Aquat. Living Resour. 22, 207–216 (2009)

DOI: 10.1051/alr/2009003
www.alr-journal.org

Aquatic Living Resources

A review of fishery-independent assessment models, and initial evaluation based on simulated data

Benoit Mesnil[1,a], John Cotter[2], Rob J. Fryer[3], Coby L. Needle[3] and Verena M. Trenkel[1]

[1] Ifremer, Département EMH, BP 21105, 44311 Nantes Cedex 3, France
[2] CEFAS Lowestoft Laboratory, Pakefield Road, Lowestoft, Suffolk NR33 0HT, UK
[3] FRS Marine Laboratory, PO Box 101, Victoria Road, Aberdeen AB11 9DB, UK

Received 15 May 2008; Accepted 8 September 2008

Abstract – Large uncertainties in catch data (officially-reported landings and discards) are undermining the ability of scientific organisations to provide valid management advice based on the conventional approach of analytical stock assessments. There is thus an urgent need to consider alternative tools that do not depend on long series of precise age-structured catch data. This paper presents four fishery-independent assessment models developed under the EU project FISBOAT (Fishery Independent Survey Based Operational Assessment Tools). It also reports on rudimentary tests based on simulated data, using the same data sets and protocol as an evaluation study conducted by the US National Research Council in 1997. The survey-based assessment models at hand are able to reliably capture the major signal in biomass and recruitment, although they smooth out transient changes. However, they cannot provide absolute abundance estimates, only relative values on an arbitrary scale. The survey-based approaches could provide more rapid updates of the state of stocks than catch-based methods.

Key words: Fish stock assessment / Catch-free models / Fishery-independent assessment methods (F-I methods) / Survey indices

1 Introduction

All stock assessment methods (whether they involve surplus-production, delay-difference, stock-reduction, Collie-Sissenwine or analytical dynamic pool models) currently used by scientific organisations to advise fisheries managers on the state of fish stocks require knowledge of total catches (commercial and recreational, where appropriate) to estimate model parameters and other quantities of management interest. Errors in the input catch figures translate directly into similar errors in stock abundance estimates (e.g., Quinn and Deriso 1999), and if their magnitude varies from year to year the assessments may not even reflect the relative changes in the state of the resource. When catch is also the basis of management control, like in TAC (Total Allowable Catch) systems, there is often a temptation for fishers or states to mis-report landings for tactical reasons, especially when catch quotas become very restrictive. Large unrecorded discards at sea by fishing vessels are another problem. For over a decade the International Council for the Exploration of the Sea (ICES) has repeatedly stated that the deterioration of the catch data was threatening its ability to provide managers with the type of advice they require to apply the current policies.

One objective of the FISBOAT (Fishery Independent Survey Based Operational Assessment Tools) project was to develop and evaluate operational fishery-independent (i.e. catch-free) assessment tools that were capable of alleviating these data problems. During the project six fishery-independent (F-I) stock assessment methods were specifically developed, or elaborated upon for use without fishery catch data. Methods are here understood to mean both the mathematical models and the procedures to estimate their parameters. Four of these models can handle age-structured data (survey indices[1]) that are compatible with the selected test data, and only this subset that went through a common testing procedure is discussed in this paper. Section 2 provides a concise overview of each model, with comments on parameter estimation and practical guidelines or caveats regarding its use in assessment and advisory groups.

Although it was recognised that an authoritative evaluation of the F-I methods should involve a simulation-testing evaluation framework (with operating model, harvest rule, etc.; see Hillary 2009), the four methods were largely novel and there was a need to understand their capabilities through a simpler benchmarking approach before proceeding further. This simpler testing exercise followed the same protocol, using artificial data with known properties, as an evaluation study conducted in the USA on catch-based (mostly age-structured) assessment models (NRC 1998). Sections 3 and 4 recount the

[a] Corresponding author: Benoit.Mesnil@ifremer.fr

[1] Throughout this article we use the term "index" as a synonym for survey CPUE.

conditions and results of these preliminary probing tests carried out on the four models. Section 5 concludes on the insight gained during the project into the potential performance of the F-I methods for assessment of stock status, and on some implications for the European (ICES) advisory system.

2 Methods considered

The F-I methods presented in this section are intended to estimate fish stock abundance, or trends thereof; one is stage-structured (BREM), and three are age-structured (SURBA, TSA and YCC).

2.1 Biomass random effects model (BREM)

Model description

The population dynamics are formulated as the difference model from Hilborn and Walters (1992, p. 336):

$$B_t = R_t + g_{t-1}B_{t-1} \quad (1)$$

where B_t is the total population biomass and R_t the recruitment in biomass in year t. g_{t-1} is the net biomass growth rate: that is, the balance between individual growth and total (natural + fishing) mortality. Recruitment is assumed to follow a lognormal distribution without any stock-recruitment relationship:

$$\log R_t \sim N(\mu_R, \sigma_R^2).$$

Biomass growth is modelled by a random walk on the log-scale, to reflect the assumption that total mortality, which is part of g, does not vary wildly from year to year:

$$\log g_t = \log g_{t-1} + \varepsilon_t^g \;\; ; \;\; \varepsilon_t^g \sim N(-0.5\sigma_g^2, \sigma_g^2).$$

Thus, both recruitment R_t and biomass growth g_{t-1} are treated as random effects with parameters μ_R and σ_R^2, and g_1 (growth rate at $t = 1$) and σ_g^2 respectively. The negative mean $-0.5\sigma^2$ is a bias-correction term for a lognormal distribution.

The observation model uses two indices, b_t for total biomass at time t (recruits included) and r_t for recruits only. Both are assumed to follow lognormal distributions with static variance and catchability coefficient:

$$\log b_t \sim N(\log(q_b B_t), \sigma_b^2)$$
$$\log r_t \sim N(\log(q_r R_t), \sigma_r^2).$$

To ensure identifiability, the following constraints are imposed: $q_b = 1$ and $\sigma_b^2 = \sigma_r^2$. Other constraints are possible, e.g. larger CV for recruits' survey.

Sensitivity and robustness issues

Convergence of the parameter estimation algorithm depends critically on sensible starting values. The above mentioned constraints allow parameter identifiability, but the effect of setting $q_b = 1$ is that biomass estimates can only be relative (not absolute). In addition, the estimates of mean recruitment μ_R and catchability for recruits q_r are confounded to some degree. This appears as a strong correlation between estimates. Furthermore, the standard deviation of biomass growth σ_g is only estimable if relatively large. Results of extensive simulation studies exploring parameter confounding can be found in Trenkel (2008).

Input and output

BREM only requires two series of survey indices in biomass, one for the total population (adults + recruits) and one for the recruits alone; splitting out the recruits can be based on age readings but there are favourable cases where a reasonable cut-off size may be identified by inspection of the length compositions. Knowledge of natural mortality M is not required and occasional gaps in survey series are unlikely to affect the estimation. An extension handling two series of indices per category (e.g. acoustic and egg surveys) has been developed (Trenkel 2007, 2008).

Seven parameters are estimated: B_1 (biomass in year 1), g_1 (biomass growth in year 1), σ_g (standard deviation of growth), μ_R (mean recruitment for normal distribution), σ_R (standard deviation of recruitment), q_r (catchability of recruits) and σ_b (standard deviation of observation error for base normal). Plugging converged estimates into Eq. (1) yields estimated time trajectories of relative total biomass and annual recruitment. In addition, standard deviations are available for biomass estimates based on the observed Fisher Information matrix, but NOT the same way for recruitment estimates as these are random effects, not real parameters.

Implementation issues

Parameter estimation by maximum likelihood is implemented in AD Model Builder (Fournier 2005) using the random effects module (Skaug and Fournier 2006). Run time for NRC set 1 (Sect. 3) was about 20 seconds. Run time does not increase with the number of years, but depends on how good the starting values are.

Relation to management indicators

Future recruitment could be predicted using the fitted lognormal distribution, either as expected recruitment or by drawing a random recruitment value from the distribution. The relationship between model predictions and commercial quantities is not obvious.

2.2 SURBA

Model description

The basis of *SURBA* is a simple survey-based separable model of mortality. This model was first applied to European research-vessel survey data by Cook (1997, 2004), but

it has a long history in catch-based fisheries stock assessment (Pope and Shepherd 1982; Deriso et al. 1985; Gudmundsson 1986; Johnson and Quinn 1987; Patterson and Melvin 1996; see Quinn and Deriso 1999 for a summary). An early version of the current implementation was presented in Beare et al. (2005).

The separable model used in *SURBA* assumes that total mortality $Z_{a,y}$ for ages a and y can be expressed as:

$$Z_{a,y} = s_a \times f_y$$

where s_a and f_y are respectively the age and year effects of mortality. This differs from the usual assumption in that total mortality Z is the quantity of interest, rather than fishing mortality F. Then, given $Z_{a,y}$, abundance $N_{a,y}$ can be derived as:

$$N_{a,y} = R_{y_0} \exp\left(-Z_{a0,y0} - Z_{a0+1,y0+1} - Z_{a0+2,y0+2} - \ldots - Z_{a-1,y-1}\right)$$

where a_0 and $y_0 = y - a + a_0$ are respectively the age and year in which the fish measured as $N_{a,y}$ first recruit to the observed population. Thus the abundance at each age and year of a cohort is given by the recruiting abundance R_{y_0} of the relevant cohort modified by the cumulative effect of mortality during its lifetime. Given assumed catchabilities $q_{a,y,i}$ for each index (survey) i, estimated abundance indices can be derived as:

$$\hat{I}_{a,y,i} = q_{a,y,i}\hat{N}_{a,y}.$$

Parameters are estimated by minimising the weighted sum-of-squares of differences between observed and estimated log abundance indices,

$$SSQ = \sum_{a,y,i} \omega_{a,y,i} \left(\log I_{a,y,i} - \log \hat{I}_{a,y,i}\right)^2$$

where $\omega_{a,y,i}$ are optional weighting terms. Corresponding SSQ terms can also be included for biomass indices. All abundance estimates are relative.

This simple basis has been expanded considerably over recent years, as the model has been road-tested in ICES assessment working groups (and elsewhere) and modified where necessary. The development is summarised in Needle and Hillary (2007) and Beare et al. (2005), but in brief:

- Biomass indices can be used, as well as multiple age-structured indices.
- The year-effect for the final year is set to the mean of the previous three year effects, as the terminal year-effect cannot currently be determined directly from the data.
- Age-structured indices are all back-shifted to the start of the year, using the current estimate of Z. This allows them to be compared directly, and ensures firstly, that abundance indices refer to January 1, and secondly, that mortality estimates relate to the calendar year rather than the year between successive cruises of a given survey.
- Biomass indices are shifted forwards to spawning time before inclusion in the parameter estimation process.
- Index catchabilities and SSQ weightings can both be defined by the user.
- Optionally, a smoothing term can be added to the SSQ to penalise excessive inter-annual variation in estimated year effects. The degree of smoothing is determined by a user-defined variable λ.
- The reference age (that is, the age at which the age-effect s is fixed to 1.0) can also be defined by the user.
- Estimated variances (and thereby confidence intervals) of mean Z and recruitment are derived from the variance-covariance matrix of the parameter estimates, using the delta method. Variances estimates for abundance and spawning stock biomass SSB are currently being implemented.
- Retrospective runs can be generated automatically, with the last year of data being moved back one year at a time until half of the original time-series remains.
- A scan facility automatically runs assessments with a range of choices for smoothing, the reference age, and catchability on the first age, to evaluate model sensitivity to these essentially *ad hoc* settings.

Sensitivity and robustness

The model is most sensitive to assumptions about catchability. In particular, estimates of Z can be very different under different assumptions about catchability; SSB estimates are more robust. Z estimates can be very uncertain in any case, and it is not uncommon for there to be no significant evidence of any changes in the levels of Z. Finally, the automated scanning routine sometimes fails – values scanned over need to be interactively defined in future.

Inputs and output

SURBA uses the Lowestoft VPA input format (Darby and Flatman 1994), and currently expects to see the full set of such files – which means that dummy catch-based data files with arbitrary values had to be set up to analyse the NRC datasets. The inputs that are actually required for fitting the model are age-structured indices, and (optionally) biomass indices. The user can also define catchability and SSQ weightings for both types of index, along with values for the smoother λ and the reference age.

Both text and graphical outputs are provided by the program. Text outputs include parameter estimates with variances, mortality and relative abundance estimates, estimated variances for mean Z and recruitment, log residuals, stock summaries (SSB etc.), results of retrospective and scan runs, and goodness-of-fit statistics. Plots include exploratory raw-data figures (such as catch curves), model fits and stock summaries, residuals, and retrospective summaries.

Implementation issues

SURBA (currently Version 3.0) is implemented in Fortran-90 with NAG library minimisers and a Windows user interface, in which diagnostic plots are automatically generated. The run time for NRC set 1 on a 1.60 GHz, 512 MB RAM laptop was 6 s (standard), 40 s (standard + 15 retrospective runs), and 7 min 47 s (105-run scan).

Relation to management indicators

Abundance estimates (and therefore biomass measures) are currently generated by *SURBA* on a relative scale only, and are usually plotted as mean-standardised values for ease of comparison. Furthermore, *SURBA* provides estimates of total mortality Z rather than fishing mortality F (although, given the tentative nature of most natural mortality estimates, this is true of catch-at-age methods also). Therefore *SURBA* can be used to provide advice on relative trends in abundance and total mortality, but not absolute levels. It is possible to generate pseudo-absolute abundance estimates by using a catch-at-age VPA to estimate survey catchabilities-at-age using data from some period in the past, and then applying these to recent *SURBA*-derived relative population estimates to scale them to a level commensurate with that indicated by catch data. However, this requires assumptions that there was a period when catch data were reliable, and that the relationship between survey and fishery catchability has remained constant ever since, and these can be hard to maintain. It is also possible, of course, to produce F estimates by subtracting fixed M values from the Z estimates produced by *SURBA*.

2.3 Time series analysis (TSA)

Model description

TSA, or "Time Series Analysis", is a state space framework for modelling a fishery. The initial implementation modelled commercial catch-at-age data with survey indices-at-age used as auxiliary information (Gudmundsson 1994). The framework has since been extended to deal with a range of data sources and applications (e.g. Fryer 2002; Gudmundsson 2004). Here, the framework is adapted to model the indices-at-age from a single survey. The state equations relate the log numbers-at-age and total mortalities-at-age in year y to those in year y–1. Log numbers-at-age in year y are given by:

$$\log N_{a,y} = \log N_{a-1,y-1} - Z_{a-1,y-1}, \quad a > 1$$
$$\log N_{1,y} \sim N(\mu_R, \sigma_R^2).$$

Total mortalities are partitioned into fishing mortalities and natural mortalities through $Z_{a,y} = F_{a,y} + M_a$. Natural mortalities are assumed known and fishing mortalities evolve according to the following model:

$$\log F_{a,y} = U_{a,y} + V_y \tag{2}$$

$U_{a,y} = U_{a,y-1} + N(0, \sigma_U^2)$ with the constraint that $\sum_a U_{a,y} = 0$

$$V_y = Y_y + N(0, \sigma_V^2)$$

$$Y_y = Y_{y-1} + N(0, \sigma_Y^2).$$

Thus, log fishing mortality is separated into an age component $U_{a,y}$ and a year component V_y, both of which can evolve over time. Finally, the state vector consists of the log $N_{a,y}$, log $F_{a,y}$, $U_{a,y}$, V_y and Y_y.

The observation equations are given by:

$$\log I_{a,y} = \log q_a + \log N_{a,y} + \varepsilon_{a,y}$$

where $I_{a,y}$ are the indices-at-age, q_a are the survey catchabilities, and the $\varepsilon_{a,y}$ are assumed to be normally distributed with zero mean and standard deviation $\sigma_I\ \lambda_a\ \delta_{a,y}$. The λ_a are initially taken to be unity, but can be adjusted later if the errors associated with some ages are larger than for others. The $\delta_{a,y}$ are also initially taken to be unity, but can be inflated to decrease the influence of outliers. It is assumed that the survey takes place at the start of the year.

The model is fitted using the Kalman Filter, with the parameters μ_R, σ_R, σ_I, σ_U, σ_V, σ_Y, q_a, $U_{a,1}$ estimated by maximum likelihood. For identifiability, log q_1, V_1 and Y_1 are taken to be zero. For stability, some constraints must be put on the q_a: for testing on the NRC data sets, the q_a (for $a > 1$) were constrained to change log-linearly with age.

Sensitivity / robustness

Good starting values are crucial, but can be hard to find for a new stock and survey, requiring some iteration and experience. Once found, however, parameter estimates are robust to the addition of further years of data.

The method works on the log scale, so zero indices must be replaced by some small positive value. Unity was used for the NRC data sets. This means that the method can only be sensibly applied to those age classes where zero indices do not often occur – typically the younger age classes. An option would be to group older age classes into a single plus group, but this has not yet been implemented.

Very large year classes can cause a problem, because they can unduly dominate the parameter estimates associated with recruitment (i.e. μ_R and σ_R). It is possible to reduce their impact on these estimates, but this is done manually following graphical inspection of standardised prediction errors.

Inputs and outputs

TSA requires survey indices-at-age. Although described above for a single survey, the method can handle multiple surveys, which do not need to overlap. Missing survey indices (and missing years) are allowed. Natural mortalities-at-age are required to estimate (relative) fishing mortalities. If these are unavailable, F is replaced by Z in Eq. (2), and (relative) total mortalities are estimated.

TSA estimates relative numbers-at-age and relative mortalities-at-age with approximate coefficients of variation. The estimates can not be combined across age classes (as there are separate scaling factors for each age class), so it is not possible to estimate (relative) biomass. However, sensible proxies for stock biomass can be constructed.

Implementation issues

TSA is written in Fortran 90, using NAG routines, but can be called from R. However, stock-specific changes to the parameter constraints sometimes need to be made in the source Fortran code, rather than in R. The run time for NRC data set 1 (Sect. 3) was about 30 s on a 1.8 GHz, 524 MB RAM laptop.

Predictive ability

The method can predict both relative numbers-at-age and fishing mortalities-at-age (with approximate coefficients of variation) as far into the future as required.

2.4 Year-class curve (YCC) method

Model description

A "year-class" curve is a plot of log indices against age for a single year-class of a species. Marine fish caught in trawls typically show nearly linear year-class curves for ages that are fully selected. The usual model of mortality over time t, assuming no net migration to or from the stock, is considered:

$$N_t = N_0 \exp(-Zt) \quad (3)$$

where Z is the instantaneous rate of total mortality. We now assume that the indices I are a constant proportion of N, i.e. $I = qN$ for all ages included in the analysis, and that Z represents an average value over time. Then, taking natural logarithms of Eq. (3), restricting attention to one year-class, c, substituting *age* for t, and adding a random error term, e, gives the basic model for a year-class curve:

$$\log I_{a,c} = \log(I_{0,c}) - Z\,age + e_{a,c} \quad (4)$$

where $I_{0,c}$ is the index for age zero, a is the age-class, i.e. the age in years as an integer index, while *age* is age in years as a real number. e is assumed to be normally distributed with zero mean and variance σ_e^2. Additional linear terms may be added to Eq. (4) to allow for varying selectivity of the survey trawl with age, for survey-specific catchabilities, and/or for gradual changes in Z over time. The latter is achieved using polynomials in *age* and *year* with a minimum of additional parameters so as to yield best precision of estimation with the available data.

Different series of survey indices are likely to estimate year-class curves with different precision depending on the season and area covered, on the precision of age-reading and other practical aspects, and on how well the chosen model fits the data. Weighting of different data sets to reflect their precision with respect to the chosen model is therefore desirable. Cotter and Buckland (2004) suggest that the weighting estimated for each index series f should be balanced with the reciprocal of the estimated residual variance specific to that survey computed after the model is fitted, i.e. $\hat{w}_f \propto \hat{\sigma}_f^{-2}$. They describe how the method can be implemented using iteratively weighted least squares (IWLS) taking into account the degrees of freedom contributed by each survey to the estimates of each parameter. Usually, 2 or 3 iterations produce stable values. Additionally, using the survey specific residual variances, the relative precision of the different surveys can be compared using F tests (Cotter 2001). Note that biased survey indices will produce biased weights (Quinn and Deriso 1999, p. 353). Surveys that appear exceptionally precise should be scrutinised to see whether biased sampling may be the cause, e.g. due to clustering of observations in restricted times or places (Cotter and Buckland 2004).

A year-class curve can be fitted repeatedly in a process called *forward validation* that is designed to find the most reliable model for predicting next year's indices. Starting from an early year and proceeding forwards in the time-series, it finds the differences between the predicted log indices and the observed log indices for one year after the time domain of the data used to fit the model. The preferred model is the one whose mean difference is closest to zero, and for which the mean square of the differences is lowest. This is merely a simulation of a fish stock assessment working group making predictions each year for the coming year, then checking them when the outcome is known. Full details of available models, survey weighting, and forward validation to find the preferred model are given by Cotter et al. (2007).

Sensitivity / robustness

Catchability must be constant over time but may vary between surveys since intercalibration factors are automatically fitted if required. Changes to the design of a survey that might cause a change in catchability (e.g. a different vessel or gear) can be accommodated simply by treating it as a new survey and fitting an extra intercalibration factor.

Only gradual changes of Z are allowed by using polynomials in *year* to a maximum degree of 3. This is intended to minimise the dangers of erroneously treating random measurement errors as trends in the year-class signal over time. However, if sudden, real changes in Z actually do occur from year to year, they might be overlooked.

Year-class curves can be fitted across surveys, or nested within. Over- and under-fitting can both cause biased estimates of parameters. Forward validation helps to eliminate such models because they tend to be poor at predicting beyond the observed domain. The AIC may also be used to help find the best model.

Inputs and outputs

The basic input is a standard VPA-type tuning file (Darby and Flatman 1994). *YCC* software operates on a flat file having survey, age, year, time-of-year, indices etc., so such a file may be used directly if preferred. Year-class curves are available as plots over time, one per year class. These allow the fitted model to be compared to the observed values to check that the fit is credible. Relative recruitments, and Z over age by fleet are also given, along with various other outputs.

No assumptions are made about natural mortality (M). Fishing mortality (F) could be estimated if they were.

Implementation issues

Software written in R is available to fit year-class curves with all the options described here. Diagnostics include prediction and residual errors over time, age, and year class. Run times are usually seconds but may increase to a minute or more when there are many surveys, iterative re-weighting, and a long period of forward validation. The model may fail to fit if there

Table 1. Specifications of the simulated data sets (expanded from NRC 1998 for sets 1-5; set 6 without error in data added for this test). q stands for catchability, M for natural mortality, Y/B for yield/biomass ratio as a proxy for fishing pressure.

Set	Population trend	Age at 50% selectivity	Misreporting rate	Survey q	CV survey q	M	Mean Y/B
1	Depletion	Lower later	0.97–1.03	Constant	0.3	0.18–0.27	0.19
2	Depletion	Lower later	0.68–0.72	Constant	0.3	0.18–0.27	0.12
3	Depletion	Lower later	0.97–1.03	Higher later	0.3	0.18–0.27	0.12
4	Depletion	Constant	0.97–1.03	Constant	0.3	0.18–0.27	0.21
5	Recovery	Constant	0.97–1.03	Constant	0.3	0.18–0.27	0.07
6	2-way trip	Constant	0	Constant	0.0	0.2	0.15

are more parameters than observed vectors of indices-at-age. Missing values may either be omitted from the data set or coded as negative indices.

Predictive abilities

Predictions one year ahead of observed data are carried out routinely with forward validation. *YCC* produces tables of predicted indices-at-age for the year after the final observed year together with prediction mean square errors.

Relation to management indicators

Predicted indices-at-age in terms of numbers may be converted to weights per unit of effort-at-age using a matrix of weights-at-age by year. These may in turn be converted to spawning stock biomass per unit of effort-at-age using a matrix of maturity-at-age by year. The software allows users to insert independent observed values for each year, if available. However, *YCC* offers no prediction of next year's recruiting year class.

3 Testing procedure

3.1 Data sets

In the absence of a better alternative at the time, we resorted to the suite of data sets concocted for the US National Research Council rounds of tests during 1997. One advantage is that the outcome has been published (NRC 1998), enabling the performance of other methods to be compared with that of the methods considered by that committee (which all made use of catch and/or catch-at-age data). The data were generated by an age-structured model, where a 15-age population was projected over some 40 years but data for only the last 30 years were retained. Details of the data generation are given in Chapter 5 and Appendix E of the NRC report, and the main features are summarised in Table 1. Each data set is a single replication of a combination of stochastic processes[2]. A special comment applies to data set 3, which involves a change in survey vessel (and a near doubling of survey q), a feature that was not explicitly disclosed to the FISBOAT analysts initially and was a clear violation of a basic assumption in their methods; however, given the knowledge of a step change in q, all methods are able to deal with this situation and most authors repeated the analysis later with each period treated as a distinct survey (run labelled "set 3.2" hereafter), which resulted in improved performance. Also note that data set 5 simulates a case with very low exploitation rate (Yield/Biomass ratio in Table 1).

Since some NRC sets are rather tough, a "clean" set (labelled # 6) was added where survey q has been strictly constant, and indices at age measured without error. This was also generated with an age-structured model comprising 15 age groups, and twenty years of data were output. This set was mostly intended to check that the methods' code worked properly.

The data sets were circulated to methods' authors in advance of a project workshop. The main information provided was the matrix of survey indices by age and year. Weights at age, natural mortality (average for the NRC sets, where M varied randomly) and a maturity ogive were also provided in case methods needed these data, but no information about catches and effort by the fishery was given. Of course, the "true" (simulated) population states were only known to the coordinator. It was proposed that analysts focus on the following outputs for comparisons: time series of recruitment (preferably in number); time series of total biomass and, if possible, of total numbers; optionally, time series of spawning stock biomass (SSB).

3.2 Performance metrics

The intention behind selecting the NRC test sets was that comparisons might be possible with the performance of catch-based assessment methods as documented in the NRC report. Since the latter methods are deemed to provide absolute estimates of key management variables, the NRC Committee chose to evaluate the methods based on relative error statistics (i.e. [(estimated – true)/true], both estimates and truth being in absolute value). For F-I methods, however, a clear message from all authors was that these could only provide estimates of relative trends in population variables, and thus the statistics above could not be used. Alternatively, the following performance metric involving relative values was considered: for each quantity of interest, the time series of estimates and of true values were first normalised by subtracting their respective means and dividing by their standard deviations

[2] The report of the 2007 Methods WG (ICES CM 2007/RMC:04, Sect. 2.1.2) may leave the impression that the test data were not corrupted with noise. We point out that the NRC sets 1-5 did include various elements of noise, with perhaps the most relevant for this test being a random lognormal error on the survey indices at age with a 30% CV. Only set 6 was "clean".

Table 2. Performance statistics (3 metrics) of the fishery independent (F-I) stock assessment methods.

		Set						
	Method	1	2	3.1*	3.2*	4	5	6
Root-mean-square of normalised deviations for recruits and biomass								
Recruits	BREM	0.559	0.435	0.775	0.744	0.540	0.548	0.001
	SURBA	0.481	0.466	0.752	0.725	0.462	0.495	0.121
	TSA	0.556	0.441	0.747		0.486	0.536	0.039
	YCC	0.504	0.781	0.621	0.542	0.722	0.461	0.361
Biomass	BREM	0.207	0.211	0.805	0.524	0.194	0.197	0.012
	SURBA	0.402	0.500	0.930	0.892	0.434	0.564	0.031
	YCC	0.182	0.187	0.869	0.347	0.135	0.146	0.152
CV (in %) on recruits and biomass estimates (average over years)								
Recruits	BREM				62.3			
	SURBA	18.7	21.7	22.7	15.5	20.7	18.1	3.0
	TSA	13.4	18.8	15.9		16.8	16.2	0.05
	YCC	44.2	10.5	10.3	15.3	11.0	8.4	23.5
Biomass	BREM	46.3	54.6	39.9	69.2	12.1	37.0	14.4
	TSA **	9.5	11.7	11.2		11.9	10.9	0.04
Relative error (in %) in depletion rate (biomass in final year / in year 1)***								
Biomass	BREM	**–22.6**	**–3.9**	193.1	121.2	**15.7**	40.6	**1.0**
	SURBA	–30.8	31.4	80.2	77.6	–39.8	**–20.0**	**–5.0**
	YCC	**–20.1**	42.5	137.9	**–3.5**	**2.0**	32.7	**0.3**

Notes * 3.1: set 3 assuming a single consistent survey; 3.2: survey split in two (before/after change in vessel); ** CV of geometric-mean stock number over ages; *** Results in boldface meet NRC ± 25% criterion.

which gives a common scaling; the square root of the mean squared deviation between the normalised estimates and the normalised truth was taken as the summary statistic (kind of root mean square error, RMSE). Although this statistic is not readily interpretable to gauge the performance against standard criteria, it enables fair comparisons between the F-I methods.

The biomass depletion rate, that is the estimate of biomass in the final year divided by that in the first year, as considered in the NRC tests should in principle be the same when based on absolute or relative estimates and was also retained as an indicator for comparisons (for those FI methods yielding biomass estimates), together with the NRC mild criterion that the relative error compared to the true rate should be within ± 25%.

As a further comparison, the estimation CVs for recruitment and biomass obtained for each data set were tabulated for those methods that could provide them.

4 Results of methods comparisons across simulated data sets

The relative performance of the F-I methods is summarised in Table 2 for each of the performance metrics described above. Graphical comparisons of the trajectories of estimates vs. the truth (both normalised) are also shown to gain more detailed insight into the behaviour of each method (Figs. 1 and 2).

The first thing to note is that most methods did very well with the clean set 6 (only *YCC* showed some inconsequential deviations for recruitment estimates), which is reassuring: this validation test indicates that there is no inherent defect in the rationale of these methods, nor in the computer code.

These methods essentially behave as smoothers for noisy indices, and may miss quick transient changes in stock abundance. However, in their expected usage to evaluate "current" stock state by comparing present and historic estimates, none would have caused managers to be misled about the situation of the stock and actions to take in the last decade of the time series. For recruitment, the position of weak or strong year-classes is generally correct, although there are cases of either over-smoothing or over-reaction to the signal in the survey.

Like most VPA tuning methods, these F-I methods make the strong assumption that survey q (by age or stage) is constant over time, and it is unsurprising that estimates were badly biased in the tests with set 3.1, where the large step change in q was ignored. In normal circumstances, the assessors would be aware of such marked changes in the survey procedure and would adjust the treatment of their data accordingly, as exemplified by the runs redone as 3.2. Nevertheless, this test highlights the fact that F-I methods are strongly dependent on the quality of the survey, notably the consistency of the survey protocol, as they use no other source of information which might counterbalance poor survey data. In reality, year-on-year variations in survey design (e.g. due to weather or logistic constraints) or gear rigging are common, and users of F-I methods should be alert that they must take them into account, however benign they may first appear.

In contrast, the test indicates no particular problem with set 5, a case with very low exploitation rate ($F \ll M$) which may cause poor convergence of VPA based methods.

Overall, based on inspection of summary statistics and patterns in the plots, all the methods tried in this test performed quite similarly and could be used interchangeably, depending on availability and familiarity with the software. There is a

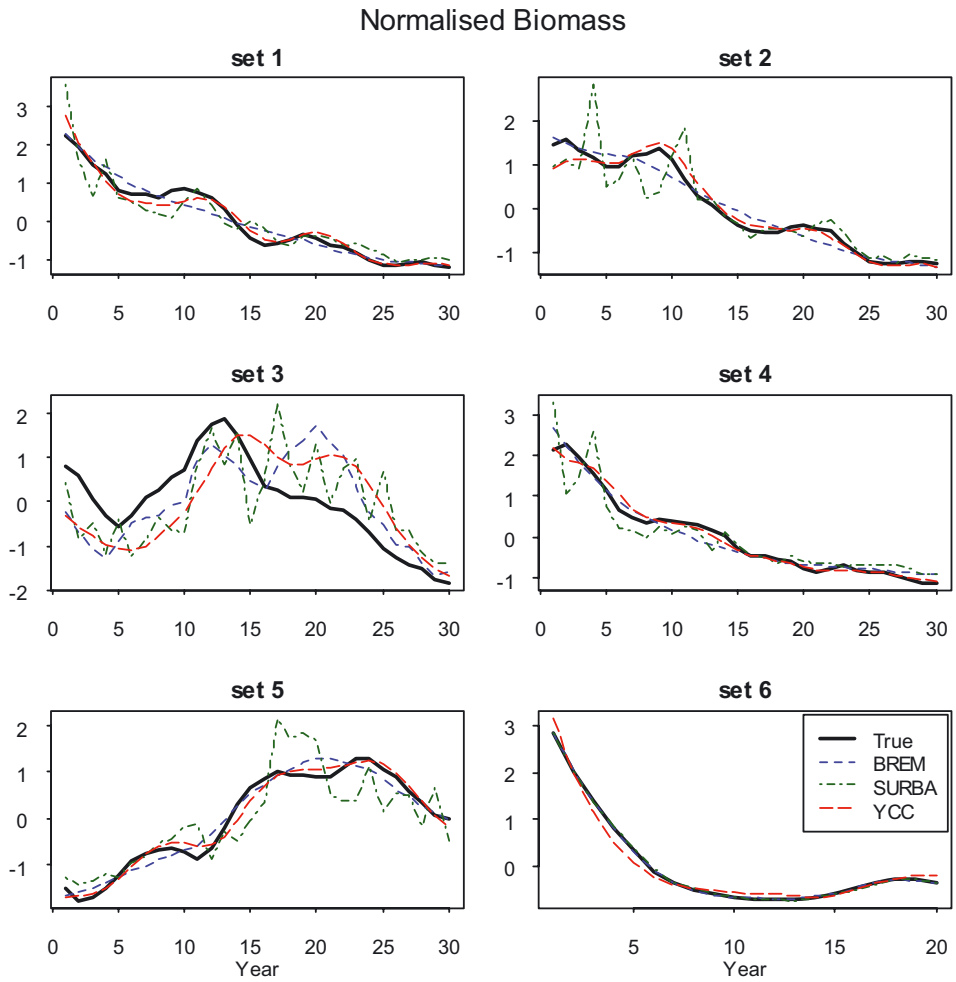

Fig. 1. Comparison of normalised series of biomass estimates across methods and datasets.

small practical advantage in favour of *BREM* which does not require extensive age compositions. Moreover, *TSA* does not (yet) provide biomass trajectories, and the plots of *SURBA* estimates show occasional wiggliness in some batches of years.

It is not straightforward to compare the performance of the F-I methods with those of the tuned catch-based methods applied to the same data in the NRC tests, since estimates from the latter are not available in tabular form. Coarse comparisons with the biomass trajectories plotted in Appendix I of NRC (1998) indicate that catch-based methods tended to consistently over- or (most often) under-estimate stock abundance relative to the truth, whereas F-I estimates wander about the true trajectory. Note in passing that with set 5, all catch-based methods under-estimated the true absolute biomass by a considerable amount, but may have preserved the relative trend. More direct, albeit not necessarily easier, comparisons can be made with the estimates of depletion rate for those NRC runs where only the survey data (not the commercial CPUE series not considered here) were used for tuning. F-I methods, notably *BREM*, perform comparatively well and were generally outperformed only by the most highly parameterised catch-based methods.

It must be kept in mind that this evaluation is contingent on, among other things, scenarios where the error in observation of the indices has a CV of 30%, a value considered reasonable for well-behaved surveys. If the methods are applied to survey data with larger errors, across the series or in specific years, their reliability in advisory contexts will obviously be poorer. In addition, the evaluation is based on a single replication of a stochastic data generation, and a proper evaluation would require summarising over many replicates (Hillary 2009). We note, however, that our protocol is the same as the one adopted by the NRC.

5 Conclusion

Although rudimentary, and awaiting further evaluation in full-fledged management strategy evaluation simulations, this exercise indicates that the F-I methods developed for this project are promising in terms of usefulness and reliability as bases for management advice.

Their main advantage, and indeed their raison-d'être, is that they are not subject to uncertainties in the commercial catches which have caused growing concern and controversies about scientific advice based on VPA approaches in recent years. Moreover, the dependence on catch data is the main reason for the current one-year delay between "data year" and

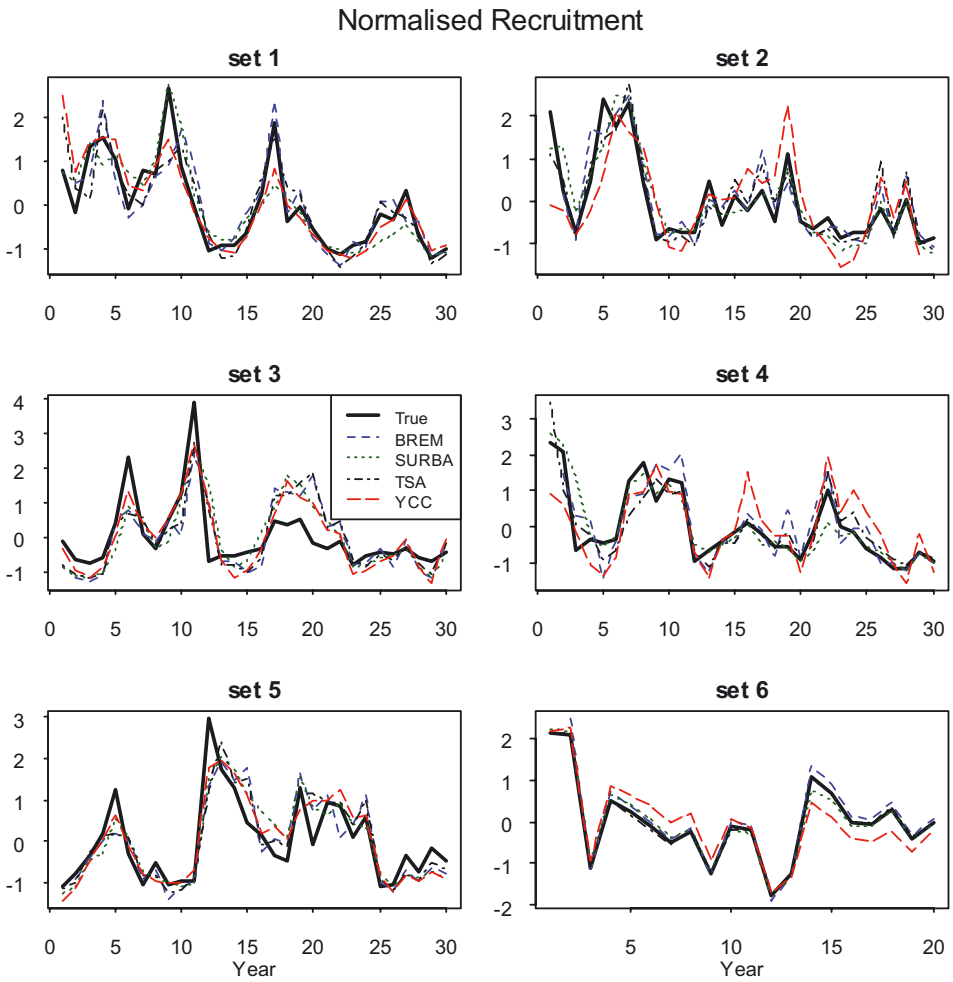

Fig. 2. Comparison of normalised series of recruitment estimates across methods and datasets.

"assessment year", which attracts criticism by managers that response from scientists to their requests is too slow. Clearly, survey-based methods can resolve this timeliness issue, as updated information on stock state is generally available in a matter of days after a survey is completed, although some overhead is still needed for data auditing, construction of the total area index when this involves more elaborate treatments than just aggregating samples, and mostly for age reading for those FI methods requiring detailed age compositions. Another bonus with all the methods reviewed here is that their fitting procedures do not require prior knowledge of the natural mortality coefficient M, which is a crucial ingredient in many other assessment methods and perhaps the most challenging parameter to estimate (M may still be needed for derived quantities, such as extracting fishing mortality F if management specifically needs it). Finally, it can be seen as an advantage that the methods reviewed have few if any "tuning knobs" to fiddle with.

Evidently, there are a few drawbacks. In particular, it is not possible to estimate absolute stock size (overall or for specific ages): all abundance estimates are relative, with an arbitrary scaling coefficient (= survey q) between actual and estimated abundance. In itself, this is not necessarily an issue, and examples might easily be found in many areas where decisions of utmost importance to society are made in response to relative indicators. The problem with fisheries management in Europe merely arises because, decades ago, scientists successfully sold the idea that they had the skills to deliver advice in absolute terms and the "system" has been built-up on these premises. One consequence is that managers were never educated to make use of alternative flows of information, such as relative indicators coupled with reference points based on past states (if only as a cross-check of the traditional advice). More seriously, scientists have never formalised and evaluated an advisory process based on such information, although many critics argue that allegedly absolute VPA estimates are effectively relative since they are scaled by input M's which are guessed rather than known. However, this is mostly a problem with the advisory system and it should not count against the performance of the F-I methods *per se*. If any survey-based approach is to be used as a management tool, there needs to be a clear idea of the management framework in which such a tool would be used. In other words, reference points for mortality and biomass would need to be redefined on the basis of total mortality and relative biomass, respectively.

A more inherent limitation of F-I methods is that they only use one source of information, and are thus critically dependent on the quality of survey protocols and data. Perceived

year-on-year changes in abundance, and ensuing effects on advised management decisions, are likely to be very fragile to inconsistencies in the conduct of surveys (dates, geographical coverage, gear, etc.), and the best professional standards must be adhered to in order to reduce biases. When survey programmes are directed at groups of species (e.g. IBTS, the International Bottom Trawl Survey programme in the North Sea), the design tries to achieve a compromise between the needs of various species, and there are often populations whose distribution is only partially covered; this potential bias has to be borne in mind when candidate species are selected for application of F-I methods (and in any case when interpreting the results for advice). Finally, despite the complaint by paymasters that surveys are by far the most costly item in the assessment process, the implication of basing management on F-I approaches may well be that more, rather than less, investment in surveys is required notably for those where the precision of indices is near the limit of acceptability. Although gaps in survey data do not technically impede estimation with the methods reviewed, it is obvious that the quality of assessments degrades quickly when gaps occur frequently, and that the "current" state of stocks cannot be appraised in those years when data are missing. As a rule, surveys should be annual to be usable safely in the deplorably polemical context of fisheries management.

Although the F-I methods seem promising, we do not propose that they should replace the conventional assessment methods in all cases. They are designed primarily to fill a gap, where catch data are incomplete or too unreliable and catch-based assessments are not feasible or too uncertain for acceptance by stakeholders. We strongly suggest, however, that they be used alongside catch-based methods to validate their results. For example, differences in estimates may be an indication that catches were misreported or natural mortality misspecified in the standard model. The technical and institutional implications of using F-I methods in fisheries management are further discussed in Cotter et al. (2009).

Acknowledgements. This work was made possible with the financial contribution of the EC to the institutes participating in the FISBOAT project (research project DG-Fish, STREP No. 502572 (2004-2007)). Dr Terry Quinn (University of Alaska, Juneau, USA) kindly provided the NRC simulated data sets. We are grateful to the referee and the editor for their precious contributions to improving the paper.

References

Beare D.J., Needle C.L., Burns F., Reid, D.G., 2005, Using survey data independently from commercial data in stock assessment: An example using haddock in ICES Division VIa. ICES J. Mar. Sci. 62, 996–1005.

Beverton R.J.H., Holt S.J., 1957, On the dynamics of exploited fish populations. UK Minist. Agric. Fish., Fish. Invest. (Ser. 2), 19, 533 p.

Cook R.M., 1997, Stock trends in six North Sea stocks as revealed by an analysis of research vessel surveys. ICES J. Mar. Sci. 54, 924–933.

Cook R.M., 2004, Estimation of the age-specific rate of natural mortality for Shetland sandeels. ICES J. Mar. Sci. 61, 159–164.

Cotter A.J.R., 2001, Intercalibration of North Sea International Bottom Trawl Surveys by fitting year-class curves. ICES J. Mar. Sci. 58, 622–632 [Erratum, Ibid. 58, 1340].

Cotter A.J.R., Buckland S.T., 2004, Using the EM algorithm to weight data sets of unknown precision when modeling fish stocks. Math. Biosci. 190, 1–7.

Cotter J., Mesnil B., Piet, G., 2007, Estimating stock parameters from trawl CPUE-at-age series using year-class curves. ICES J. Mar. Sci. 63, 234–247.

Cotter A.J.R., Petitgas P., Abella A., Apostolaki P., Mesnil B., Politou C.-Y., Rivoirard J., Rochet M.J., Spedicato M., Trenkel V.M., Woillez M., 2009, Towards an ecosystem approach to fisheries management (EAFM) when trawl surveys provide the main source of information. Aquat. Living Resour. 22, 243–254.

Darby C.D., Flatman S., 1994, Virtual population analysis: version 3.1 (Windows/DOS) user guide. CEFAS, Lowestoft, UK. Information Technol. Ser. N° 1.

Deriso R.B., Quinn T.J.II, Neal, P.R., 1985, Catch-age analysis with auxiliary information. Can. J. Fish. Aquat. Sci. 42, 815–824.

Fryer R.J., 2002, TSA: is it the way? Appendix D in Report of Working Group on Methods of Fish Stock Assessment, Dec. 2001. ICES CM 2002/D:01, 86–93.

Fournier D., 2005, An introduction to AD MODEL BUILDER version 7.0.1 for use in nonlinear modeling and statistics. Available from http://otter-rsch.com/admodel.htm.

Gudmundsson G., 1986, Statistical considerations in the analysis of catch-at-age observations. J. Cons. Internat. Explor. Mer 43, 83–90.

Gudmundsson G., 1994, Time series analysis of catch-at-age observations. Appl. Stat. 43, 117–126.

Gudmundsson G., 2004, Time-series analysis of abundance indices of young fish. ICES J. Mar. Sci. 61, 176–183.

Hilborn R., Walters C.J., 1992, Quantitative fisheries stock assessment: Choice, dynamics and uncertainty. New York, Chapman and Hall.

Hillary R., 2009, An introduction to FLR fisheries simulation tools. Aquat. Living Resour. 22, 225–232.

Johnson S.J., Quinn T.J. II, 1987, Length frequency analysis of sablefish in the Gulf of Alaska. Technical Report UAJ-SFS-8714, University of Alaska, School of Fisheries and Science, Juneau, Alaska. Contract report to Auke Bay National Laboratory.

Needle C.L., Hillary R., 2007, Estimating uncertainty in nonlinear models: applications to survey-based assessments. ICES CM 2007/O:36.

NRC, 1998, Improving fish stock assessments. Washington, D.C., National Academy Press. (Appendix E describes the data generation; Appendix I shows plots of biomass trajectories).

Patterson K.R., Melvin G.D., 1996, Integrated Catch At Age Analysis Version 1:2. Scottish Fisheries Research Report. FRS: Aberdeen.

Pope J.G., Shepherd J.G., 1982, A simple method for the consistent interpretation of catch-at-age data. J. Cons. Internat. Explor. Mer 40, 176–184.

Quinn T.J.II, Deriso R.B., 1999, Quantitative Fish Dynamics. Oxford, Oxford University Press.

Skaug, H.J., Fournier, D.A., 2006, Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. Comput. Stat. Data Anal. 51, 699–709.

Trenkel V.M., 2007, A biomass random effects model (BREM) for stock assessment using only survey data: application to Bay of Biscay anchovy. ICES CM 2007/O:03.

Trenkel V.M., 2008, A two-stage biomass random effects model for stock assessment without catches: What can be estimated using only biomass survey indices? Can. J. Fish. Aquat. Sci. 65, 1024–1035.