types of model mis-specifications tend to compound over time such that adding more data produces more biased estimates of historic stock size. Even in these situations the current trends in stock size estimated by SPA may be overly optimistic and lead to a TAC that is too high.

Mohn (1999) presented simulation results to explore the types of retrospective patterns that might arise from various SPA model mis-specifications. He also presented *ad hoc* diagnostics to help discriminate between the possible sources of mis-specification that cause the retrospective problem for the ESS cod stock. One type of diagnostic involved examining the magnitude of perturbations to model components required to remove the retrospective pattern. Mohn (1999) considered simple perturbations that involved adding a common "effect" to part of a model

component. For example, he considered catch perturbations of the form

$$C_{a,y}(\omega) = \begin{cases} C_{a,y}, & y < y_o, \\ C_{a,y}\omega\phi_a, & y \geq y_o, \end{cases}$$

where $C_{a,y}$ and $C_{a,y}(\omega)$ are the observed and perturbed catch at age a in year y. The perturbation $\omega\phi_a$ was applied only after some specified year y_o. The magnitude of the perturbation was controlled by ω, while ϕ_a was an age effect that was fixed in all perturbations. This perturbation scheme was used to explore whether unreported discarding of catches starting in year y_o could be the source of the retrospective problem. The magnitude of the perturbation required to remove the pattern could be assessed to judge whether discarding was a plausible causal mechanism. Mohn (1999) considered perturbations to other model components as well. The "parameters" of his perturbation analyses were ω and y_o. Mohn (1999) profiled over these parameters to find values that removed the retrospective problem. He measured the retrospective problem using

$$\rho = \sum_{y=1985}^{1994} \frac{S_{y,y} - S_{y,1994}}{S_{y,y}}. \tag{1}$$

Recall that $S_{y,1994}$ is the estimate of stock size in year y based on all catch and other stock data up to 1994. If the retrospective estimates for year y based on data only up to year y (i.e. $S_{y,y}$) fluctuate randomly about $S_{y,1994}$ then ρ will be approximately zero. For ESS cod SSB, $\rho = 3.59$.

In this paper we also use perturbation analyses to diagnose more likely causes of the retrospective pattern for ESS cod. We improve upon Mohn (1999) by using more realistic perturbations; for example, we investigate catch perturbations of the form

$$C_{a,y}(\omega) = (C_{a,y} + \delta) \times \omega_{a,y}. \tag{2}$$

We add a small offset, $\delta = 0.1$, to the catch so that zero catches are perturbed. In (2) we perturb each $C_{a,y}$ separately, whereas Mohn (1999) considered more simple and arbitrary perturbations. We also find perturbations that remove the retrospective pattern. The advantage of searching over a higher dimensional perturbation space is the potential of finding smaller and more realistic perturbations to remove the retrospective pattern than those presented in Mohn (1999).

We use the local influence approach for perturbation analyses. This method is briefly described in Section 2, and more fully discussed in Cadigan and Farrell (2002). The local influence approach is computationally more convenient for perturbation analyses because it does not require re-estimation of the SPA over the very large number of corners in the perturbation space. This is possible because, as we show in Section 4, the perturbation surface of ρ around a

relevant neighborhood of the origin is usually quite linear. Similar to Cook (1986), we use basic concepts in differential geometry to study the effect of a perturbation on ρ. More specifically, we find the direction at the perturbation origin that results in the greatest reduction in ρ, and then examine perturbations in this direction that reduce ρ to zero. The directions are based only on the unperturbed SPA parameter estimates. When the perturbation surface of ρ is linear, our method finds the smallest perturbation that removes the retrospective pattern, as measured by ρ. In Section 4, we consider this for four distinct perturbation schemes on catches, mortality, survey catchability, and case weights in order to determine if the retrospective pattern in the ESS cod stock SPA is more likely caused by any of these components.

2. Local Influence Approach

Cadigan and Farrell (2002), hereafter referred to as C&F, considered local influence diagnostics for problems that involved estimating a $p \times 1$ parameter vector θ by maximizing a fit function $l(\theta)$ that has basic smoothness properties. The estimate of θ, denoted as $\hat{\theta}$, was the solution to

$$\dot{l}(\hat{\theta}) = \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0.$$

They perturbed model components with a $k \times 1$ perturbation vector ω and studied the influence of the perturbations on key model results. The perturbation ω was of the form $\omega = \omega(h) = \omega_o + hd$, where ω_o was the null perturbation, d was a fixed direction vector of length one, and h was a scalar that determined the magnitude of the perturbation. The dimension of k could be large; for example, in one of the SPA applications in C&F, k was equal to 520.

C&F measured influence on an important, but arbitrarily specified, scalar SPA result denoted as $g(\hat{\theta})$. They considered some basic geometric properties of the influence surface of $g_\omega(\hat{\theta}_\omega)$ versus ω near the origin, ω_o. The primary diagnostic used by C&F was the local slope in the direction d,

$$S(d) = \left.\frac{\partial g_\omega(\hat{\theta}_\omega)}{\partial h}\right|_{h=0} = d' \left.\frac{\partial g_\omega(\hat{\theta}_\omega)}{\partial \omega}\right|_{\omega=\omega_o} = d'\dot{g}_o.$$

A particularly interesting diagnostic was the direction of maximum slope,

$$s_{\max} = \dot{g}_o / \sqrt{\dot{g}_o' \dot{g}_o}. \tag{3}$$

Further computational details are given in Section 2.1 of C&F.

Some minor modifications of the methods in C&F are required to study local influence on ρ. The influence measure $g_\omega(\hat{\theta}_\omega)$ used by C&F involves parameter estimates from the optimization of a single fit function, whereas ρ is based on parameter estimates from multiple optimizations. For example, ρ for ESS cod is based on ten optimizations. Let $S_{\omega y,t}$ denote a perturbed estimate

of stock size. Using $S_{\omega y,t}$'s in (1) gives the perturbed value of ρ_ω. Note that $S_{\omega y,t}$ is a particular form for $g(\hat{\theta})$ using the notation in C&F. Let

$$\dot{S}_{o\,y,t} = d' \frac{\partial S_{\omega\,y,t}}{\partial \omega}\bigg|_{\omega=\omega_o}.$$

Each $\dot{S}_{o\,y,t}$ can be computed using the methods in C&F. The local slope of ρ_ω in the direction d is given by

$$\dot{\rho}_o = d' \left\{ \sum_y \left(S_{o\,y,Y} S^{-1}_{o\,y,y} \dot{S}_{o\,y,y} - \dot{S}_{o\,y,Y} \right) S^{-1}_{o\,y,y} \right\}, \tag{4}$$

where Y is the last year under consideration. Let $\dot{\rho}_{\max}$ denote the maximum local slope. The direction of maximum slope, $d_{\max}$, can be found using (3) with $\dot{g}_o$ equal to the $\{\cdot\}$ term in (4).

3. SPA for ESS cod

We used essentially the same SPA formulation for ESS cod as outlined in Mohn (1999). The cohort model we used was

$$N_{a,y} = \left(N_{a+1,y+1} e^{M/2} + C_{a,y} \right) e^{M/2}, \tag{5}$$

where $M = 0.2$ is the annual mortality rate due to sources other than the reported catch. The catches are presented in Fanning *et al.* (1995). Equation (5) can be used to estimate the $N_{a,y}$'s for all ages and years if the $N_{A,y}$'s (numbers at the oldest age A for all years) and the $N_{a,Y}$'s, or the survivors in the last year Y, are both known. The ESS cod SPA we considered covered the years 1970 to 1994, and ages 3 to 15; that is, $Y = 1994$ and $A = 15$. Following Mohn (1999), all of the $N_{A,y}$'s and some of the $N_{a,Y}$'s were approximated using assumptions about fishing mortalities.

The fishing mortality at age a in year y is defined as

$$F_{a,y} = \log(\frac{N_{a,y}}{N_{a+1,y+1}}) - M.$$

The $N_{A,y}$'s were constrained so that their fishing mortalities were proportional to the average for some range of younger ages. In effect,

$$N_{A,y} = \frac{C_{A,y} e^{M_{A,y}/2}}{1 - e^{-\alpha \bar{F}_y}}.$$

Note that $\bar{F}_y$ is a function of $N_{a,y+1}$'s, so the result of the F constraint was that the $N_{A,y}$'s for $y < Y$ were constrained to be functions of unknown survivors. Hence, the only unknowns in the cohort model were the $N_{a,Y}$'s. For convenience we replaced any zero catch with 0.1. For ESS cod α was set at 0.95 and $\bar{F}_y$ was based on ages 7 to 9. A similar F-constraint procedure

was used to fix the values for the survivors at ages 9 to 15. Their fishing mortalities were set equal to $F_{8,Y}$. Hence, the only cohort parameters to estimate were $N_{a,Y}$ for $a = 3, ..., 8$.

A summer research survey index of stock abundance was used to estimate survivors. Survey indices at ages 3 to 8 and for years 1970-1994 were used for estimation. These are also available in Fanning *et al.* (1995). The average survey catch ($R_{a,y}$) was assumed to be proportional to stock abundance; that is, $R_{a,y} \approx q_a N_{a,y}$, where q_a is the unknown age-dependent catchability coefficient of the research gear. The catchabilities for ages 3 to 8 were also estimated. The fit function used for estimation was

$$l(\theta) = \sum_{a,y} \left[\log(R_{a,y}) - \log(q_a) - \log\left\{N_{a,y}(t)\right\}\right]^2, \tag{6}$$

where $\log\left\{N_{a,y}(t)\right\} = \left\{\log(N_{a,y}) + \log(N_{a+1,y+1})\right\}/2$ was a cohort approximation of mid-year or summer abundance, which coincides with the timing of the survey.

4. Results

4.1. SPA for ESS cod

We present some summaries of the SPA estimates of $N_{a,y}$'s based on the entire time series of catch and survey data in Figure 2. Estimates of recruitment ($N_{3,\,y}$) and total abundance ($\sum_a N_{a,y}$) for the entire time series are plotted in the top panel. Estimates of total biomass ($\sum_a w_{a,y} N_{a,y}$) and SSB are plotted in the bottom panel. Note that $w_{a,y}$ is the beginning-of-year weight-at-age, and SSB is the sum of biomass for ages 5 to 15. The estimates of SSB in Figure 2 are identical to the estimates of SSB in Figure 1 based on the entire time series of catches and survey indices. Clearly in 1994 the size of this stock was estimated to be at very low levels.

The estimated survey catchabilities ($\times 10^3$) were 0.20 and 0.22 for ages 3 and 4, and then decreased almost linearly to 0.13 for age 8. We computed standardized residuals based on (6) using the estimated q's and N's. Time series plots of these residuals are shown in Figure 3. There are year effects in the surveys, particularly in 1974-1975, 1982-1985, and 1993. No age dependent patterns in the residuals were apparent.

Retrospective patterns in annual total abundance (N_+) and average F for ages 7 to 9 ($\bar{F}$) are shown in Figures 4 and 5. The retrospective trends in the estimates of N_+ shown in Figure 4 are similar to those for SSB shown in Figure 1. The trends are reversed for $\bar{F}$ (see Figure 5), because over-estimation of population numbers causes under-estimation of fishing mortalities. Retrospective patterns in the model predicted surveys indices, summed for ages 3 to 8, are shown in Figure 6. It appears that, at least partially, the retrospective pattern is caused by the high surveys in 1982-1984. The SPA attempts to fit to these surveys, but the lower survey indices observed after 1982-1984 results in a decreasing trend in predictions each year. Note that there is no "convergence" in the predicted survey indices in the early 1970's

because the estimated catchabilities change for each retrospective year. The estimates of stock size presented in Figures 1, 4, and 5 do converge.

4.2. Catch perturbations

In this section we present the direction of maximum slope, $d_{\max}$, based on the catch perturbations described by equation (2). We do this using ρ's for the retrospective patterns in N_+, $\bar{F}$, and SSB. The elements of the direction vectors are shown in Figure 7. They are similar for N_+ (panel a) and SSB (panel c). Note that $\dot{\rho}_{\max}$'s (i.e. the slopes) for N_+ and SSB are positive, which indicates that perturbations in these directions will increase ρ; hence, these results suggest that reducing reported catches from the 1962-1975 cohorts and increasing reported catches from the 1977-1987 cohorts will result in relatively large reductions in retrospective patterns. The elements of $d_{\max}$ for $\bar{F}$ are opposite in sign to those for N_+ and SSB, although similar in magnitude. This is because the retrospective trends in $\bar{F}$ ($\rho = -10.9$) are the reverse of the trends in N_+ ($\rho = 3.25$) and SSB ($\rho = 3.59$). Changes to the catches in the direction of $d_{\max}$ in panel b of Figure 7 will reduce the retrospective pattern in $\bar{F}$.

Changes in the catches from the 1988-1991 cohorts have relatively little effect on retrospective patterns. Note also that changes in the catches for the 1955-1961 cohorts have no effect on ρ. These catches only affect the SPA outside of the survey "tuning" block (ages 3 to 8), and changes to them can only affect SPA population size estimates prior to 1977.

Changes in the catches for the 1955-1961 cohorts have no effect on ρ. This is because these catches only affect the SPA outside of the survey "tuning" block (ages 3 to 8), and changes to these catches can only affect SPA population size estimates prior to 1977. This does not affect ρ because, like Mohn (1999), we measure retrospective patterns from 1985 to 1994.

The practical utility of $d_{\max}$ for finding perturbations that remove retrospective patterns depends on the linearity of the influence surface. We investigated the effectiveness of $d_{\max}$ in this regard, along with other perturbation schemes presented in the next three sections. We found that the influence surfaces were indeed almost linear within a reasonable neighborhood of the origin, ω_o. For example, in Figure 8 we plot the percent change in the SSB ρ based on perturbations to some individual catches and perturbations to all catches using $d_{\max}$. We used equation (2) for the perturbations, with $\omega = 1 + h$ for individual catch perturbations and $\omega_{a,y} = 1 + h d_{a,y}$ for perturbations in the direction $d_{\max}$, where $d_{a,y}$ was the appropriate element of $d_{\max}$. The value of h controlled the amount of perturbation; for example, $h = -0.5$ with an individual catch perturbation meant that the catch was reduced by 50%, whereas $h = 1$ meant that the catch was doubled. We refer to such perturbations as global because they involve "non-local" changes to SPA inputs. All of the influence graphs in Figure 8 are almost linear.

The local influence diagnostics in panel c of Figure 7 provide a very good description of the global influence graphs in Figure 8. For example, the results in Figure 7 suggest that increasing

$C_{10,1994}$ will decrease ρ, but at a much smaller rate than increasing $C_{8,1990}$. This is exactly what is shown in Figure 8. We produced 12 graphs analogous to Figure 8, one for each combination of the three ρ's with the four perturbation schemes. The influence graphs were almost always linear, and well described by local influence diagnostics.

If we reduce catches in the direction of $d_{\max}$ in Figure 7 (panel c) by the amount $h = -3.8$ then the value of $\dot{\rho}_{\max}$ suggests that ρ will change by approximately -100%; that is, ρ should be near zero for this perturbation. Note that $\dot{\rho}_{\max}^{-1} = 3.8$, and we take the negative of this value for h because we wish to reduce ρ. However, we found that $h = -3.5$ produced better results. This global perturbation greatly reduced the retrospective pattern in ESS cod. Differences between the total observed and perturbed catches are shown in Figure 9. Perturbed catches are smaller during 1971-1975 and larger during 1984-1992. The age compositions of the perturbed and reported catches are also different, and these differences are reflected by Figure 7 (panel c). The perturbed retrospective patterns are shown in Figure 10. Note that the scales in panels a and c in Figure 10 are identical to those in Figures 1 and 4, while the scale in panel b is slightly larger than that in Figure 5. Clearly the catch perturbations substantially reduced the retrospective pattern. The ρ statistics dropped from 3.25 to 0.12 for N_+, from -10.9 to -0.16 for $\bar{F}$, and from 3.59 to -0.49 for SSB. An improvement in the overall fit of the catch perturbed SPA was also realized. The mean square error (MSE) decreased from 0.36 to 0.32.

4.3. M perturbations

The elements of the direction vectors are shown in Figure 11. Generally speaking, they are similar to the catch perturbation results in Figure 7. The results in Figure 11 suggest that a reduction in M for the 1962-1975 cohorts and an increase M for the 1979-1987 cohorts will result in a significant reduction in the retrospective pattern. The value of $\dot{\rho}_{\max}$ for SSB suggests that a perturbation in the direction shown in panel c of Figure 11 and $h = -1.8$ will reduce ρ to near zero. After a few trials we chose $h = -1.4$. Average annual perturbed M's are shown in Figure 12. The perturbed M's tend to be less than 0.2 prior to 1984, and greater than 0.2 after 1985. Age-specific perturbations in M are reflected by Figures 11 and 12. The perturbed retrospective patterns are shown in Figure 13. Note that they are substantially reduced when compared to those in Figures 1, 4, and 5.

The MSE for the perturbed SPA based on the SSB $d_{\max}$ dropped slightly from 0.36 to 0.35. This is a smaller reduction compared to the catch perturbations in Section 4.2. The M and catch perturbed SPA's give substantially different estimates of population size; for example, the maximum total abundance from the catch-perturbed SPA is 2.79×10^8, whereas the maximum total abundance form the M-perturbed SPA is 3.94×10^8.

The M and catch perturbations were based on the SSB $d_{\max}$. Note that we also could have used $d_{\max}$ for N_+ or $\bar{F}$. For example, the results in Figure 11 suggest that a perturbation in

the direction shown in panel a with $h = -1.7$, will remove the retrospective pattern. This direction vector is similar to the one shown in panel c; however, we found that perturbations based on the N_+ $d_{\max}$ were less satisfactory. Values of h that removed the retrospective pattern in N_+ still left a retrospective pattern in SSB, with ρ close to one. While this value of ρ is substantially smaller than the unperturbed value $\rho = 3.59$, it is not as small as the value in Figure 13.

4.4. Catchability perturbations

Another interesting perturbation scheme involves the survey catchabilities. A basic assumption used in the ESS cod SPA is that the survey mean numbers-at-age per tow, or the survey indices, are proportional to absolute stock numbers-at-age. The constant of proportionality, q, was assumed to depend on age but not year. This assumption may not be true; for example, it is possible that the catchability of the survey changes over time. Mohn (1999) showed that violations of the constant catchability assumption could cause retrospective patterns. To assess the potential for this we examined influence diagnostics for multiplicative q perturbations,

$$q_{\omega\, a,y} = q_a \omega_{a,y},$$

where $\omega_o = 1$. This involved perturbations to unknown model parameters. We estimated the q_a's, and hence the $q_{\omega\, a,y}$'s; however, the $\omega_{a,y}$'s were fixed.

The elements of $d_{\max}$ for the three retrospective measures (ρ's) are shown in Figure 14. They suggest that an increase in the survey catchability around 1981-1983 can reduce the retrospective pattern. The reduction in ρ is even larger if the survey catchability is allowed to decrease in 1992-1994, particularly for ages 6 to 8. We applied the SSB $d_{\max}$ perturbation with $h = -3$. The estimated $q_{\omega\, a,y}$'s are shown in Figure 15. The estimates are averaged for two age groups, and within each age group the annual trends are very similar (see Figure 14). The estimates of q_a were almost identical to the unperturbed estimates. The perturbed retrospective patterns are presented in Figure 16. Note that they are much smaller than those in Figures 1, 4, and 5.

4.5. Case weight perturbations

Many methods for assessing influence involve the perturbation of case weights. A case refers to a term in the sum in (6), which we perturb as

$$l_\omega(\theta) = \sum_{a,y} \omega_{a,y} \left[\log(R_{a,y}) - \log(q_a) - \log\left\{N_{a,y}(t)\right\}\right]^2.$$

This perturbation scheme can be used to assess, for example, the impact of deleting survey indices for a particular year. In this section we assess whether changes in case weights can reduce the retrospective pattern.

The results in Figure 17 suggest that larger changes to case weights are required to reduce the retrospective pattern compared to the other perturbation schemes in Sections 4.2 to 4.4. This is because the values of $\dot{\rho}_{\max}$ are relatively small. The results in Figure 17 do not suggest that the assumptions on the error terms in the model used for estimation are a likely source of the retrospective pattern in ESS cod. The case weight perturbations that remove the retrospective patterns do not seem informative, and are therefore not presented. Nonetheless, it is apparent that ρ is affected more by the residuals at age 3 and 8 than at other ages. Also, ρ is affected more by residuals prior to 1976; that is, changing case weights in these years tend to have a greater affect on ρ than changing the weights after 1976.

In addition, the case weight influence analysis provided further information. Recall that in Section 4.1 we concluded that the high surveys in 1982-1984 may have caused some of the retrospective pattern; however, the results in Figure 17 suggest otherwise. Changing the case weights in 1982 to 1984 have a relatively small effect on ρ. For example, we estimated the SPA with case weights equal to zero for the 1982 and 1983 survey indices. The resulting retrospective patterns were nearly identical to the original patterns. This is not intuitively obvious from Figure 6.

5. Discussion

We have presented a practical methodology based on local influence diagnostics to assess the potential magnitude of changes in SPA inputs required to remove retrospective patterns. We showed how to use these methods to find relatively small perturbations to SPA component inputs that result in greatly reduced retrospective patterns. The rationale for doing this is if the smallest perturbation that removes the retrospective pattern is unrealistic then we can reasonably conclude that the pattern is not caused by the component.

We applied the methods to the ESS cod stock, and studied the influence of commercial catches, natural mortality, survey catchability, and estimation case weights on retrospective patterns. We concluded that it seemed unlikely that reasonable changes in estimation case weights could reduce the retrospective pattern substantially; that is, the retrospective pattern in ESS cod does not appear to be caused by model error assumptions. We found potentially reasonable perturbations to the other SPA inputs that greatly reduced the retrospective pattern; however, the plausibility of the perturbations, or the SPA perturbed stock estimates, is best assessed by ESS cod experts who are knowledgeable about the fishery and other scientific information for this stock.

Our catchability perturbation results agreed with the results in Mohn (1999). He used a much simpler perturbation scheme and concluded that a change in survey catchability in 1982 could explain the retrospective pattern in ESS cod. Mohn (1999) considered simple perturbations to catch and other model components and concluded that it was difficult to

discriminate between the various possible causes of retrospective patterns; however, based on other information, Mohn (1999) concluded that survey catchability assumptions appeared to be the principal source of the retrospective pattern. We also could not discriminate between catch, natural mortality, or survey catchability as the possible source of the retrospective pattern.

If the perturbations we found to remove retrospective patterns are equally plausible then we can demonstrate some retrospective "corrected" stock scenarios. In Figure 18 we show estimates of SSB based on the perturbations in Figure 9, 12, and 15. It is interesting that catch and M perturbations lead to somewhat different estimated stock trajectories, since they both play a very similar "role" in SPA by accounting for population deaths. The differences in the stock trajectories are probably the consequence of differences in the catch and M perturbation schemes. It is also interesting that the catchability perturbations resulted in a stock trajectory that is very similar to the unperturbed result. Nevertheless, the important message in Figure 18 is that in 1994 the stock was at a very low level, and this conclusion is the same for the four scenarios.

We found that the retrospective diagnostics were fairly similar when ρ was based on different stock quantities (e.g. $N+$, $\bar{F}$, or SSB), although we observed some differences in the effect of the perturbations on retrospective patterns. It would be useful to assess how sensitive the diagnostics are to the retrospective metric used; for example, an investigation of the effect of the choice of metric on $d_{\max}$ would be of interest.

It is possible that retrospective patterns are caused by mis-specifications of two or more model components. For example, the retrospective patterns could be caused by mis-reported catches and incorrect assumptions about survey catchability. The potential influence of this on retrospective patterns could be investigated using local influence diagnostics based on perturbations to multiple components. This is relatively straightforward to implement; however, scaling the perturbations is a problem. For example, multiplicative perturbations to catches may not be comparable to multiplicative perturbations to survey catchabilities. Local influence diagnostics based on multiple component perturbations will likely be very sensitive to the relative scaling of the perturbations to different components. Further discussion is presented in C&F.

References

Cadigan, N. G. and Farrell, P. J. 2002. Generalized local influence with applications to fish stock cohort analysis. Applied Statistics, 51: 469-483.

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

Evans, G. T. 1996. Using the elementary operations of sequential population analysis to display problems in catch or survey data. Canadian Journal of Fisheries and Aquatic Science, 53: 239-243.

Fanning, L. P., Mohn, R. K., and MacEachern, W. J. 1995. An assessment of 4VsW cod in 1994 with consideration of ecological indicators of stock status. DFO Atlantic Fisheries Research Document 95/73. 29 pp.

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.

Pope, J. G. 1972. An investigation of the accuracy of virtual population analysis using cohort analysis. ICNAF Research Bulletin, 9: 65-74.

Sinclair, A., Gascon, D., O'Boyle, R., Rivard, D., and Gavaris, S. 1991. Consistency of some Northwest Atlantic ground fish stock assessments. NAFO Scientific Council Studies, 16: 59-77.

Vaughan, D. S., and Prager, M. H. 2002. Severe decline in the abundance of the red porgy (*Pagrus pagrus*) population off the southeastern United States. Fisheries Bulletin, 100: 351-375.

Figures

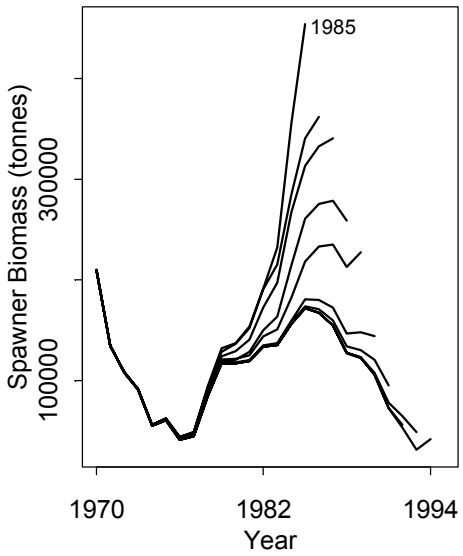


Figure 1: Retrospective estimates of spawner biomass (SSB).

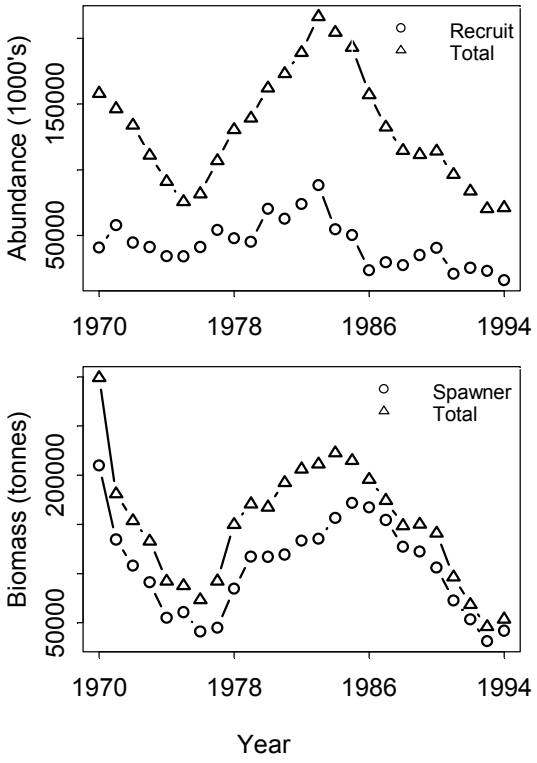


Figure 2: Estimates of recruitment and total abundance (top panel), and total and spawner biomass (bottom panel).

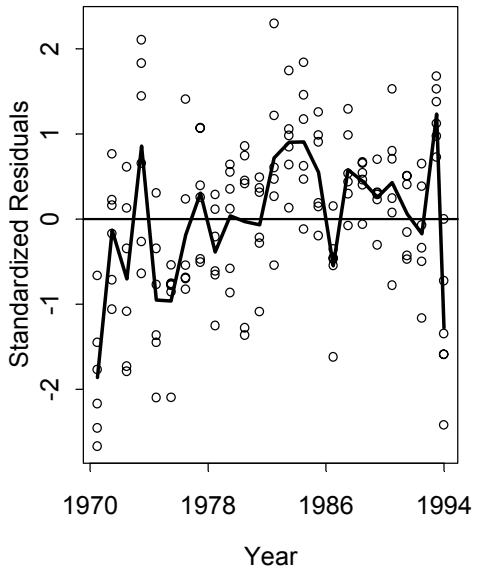


Figure 3: Time series of standardized survey residuals. The heavy solid line connects the average of the residuals each year.

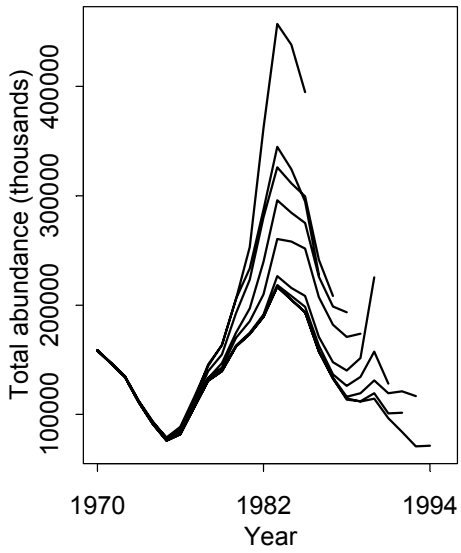


Figure 4: Retrospective estimates of total abundance (N_+) for ages 3-15.

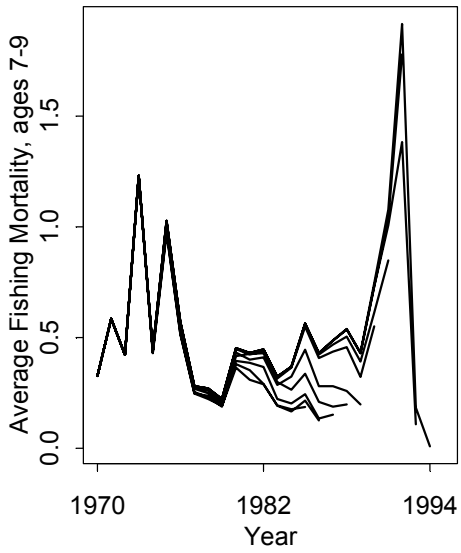


Figure 5: Retrospective estimates of average fishing mortality at ages 7 to 9 ($\bar{F}$).

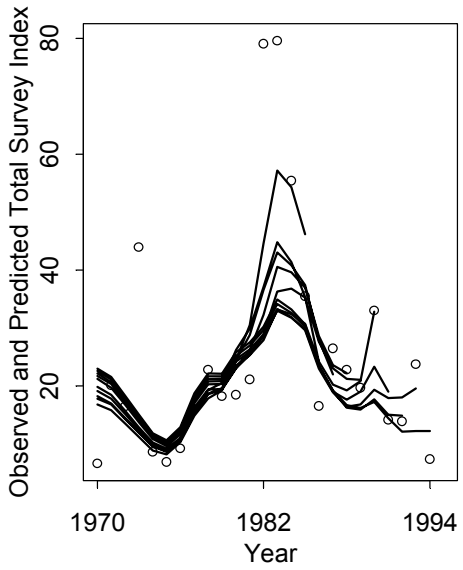


Figure 6: Observed and retrospective predicted total (ages 3-8) annual survey abundance.

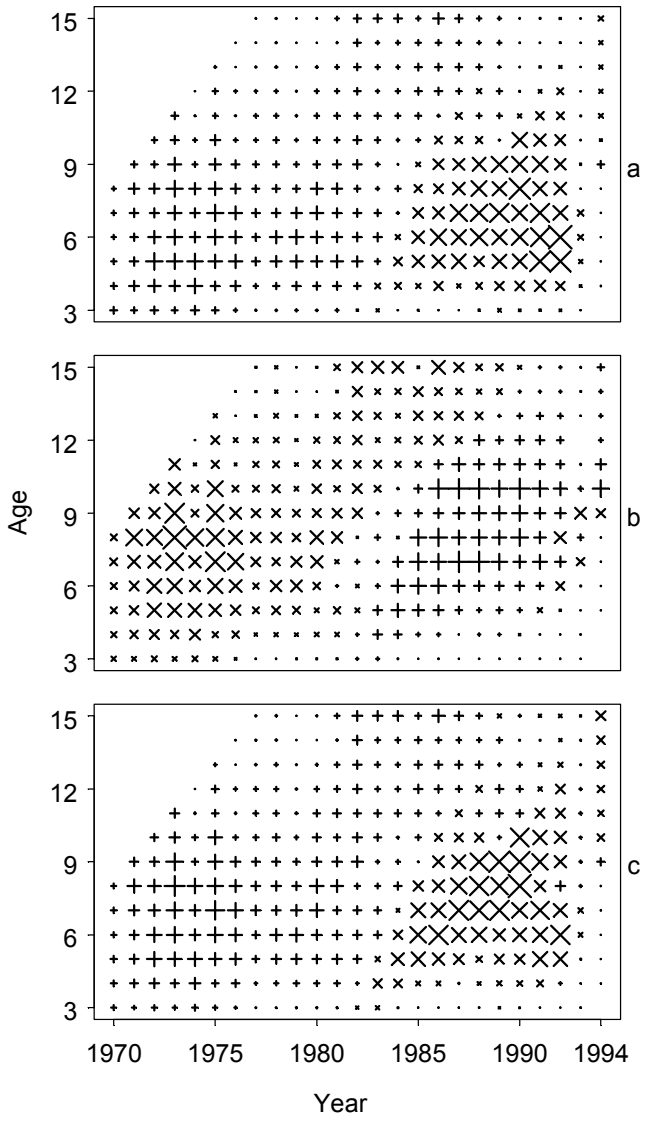


Figure 7: Catch local influence results for ρ. Each panel shows the elements of $d_{\max}$. The size and type of the plotting symbols are proportional to the absolute value and sign of the elements, respectively. Negative is denoted by an $\times$. Panel a: N_+, $\dot{\rho}_{\max} = 24.6\%$ of $\rho = 3.25$. Panel b: $\bar{F}$, $\dot{\rho}_{\max} = -51.0\%$ of $\rho = -10.9$. Panel c: SSB, $\dot{\rho}_{\max} = 26.3\%$ of $\rho = 3.59$.

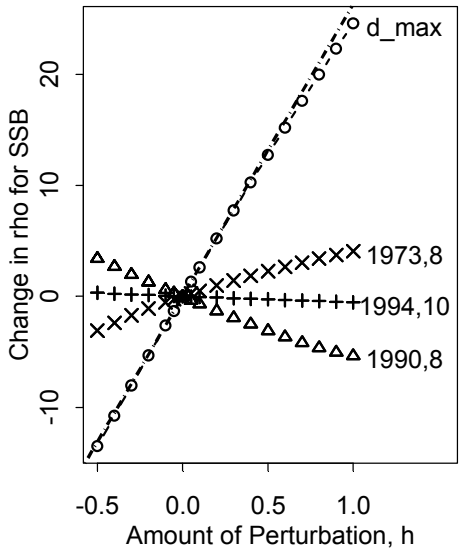


Figure 8: Displacement in the SSB ρ based on some catch global perturbations. All results are in percent of unperturbed estimates. The (year,age) indicate perturbations of individual catches. The dashed line is a straight line with a slope corresponding to $\dot{\rho}_{\max}$.

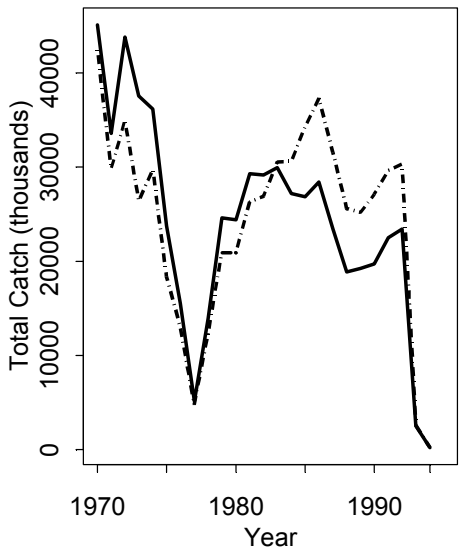


Figure 9: Total observed annual catch (solid line), and total perturbed annual catch (dashed line).

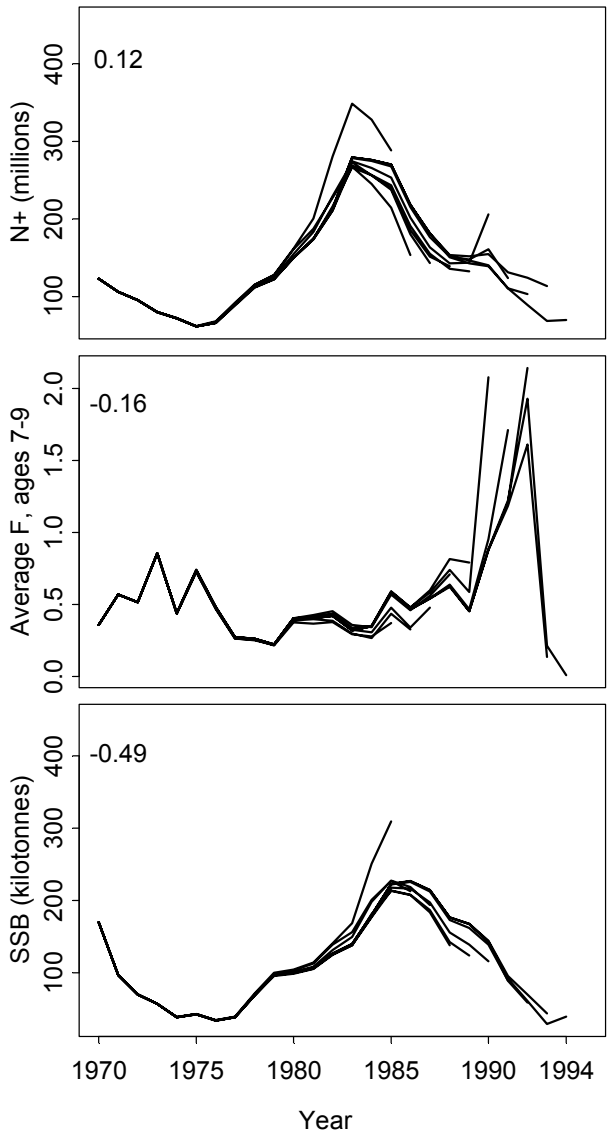


Figure 10: Retrospective estimates based on $d_{\max}$ perturbed catches. Top panel: total abundance, N_+; middle panel: Average fishing mortality, $\bar{F}$; bottom panel: Spawning stock biomass, SSB. The value of ρ is shown in the top left-hand corner.

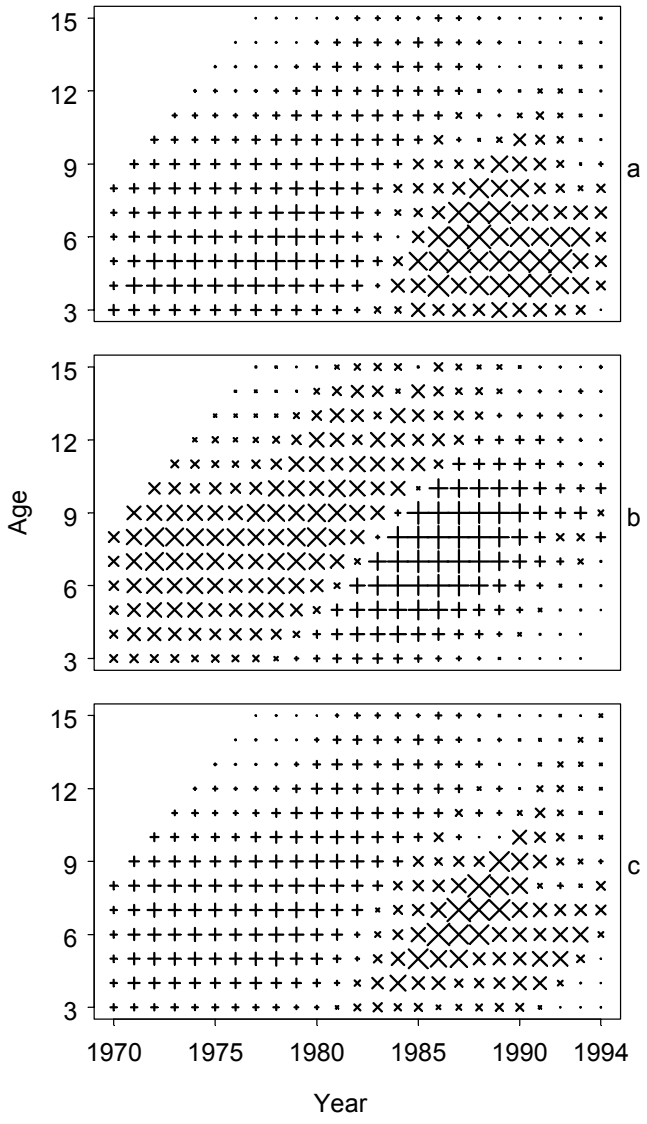


Figure 11: M local influence results for ρ. Each panel shows the elements of $d_{\max}$. The size and type of the plotting symbols are proportional to the absolute value and sign of the elements, respectively. Negative is denoted by an $\times$. Panel a: N_+, $\dot{\rho}_{\max} = 58.8\%$ of $\rho = 3.25$. Panel b: $\bar{F}$, $\dot{\rho}_{\max} = -108.0\%$ of $\rho = -10.9$. Panel c: SSB, $\dot{\rho}_{\max} = 54.7\%$ of $\rho = 3.59$.

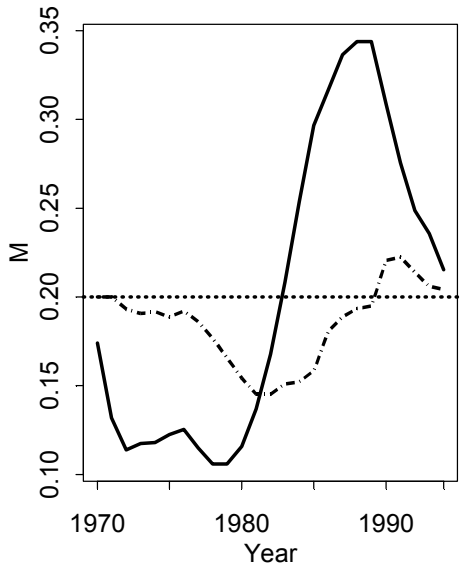


Figure 12: Average perturbed natural mortality, M_ω, at ages 3-9 (solid line) and ages 10-15 (dashed line). The nominal value $M = 0.2$ is shown as a dotted line.

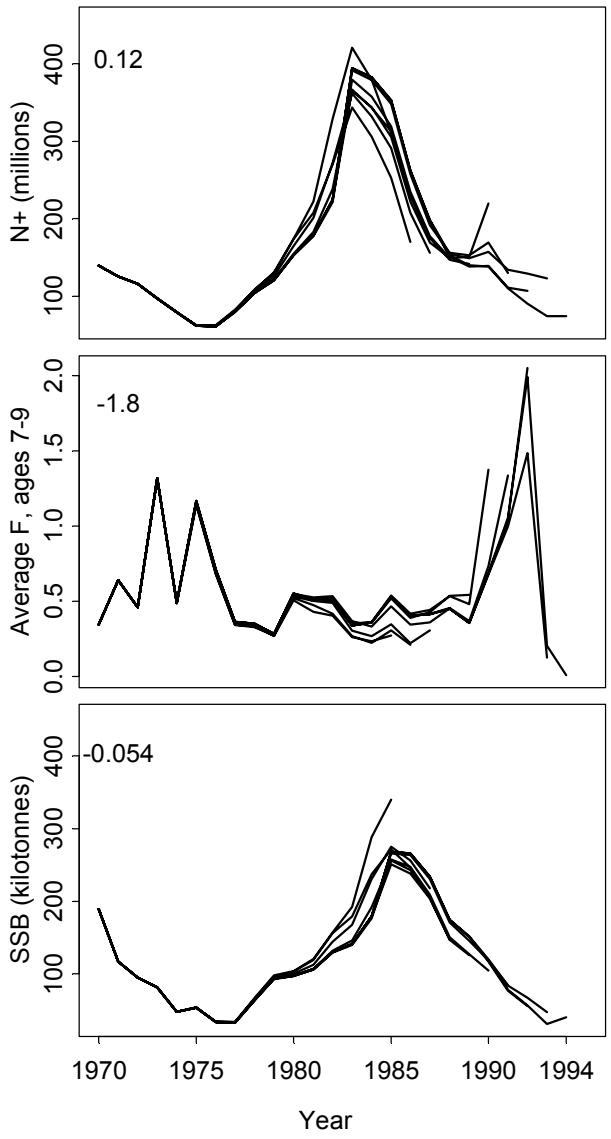


Figure 13: Retrospective estimates based on $d_{\max}$ perturbed natural mortalities, M_ω. Top panel: total abundance, N_+; middle panel: Average fishing mortality, $\bar{F}$; bottom panel: Spawning stock biomass, SSB. The value of ρ is shown in the top left-hand corner.

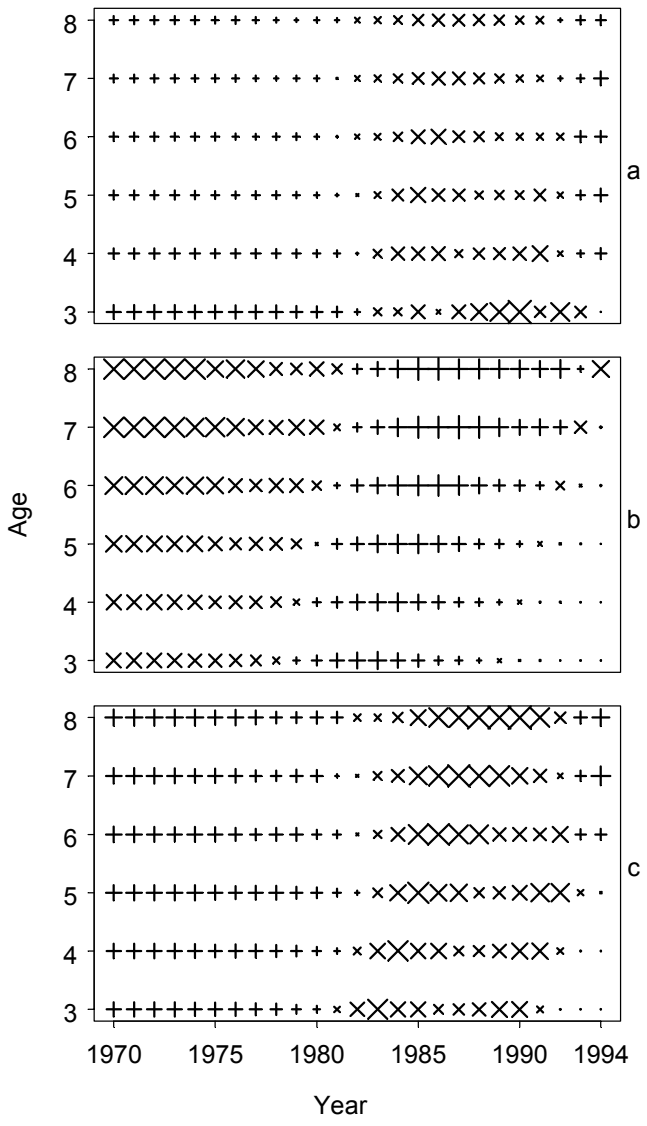


Figure 14: Survey catchability influence results for ρ. Each panel shows the elements of $d_{\max}$. The size and type of the plotting symbols are proportional to the absolute value and sign of the elements, respectively. Negative is denoted by an $\times$. Panel a: N_+, $\dot{\rho}_{\max} = 35.4\%$ of $\rho = 3.25$. Panel b: $\bar{F}$, $\dot{\rho}_{\max} = -57.7\%$ of $\rho = -10.9$. Panel c: SSB, $\dot{\rho}_{\max} = 31.3\%$ of $\rho = 3.59$.

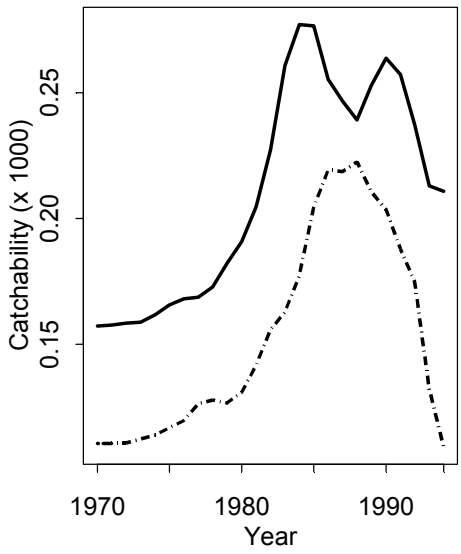


Figure 15: Average perturbed survey catchabilities at ages 3-5 (solid line) and ages 6-9 (dashed line).

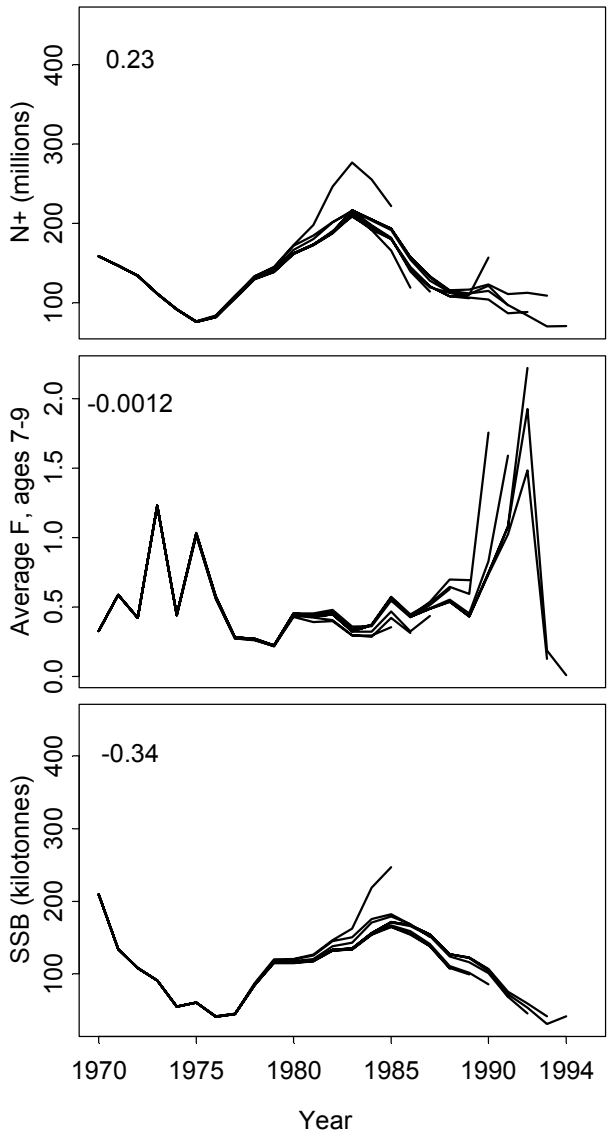


Figure 16: Retrospective estimates based on $d_{\max}$ perturbed survey catchabilities, q_ω. Top panel: total abundance, N_+; middle panel: Average fishing mortality, $\bar{F}$; bottom panel: Spawning stock biomass, SSB. The value of ρ is shown in the top left-hand corner.

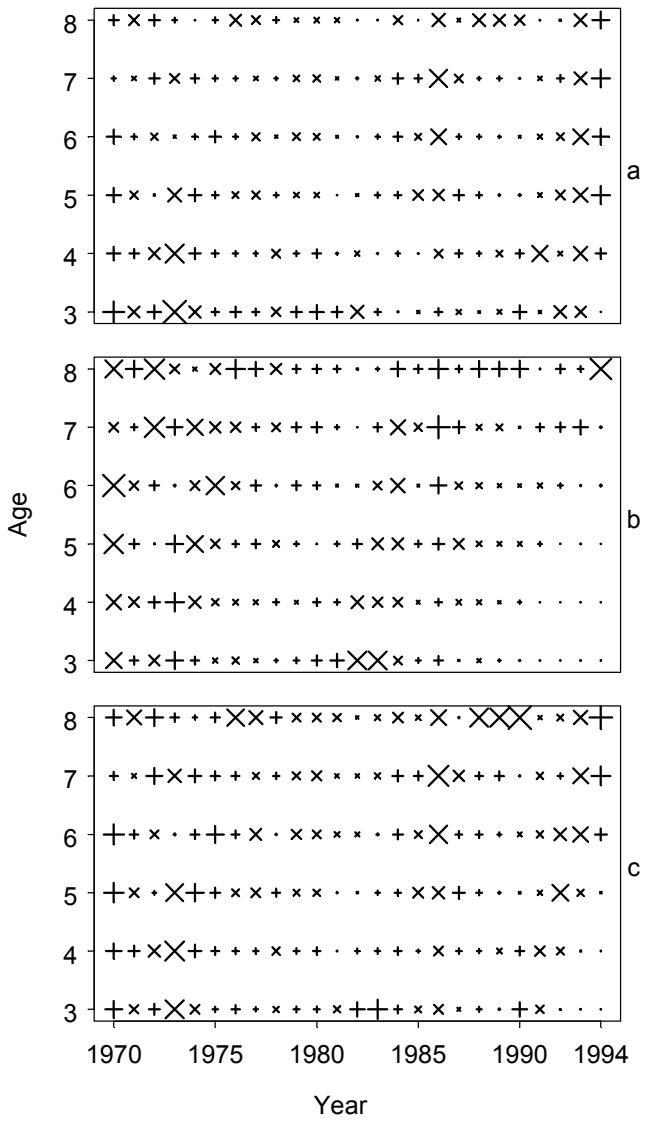


Figure 17: Case weight local influence results for ρ. Each panel shows the elements of $d_{\max}$. The size and type of the plotting symbols are proportional to the absolute value and sign of the elements, respectively. Negative is denoted by an $\times$. Panel a: N_+, $\dot{\rho}_{\max} = 18.3\%$ of $\rho = 3.25$. Panel b: $\bar{F}$, $\dot{\rho}_{\max} = -33.7\%$ of $\rho = -10.9$. Panel c: SSB, $\dot{\rho}_{\max} = 18.5\%$ of $\rho = 3.59$.

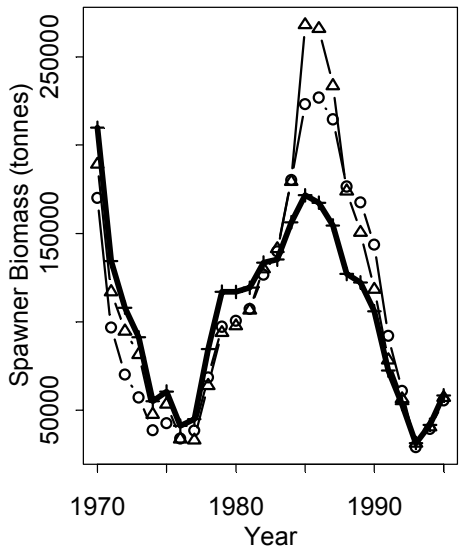


Figure 18: Comparison of spawning stock biomass (SSB) estimates: unperturbed (heavy solid line), M-perturbed ($\triangle$), catch perturbed ($\circ$), and q-perturbed ($+$).