African Journal of Marine Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tams20

A bootstrap method for estimating bias and variance in statistical fisheries modelling frameworks using highly disparate datasets

BÞ Elvarsson[ab], L Taylor[a], VM Trenkel[c], V Kupca[d] & G Stefansson[a]

[a] Science Institute, University of Iceland, Reykjavik, Iceland

[b] Marine Research Institute, Reykjavik, Iceland

[c] Ifremer [French Research Institute for the Exploitation of the Sea], Nantes, France

[d] HPC2N [High Performance Computing Center North], Umeå University, Umeå, Sweden

Published online: 24 Apr 2014.

To cite this article: BÞ Elvarsson, L Taylor, VM Trenkel, V Kupca & G Stefansson (2014) A bootstrap method for estimating bias and variance in statistical fisheries modelling frameworks using highly disparate datasets, African Journal of Marine Science, 36:1, 99-110, DOI: 10.2989/1814232X.2014.897253

To link to this article: http://dx.doi.org/10.2989/1814232X.2014.897253

African Journal of Marine Science 2014, 36(1): 99–110
Printed in South Africa — All rights reserved

AFRICAN JOURNAL OF MARINE SCIENCE
ISSN 1814-232X EISSN 1814-2338
http://dx.doi.org/10.2989/1814232X.2014.897253

A bootstrap method for estimating bias and variance in statistical fisheries modelling frameworks using highly disparate datasets

BÞ Elvarsson[1,2]*, L Taylor[1], VM Trenkel[3], V Kupca[4] and G Stefansson[1]

[1] *Science Institute, University of Iceland, Reykjavik, Iceland*
[2] *Marine Research Institute, Reykjavik, Iceland*
[3] *Ifremer [French Research Institute for the Exploitation of the Sea], Nantes, France*
[4] *HPC2N [High Performance Computing Center North], Umeå University, Umeå, Sweden*
* *Corresponding author, e-mail: bjarki.elvarsson@gmail.com*

Statistical models of marine ecosystems use a variety of data sources to estimate parameters using composite or weighted likelihood functions with associated weighting issues and questions on how to obtain variance estimates. Regardless of the method used to obtain point estimates, a method is required for variance estimation. A bootstrap technique is introduced for the evaluation of uncertainty in such models, taking into account inherent spatial and temporal correlations in the datasets, which are commonly transferred as assumptions from a likelihood estimation procedure into Hessian-based variance estimation procedures. The technique is demonstrated on a real dataset and the effects of the number of bootstrap samples on estimation bias and variance estimates are studied. Although the modelling framework and bootstrap method can be applied to multispecies and multiarea models, for clarity the case study described is of a single-species and single-area model.

Keywords: bootstrapping, correlated data, fish population dynamics, non-linear models

Online supplementary material: The technical details of the model are available in Supplementary Appendix A at http://dx.doi.org/10.2989/1814232X.2014.897253.

Introduction

Statistical models consolidate data from various sources by using them simultaneously to estimate parameters. The importance of using all data in a single model has been emphasised by several authors (Methot 1989; Demyanov et al. 2006) but, although the benefits are clear, it is certainly not without problems, including the question of variance estimation, model mis-specification and weighting of all data sources (Stefansson 2003; Francis 2011; Maunder and Punt 2013). In the context of complex population dynamics models of exploited marine species, multiple data sources with widely different properties are used routinely in the estimation process.

Variance estimates of parameters in non-linear models have commonly been derived from the inverted Hessian matrix at the optimum, when the method of least squares (or maximum likelihood) is employed for parameter estimation. Alternatively the Jacobian matrix of the residuals can be used. Several conditions need to be satisfied for statistical inference, e.g. confidence statements to hold in the finite-sample case. First, the model needs to be correct. Second, variance assumptions, i.e. homoscedasticity and knowledge of the ratios of variances in individual datasets, need to be appropriate.

Methods of estimating variances in fish stock assessment models have been discussed and evaluated by many authors, including Gavaris et al. (2000), Gavaris and Ianelli (2001), Magnusson et al. (2013) and Patterson et al. (2001). When the distributional properties of the data are not well understood or the models are incorrect, Hessian-based approaches have been seen to fail in several examples in fishery science (Patterson et al. 2001). Although this may seem to contradict the theoretical statements, the assumptions – e.g. in Jennrich (1969) – include independence of observations, a unique minimum, identically distributed errors and, of course, the results are only asymptotic. Any of these assumptions may fail. It follows that for problems in fishery science one cannot assume *a priori* that a Hessian-based method will give reasonable results. For example, disregarding correlation structure when present has been found potentially to lead to incorrect conclusions in single-species assessments, sometimes with serious consequences (Myers and Cadigan 1995). Similarly, multimodal likelihood functions have been seen in real applications (Richards 1991) and typically correspond to incorrect model assumptions that are not detected with traditional analysis (Stefansson 2003) but may potentially be detected if histograms of bootstrap parameter estimates also become multimodal (see example in Hannesson et al. 2009).

Many of the limitations of the Hessian-based approaches have been met by alternative methods. In particular, models developed using the Bayesian framework (as discussed in

e.g. Punt and Hilborn 1997) provide an elegant formulation of uncertainty as posterior distributions of the quantity of interest. In all but trivial cases the posterior distribution must be estimated numerically with methods such as Markov chain Monte Carlo. With the commoditisation of computers in conjunction with the development of frameworks such as BUGS (Spiegelhalter et al. 1996) and ADMB (Fournier et al. 2012), the Bayesian framework has become a popular alternative to Hessian-based uncertainty methods. The attraction of the Bayes inference stems, to some degree, from the ability to include prior belief/knowledge into the model as explicit distributions. Various sources (e.g. Chen et al. 2000; Millar 2002) suggest, however, that considerable care must be taken when choosing model priors to avoid mis-specification and suggest a suite of robust priors applicable in fisheries model setting.

Alternative frequentist approaches to Hessian-based parameter variance estimation include bootstrap methods (Efron 1979; Efron and Tibshirani 1994). The simplest bootstrap method assumes that the data are independent measurements without correlation. However, semi-parametric approaches have also been developed to sample residuals from a model, possibly from a distribution (parametric bootstrap) or with a known correlation structure (Davison and Hinkley 1997).

This paper demonstrates a novel use of bootstrapping to address complex and disparate data issues. The approach is generic, but it has special application to statistical models of (multiple and interacting) marine populations such as those developed within the Gadget framework. Gadget is a statistical age–length-structured modelling environment originally proposed by Stefansson and Palsson (1998), combining concepts from several earlier methods (Gavaris 1988; Methot 1989; Tjelmeland and Bogstad 1989; Bogstad et al. 1992), described in Begley (2004) and subsequently used in multiple fisheries applications (e.g. Björnsson and Sigurdsson 2003; Taylor et al. 2007; Lindstrøm et al. 2009). The protocol used in Gadget to estimate likelihood component weights and optimise model parameters is described in detail in Taylor et al. (2007) and the weighting protocol is based on that described in Stefansson (1998) and Stefansson (2003).

In the following sections the development of an elementary sampling unit used in the bootstrap is described. The methodology is applied to a Gadget model for cod in Icelandic waters (the standard model from Taylor et al. 2007) and contrasted with a more traditional Hessian-based approximation of variance.

Methods

Development of an elementary sampling unit

Statistical fisheries models may involve the use of a large number of data from a variety of sources. Every sample from each data source can be classified according to sampling location and time. A model such as Gadget operates on certain time-steps and also uses some spatial units. Within any modelled spatio-temporal unit there will normally be several data samples. For any bootstrap method the first question is, therefore, what the sampling unit should be. A unit of measurement in marine studies tends to be based on a single fish and elementary resampling might bootstrap on individual fish (as in e.g. Gudmundsdóttir et al. 1988). Doing this assumes that all individually measured fish are independent, which is invalid for several reasons (Pennington and Volstad 1994; Hrafnkelsson and Stefansson 2004). Resampling entire fish samples within a haul (W Singh, Science Institute, University of Iceland, Reykjavik, unpublished data) can potentially be used to account for this intra-haul correlation. Appropriate analyses of variance can correspondingly be used to evaluate these effects (Helle and Pennington 2004; De Croos and Stefansson 2011) and, when combining samples, alternatives to simple sums or means may be needed for aggregation (Babak et al. 2007). However, considering samples as units may not be quite enough, since fish at close geographic locations will also tend to be similar due to a fine-scale spatial structure that cannot be modelled easily (e.g. Stefansson and Palsson 1997a).

In addition to the sampling unit problem, one needs to take into account the variety of data sources. Biological samples from commercial catches may be collected on a fine temporal and spatial scale whereas scientific surveys are typically conducted only once or twice a year and different surveys may or may not overlap spatially. Other datasets such as species composition of stomach contents or tagging experiments may be collected at completely different resolutions to age or length data.

Here, the proposed sampling unit is based on spatial structure on the Icelandic coastal shelf developed by Taylor (2003), shown in Figure 1, where the areas within the gridlines are referred to as ‘subdivisions’. The spatial structure is based mainly on bathymetry, hydrography and species assemblages with some further disaggregation defined by fishing regulations. In this context an ‘elementary sampling’ unit consisted of all data collected inside a subdivision within a time period of interest. Subdivisions and elementary sampling units are therefore used interchangeably. In order to reduce correlations between the elementary sampling units, aggregations are made. For example, to remove within-sample correlations between length groups (Hrafnkelsson and Stefansson 2004), only (combinations of) entire length samples are used, rather than lengths of individual fish. Similarly, data are aggregated within the fairly large spatial areas and the shortest time-step is at least one month. This is intended to eliminate intra-haul correlations (Pennington and Volstad 1994) and those correlations between age-groups (Myers and Cadigan 1995) that are related to local shoals or small feeding patches.

To generate input files for Gadget, a second aggregation method is applied on the elementary sampling units, that is all data from a particular subdivision, which varies somewhat depending on the data source. Some data types, e.g. length distributions, are simply added, whereas others, such as mean length-at-age, may go through a computational mechanism involving age–length keys. A description of a fisheries database, able to handle data aggregations in this manner, is provided by Kupca and Sandbeck (2003) and Kupca (2004).

Here, the fundamental idea is the aggregation of elementary sampling units in the creation of model inputs. These sets of elementary sampling units can therefore be sampled

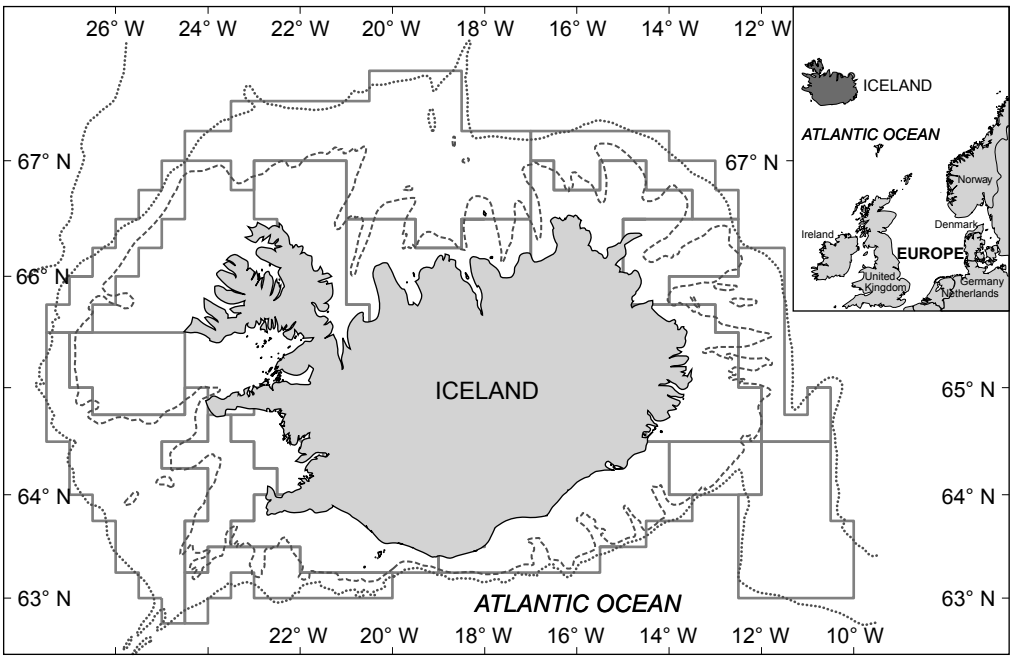

Figure 1: The spatial structure of data storage on the Icelandic coastal shelf together with 200 m (broken line) and 500 m (dotted line) depth contours. The areas within the gridlines are referred to as 'subdivisions'. A given time period, time-step size and subdivision are referred to collectively as an 'elementary sampling unit'

(with replacement) before aggregation, with each resample leading to a new model input dataset. A typical model run for parameter estimation based on such a resampled dataset will result in a resampled parameter estimate. The collection of all such estimates forms a bootstrap sample. The procedure could be called a 'spatio-temporal block bootstrap with unequal block size'.

A fisheries example

The setting

The example marine system used in this paper is based on cod in Icelandic waters (Figure 1) with an approach very similar to Taylor et al. (2007). The model consists of two stock components of cod, i.e. immature and mature cod in a single area. Modelling maturity enables the calculation of spawning stock biomass and allows different weight–length relationships to be used for immature and mature fish.

Two fixed-station surveys are used to monitor the stock – in spring and autumn – providing population indices as well as biological samples. Landings information is available from official databases and raw biological data (length distributions, age compositions), together with survey data, are in the Marine Research Institute's (MRI) databases (see e.g. Palsson et al. 1989, Sigurdsson et al. 1997, Taylor et al. 2007 and ICES 2011 for a description of data and surveys). The technical details of the model are described in Supplementary Appendix A (available online).

The dataset and parameters

The model is a parametric and deterministic forward population dynamics simulation model. A single simulation results in a complete population structure, including predictions of all datasets, as described in Begley (2004) and Taylor et al. (2007), and a corresponding evaluation of a (negative log-) likelihood function (sums of squares in the present paper).

With the exception of landings data, datasets are used only in the likelihood components. For simplicity, landings data are used directly in the population models, whereby the populations are simply reduced in numbers to be in accordance with the corresponding landed weight. Note that, in the approach proposed here, the landings data are not resampled.

An overview of the datasets and model parameters used in this case study is shown in Tables A.1 and A.2, respectively, in Supplementary Appendix A.

Estimation protocol

The weights on the likelihood components are calculated for each model (i.e. each bootstrap run), according to the protocol described in Supplementary Appendix A, Section A.2.4 (available online), with arbitrary starting parameters. This is a two-stage estimation method, where the error variances, within a dataset, are estimated by increasing the weight on that particular component of the total sum of squares, followed by a final minimisation using those inverse variances as weights. For a full description of this procedure refer to Supplementary Appendix A.

The bootstrapping approach consists of the following:

- The base data are stored in a standardised database:
  - Time aggregation: 3 months
  - Spatial aggregation: subdivision
  - Further disaggregation is based on a range of categories including fishing gear, fishing vessel class, sampling type (e.g. harbour, sea or survey). A full listing

of data types used in the case study can be found in Table A.1 of Supplementary Appendix A. These data are stored subdivision-disaggregated to allow for use in a bootstrap.

- To bootstrap the data, the list of subdivisions, depicted in Figure 1, required for the model is sampled (with replacement) and stored. For a multi-area model one would conduct the resampling of subdivisions within each area of the model.
- The list of resampled subdivisions is then used to extract data (with replacement so the same dataset may be repeated several times in a given bootstrap sample).
- For a single bootstrap Gadget model, the same list of resampled subdivisions is used to extract each likelihood dataset; i.e. length distributions, survey indices and age–length frequencies are extracted from the same spatial definition.
- A Gadget model is fitted to the extracted bootstrap dataset using the estimation procedure described above.
- The resampling process is repeated until the desired number of bootstrap samples is extracted.

When resampling, data are forced to remain in the correct year and time-step, so resampling is based on sampling spatially the elementary data units within a given modelled unit of time and space. Thus, within a modelled spatial unit, the bootstrap is a resampling of subdivisions. This implicitly assumes data contained within each area of the model to be independent and identically distributed. Independence is justified by the definition of subdivisions. Furthermore, treating them as if they were from the same distribution, i.e. bootstrap replicates, appears to have little negative effect when compared to more traditional methods (Taylor 1999).

The entire estimation procedure is repeated for each bootstrap sample. In particular, since the estimation procedure includes an iterative reweighting scheme, this reweighting is repeated for every bootstrap sample. The point of this is that the bootstrap procedure is no longer conditional on the weights. The procedure as a whole is quite intensive computationally but can easily be run in parallel, e.g. on a computer cluster.

In stark contrast to this, Hessian-based approaches usually compute the Hessian only at the final solution. Thus, they completely omit the effect of reweighting likelihood components when estimating uncertainty. Such methods are thus conditional on the weights obtained in a pre-estimation stage.

Application of the bootstrap procedure and its variants

The bootstrap procedure presented here is, as noted earlier, quite demanding computationally as the number of bootstrap samples increases. In this exercise 1 000 bootstrap samples were chosen as the 'baseline' simulation. This number of iterations was chosen as a practical upper limit, as a single optimisation run for a Gadget model takes a substantial amount of time. In addition to the baseline simulation, two sensitivity tests are considered in the present case study. Here it is of considerable interest to study possible reduction in the number of bootstrap samples and other means to reduce the number of calculations. An interesting comparison to the baseline simulation would be to reduce the number of bootstrap samples to 100 samples. A more thorough analysis of the effects of sample size is described below.

Another interesting sensitivity test would be a bootstrap procedure conditional on weights obtained at the pre-estimation stage, i.e. using the same (fixed) likelihood weights throughout the simulation. The reason for this comparison would be twofold: (1) computationally, the number of calculations required would be drastically reduced and (2) a comparison would be made in relation to Hessian-based approaches. One should note, however, that with this bootstrap the estimation is not the same function of the data as the procedure where the weighting takes place for each dataset. This may lead to inappropriate weights for a given dataset which in turn can, as mentioned earlier, lead to inaccurate parameter estimates.

Hessian-based inference

For illustrative purposes the inferences arising from the bootstrap procedure presented here are compared to a Hessian-based confidence interval (described by Tinker et al. 2006, and references therein). In particular, central differences were used to calculate the second derivatives needed to obtain an estimate of the variance–covariance matrix and a multivariate delta method (Oehlert 1992) was used to obtain the confidence interval for derived biomass.
The effects of sample size on the inferences obtained from the inverted Hessian matrix were studied using an artificial increase in measurements. The time-step length was varied between one, two and the baseline three months, with input files being adjusted accordingly. The resulting CVs for the recruitment parameters were estimated and the effects of the different step lengths contrasted. Similar analysis was conducted for the proposed bootstrap procedure but, for the sake of clarity, is discussed only in connection with the Hessian-based approach.

Number of bootstrap samples

With regards to the bootstrap procedure itself this study also examines the effect of the number of bootstrap samples on the variance and bias estimates using a retrospective bootstrap. For a sample number n, ranging from 25 to 1 000 bootstrap samples, n vectors of parameter estimates from the baseline bootstrap were sampled with replacement 100 times. From those 100 samples the coefficient of variation (CV) was calculated for the mean and standard deviation of each parameter. Uncertainty in bias estimation is harder to quantify in a similar way because parameter bias is often estimated close to zero.

Model output

Given the optimised parameter estimates it is possible to output a wide range of descriptors of the model ecosystem because Gadget operates on and stores the number in each age–length cell for each time-step of the model. For this study, the estimated parameters along with a derived biomass trajectory (age 4+) are considered. Comparisons of uncertainty estimates will be made, as noted earlier, using the three bootstrap variants, i.e. both 1 000 and 100 bootstrap simulations with the iterative reweighting

procedure applied to all bootstrap samples and 1 000 bootstrap simulations using fixed weights, as well as the Hessian-based approach. A schematic overview of all calculations performed here is shown in Figure 2.

Results

The simplest model outputs are the point estimates of model parameters. Figure 3 gives histograms of bootstrap

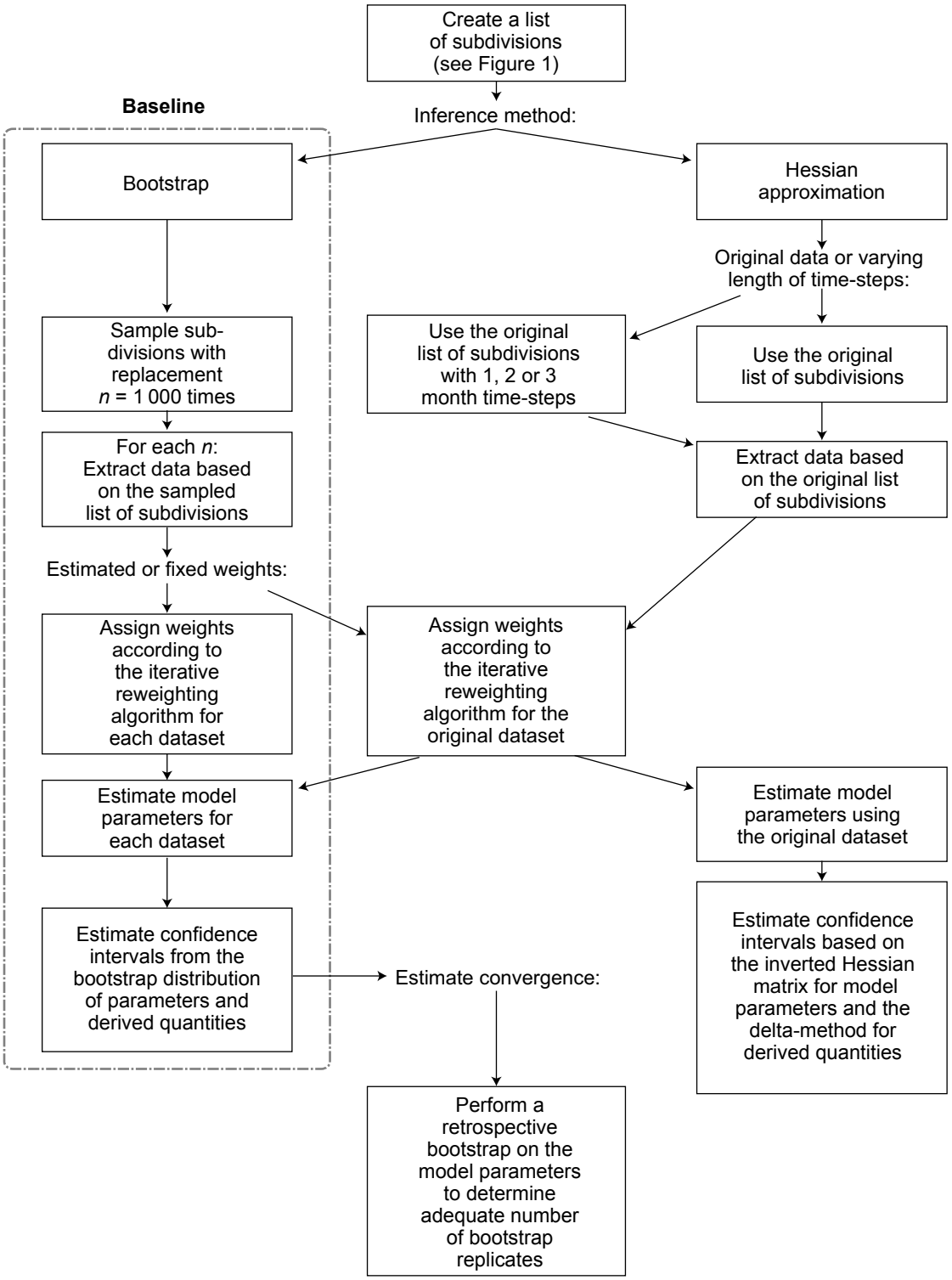

Figure 2: A flowchart of the calculations performed. Boxes indicate action and unbounded text possible uncertainty estimation variants or decisions

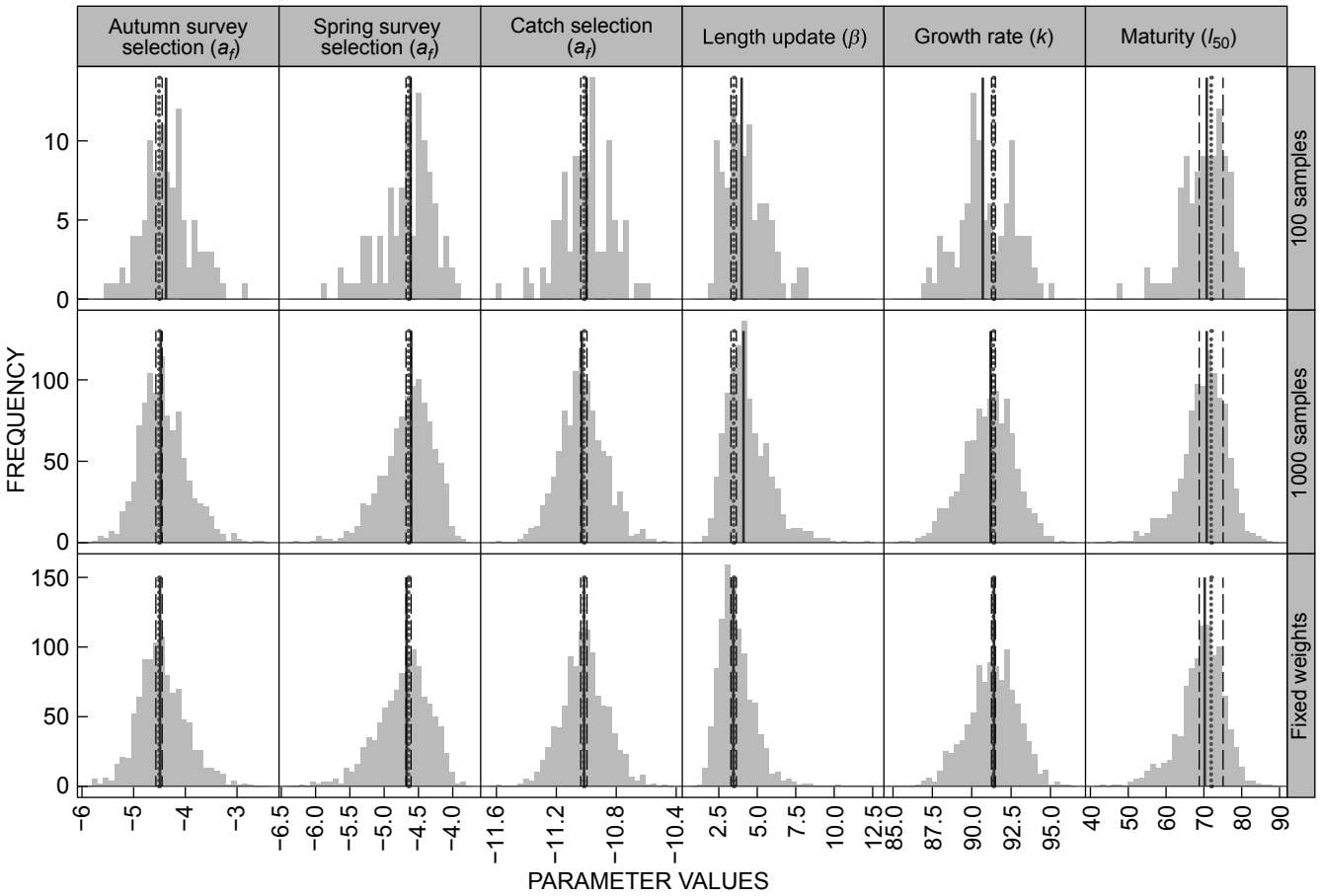

Figure 3: Histograms of the estimated fleet selection parameter a_f for the three fleets (October [autumn] survey, March [spring] survey, commercial catch), β the parameter defining the length update matrix, k the growth rate and l_{50} the maturity. The parameter estimates were obtained from 1 000 bootstrap samples, compared to a smaller number of bootstrap samples, 100, where for both numbers of samples iterative weighting applied to all bootstrap samples. This is then all compared to 1 000 bootstrap samples where, in the parameter estimation, the weighted likelihood function is conditioned on the original weights. The point estimate (grey broken line) and bootstrap mean (black solid line), together with 95% confidence bounds obtained from a Hessian-based approximation to the variance–covariance matrix (black dashed lines), are indicated

estimates of several parameters. It compares the distributions of those parameter estimates from 1 000 bootstrap samples, either using reweighting for each dataset or fixed weights, to those using only 100 samples with reweighting. For each parameter, the point estimate from the full dataset, the median of the bootstrap estimates and 95% confidence intervals from a Hessian-based approximation are indicated. The differences between the point estimate and the bootstrap mean are relatively minor, i.e. there is no obvious sign of an estimation bias, in all cases except for the length update (see β in Supplementary Appendix A, Eqn 3; available online). It should be noted that the maturation parameters are correlated, affecting the relationship between the point estimate and bootstrap mean for the maturation. The different bootstrap methods exhibit similar distribution of parameter estimates, with the exception of the length update, where the bootstrap mean based on the original weights falls closer to the point estimate, thus failing to detect bias in the length update.

Boxplots can be used to illustrate bootstrapped trajectories of various abundance or biomass measures. The estimate of the 4+ biomass is shown in Figure 4. The main variation appears, in absolute terms, in the initial and final years, while only the final year shows a considerable amount of variation in terms of CV. The initial and final years are, of course, considerably different from the intermediate ones, but in different ways. The number of fish in the initial year is part of the estimation procedure and therefore of a different nature when compared to subsequent years. Further, the survey starts in 1985 (with the model starting in 1984), which makes the initial conditions somewhat poorly determined. The final years, on the other hand, are poorly determined, since there is relatively little information in the objective function for the younger year classes because they have been surveyed for only a few years.

The same effects are seen for estimated recruitment at age 1 (Figure 5) where there is less variation in the earlier and intermediate years than the later years. As for the other parameters, the Hessian-based confidence estimates are considerably smaller than those obtained using bootstrap methods. The CVs of the Hessian-based approach followed roughly the same pattern as for those arising from the

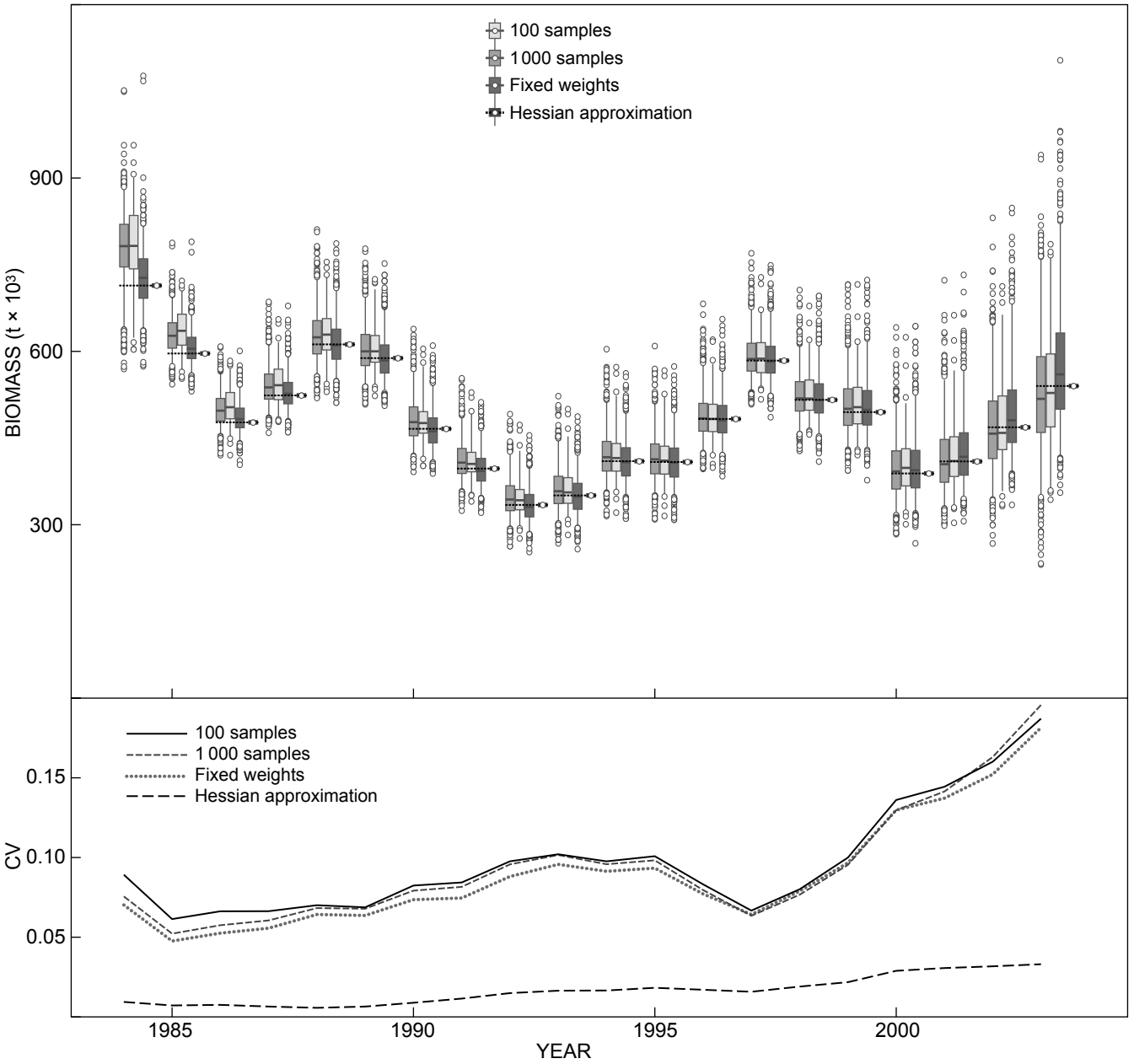

Figure 4: Boxplot (top panel) of the end-of-year biomass for cod of age 4 and older estimated on 1 000 bootstrap samples, both using iterative weighting for each sample and using the fixed weights for all samples, compared to 100 bootstrap samples. The fixed weights were obtained using iterative weighting for the original dataset. The point estimate is indicated by the central black broken line through the boxes. The box indicates the interquartile range and the whiskers 95% confidence intervals. Any further outlying data points are indicated as points. Bottom panel shows the estimated CV for the age 4+ biomass using the same methods as above

various bootstrap approaches but generally were around 12% of the corresponding bootstrap CV.

In Figure 6, CVs for the mean and standard deviation of the model parameters are shown as a function of the number of bootstrap samples, n, where the separate panels show different groups of parameters. The CV estimates appear to fall close to $1/\sqrt{n}$, as shown in the figure, and most of them are less than 15% for 100 bootstrap samples. The initial conditions, i.e. the numbers at age in 1984, had a somewhat higher CV for the mean and standard deviation than the other parameter groups. The initial numbers at age 8 and 9 in 1984 in particular, showed a considerably higher CV for all sample sizes. Those two age groups were, as noted earlier, poorly determined, and had a very low estimate compared to other initial numbers, as the corresponding year classes were present in the data for only the first few years of the model.

Hardly any biases were observed in this analysis. Notable exceptions were the length update parameter, shown in Figure 3, and the first two years of the 4+ biomass, which

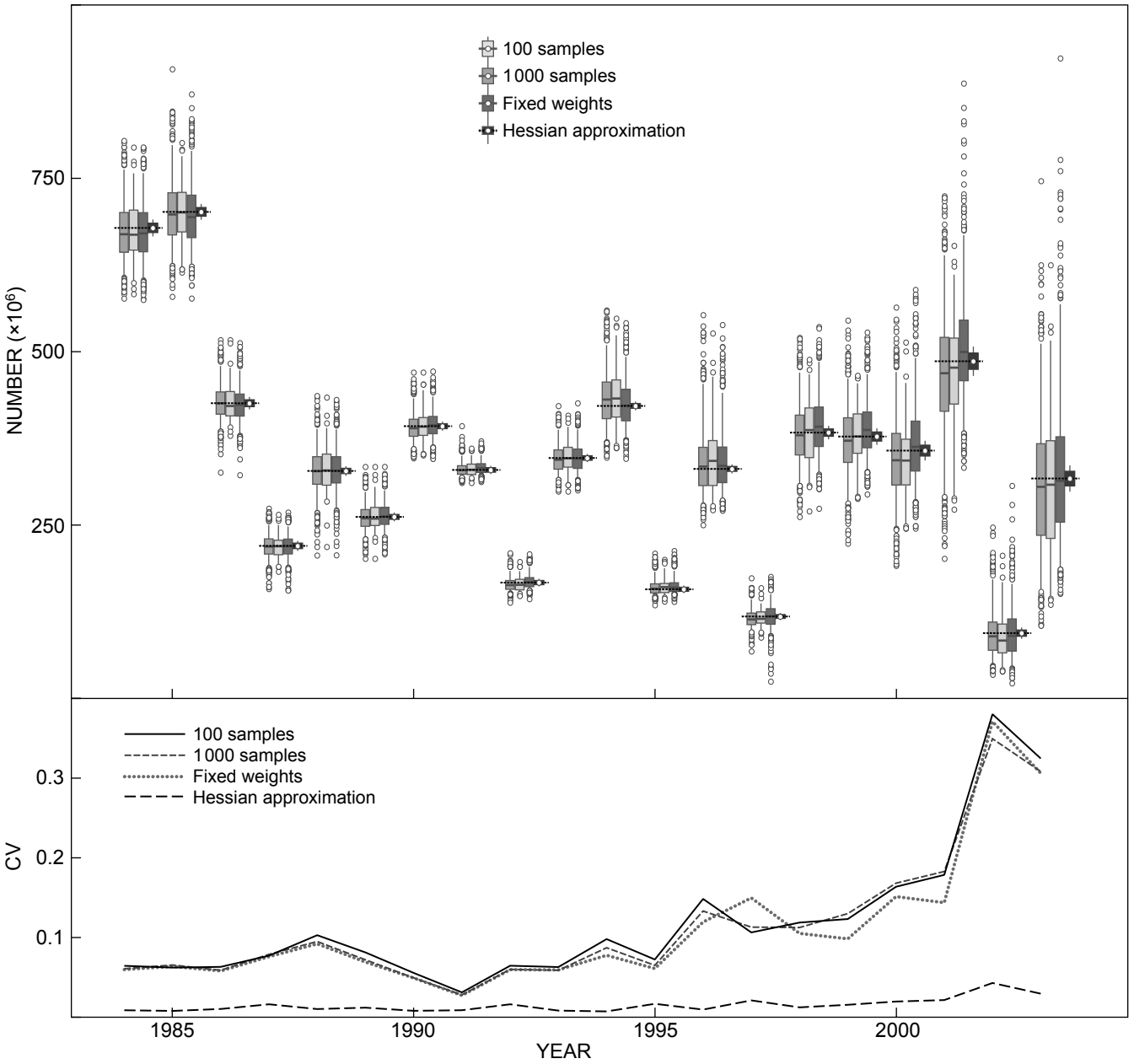

Figure 5: Boxplot (top panel) of the number of recruits (age 1) in each year estimated by 1 000 and 100 bootstrap samples compared to 1 000 bootstraps with fixed weights and a Hessian-based approximation to the 95% confidence interval. The point estimate is indicated by a central black broken line through the boxes. Bottom panel shows the estimated CV for the recruitment using the same methods as above

appeared to have a measurable bias. This was only detected in the bootstrap simulations where the iterative reweighting scheme was applied to all bootstrap samples. The fixed-weight run and the Hessian-based approach failed to detect these differences.

The effects of the number of time-steps within a year can been seen in Figure 7. There the CV of recruitment is illustrated as a function of the number of (intra-year) time-steps in the model. The number of time-steps appears to be inversely proportional to the CV size. These effects were not, when varying the time-step, observed when conducting a similar analysis using the bootstrap.

Discussion

This paper has presented a novel bootstrap method suitable for models of population dynamics. Several modifications and alternatives to the original bootstrap methodology (Efron 1979; Efron and Tibshirani 1994) have been presented. For example, to account for correlations in simple non-replacement sampling schemes (as used for most questionnaires or 'sample surveys'), without-replacement bootstraps and with-replacement bootstraps have been suggested, along with somewhat-more-general resampling procedures for complex survey data (Gross 1980; McCarthy

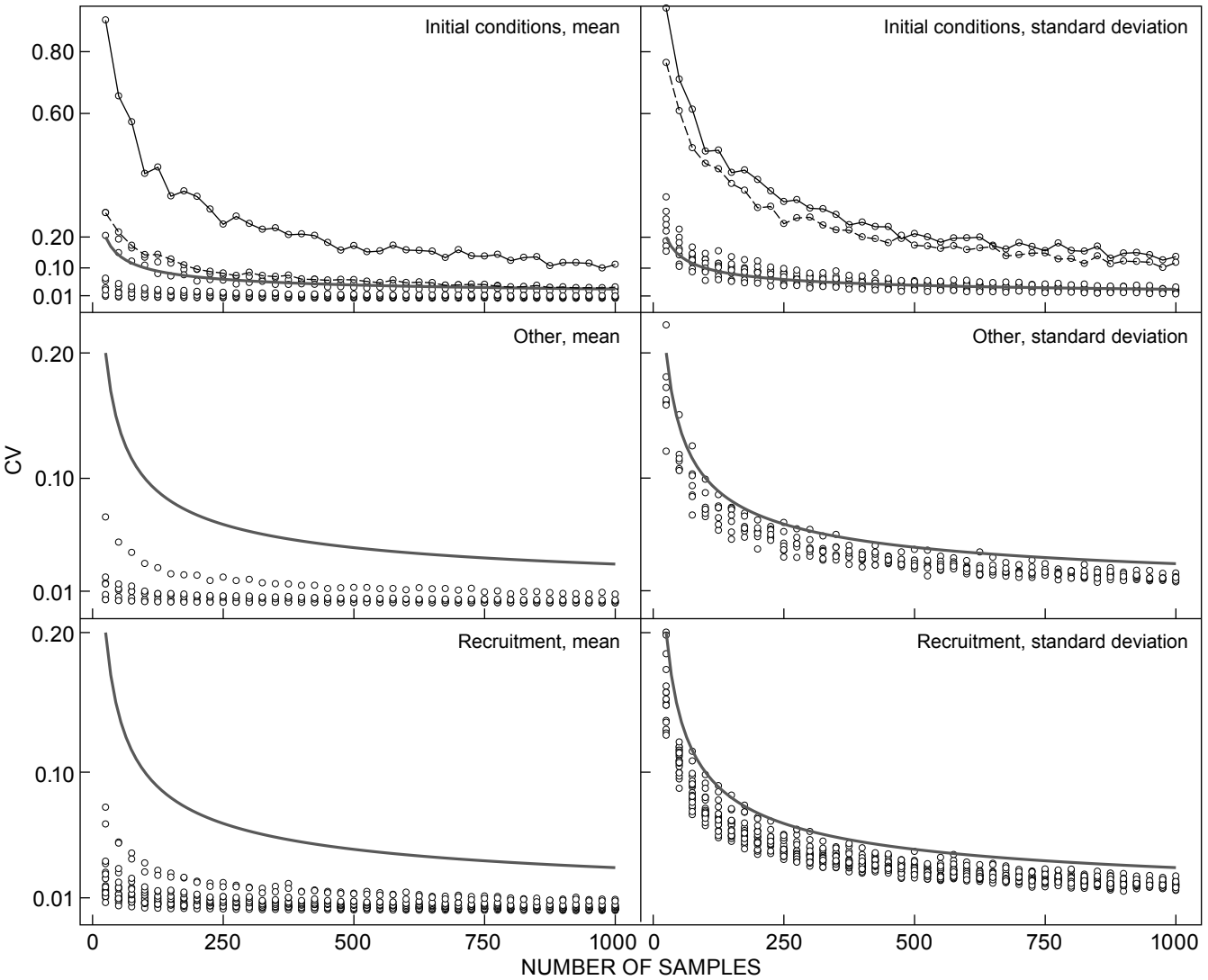

Figure 6: Results of a retrospective bootstrap sampling on the parameter estimates from the 1 000 bootstrap samples, with iterative weighting applied to all samples. This retrospective bootstrap studies the variation of the mean and standard deviation of each parameter estimate by calculating the coefficient of variation (CV) as a function of the number of bootstrap samples, n, of both the mean and standard deviation (SD). A point on the graph shows the CV of the mean (panels on the left hand side) or SD (panels on the right hand side) for a particular parameter and number of samples, n. The different panels contain the CVs of the initial number at age (v_a in Supplementary Appendix A, Eqn 6, available online), other variables i.e. the variables which are shown in Figure 3, and yearly recruitment shown in Figure 5 (R_y in Supplementary Appendix A, Eqn 5). CV of the initial number at ages 8 and 9 are illustrated with solid and broken lines respectively. For comparison, $1/\sqrt{n}$ is shown (thick grey line) on all panels

and Snowden 1985; Rao and Wu 1988; Sitter 1992). Theoretical assumptions and derivations behind these approaches do not easily extend to the present situation with disparate datasets, composite likelihoods in the estimation phase and, last but not least, the highly non-linear population dynamics models used as a basis for obtaining predicted values and error sums of squares or likelihood functions. The 'trick' in the current proposal is not a theoretical development but is the methodology of having the bootstrap sampling unit y_i as a collection of all relevant datasets sufficiently aggregated such that they can be assumed to be independent.

Some of the modifications of the original bootstrap have been developed for marine surveys (Smith 1997) but this has been intended to reflect e.g. the sampling design used for the surveys and simple estimation of quantities such as a stratified mean. In the present setting the data need to go through an aggregation procedure to be used in a non-linear population dynamics model and it is the output of this model which is of interest, not variances in the input. Thus, there is a need for the bootstrap to mimic this aggregation procedure for the full data from raw data or finer-scale aggregates. This is the case with any population dynamics or assessment model, used in fisheries or other areas of resource harvesting, particularly in a multispecies and multi-area context.

The methodology proposed here is certainly computationally intensive. However, this is also the case for many other

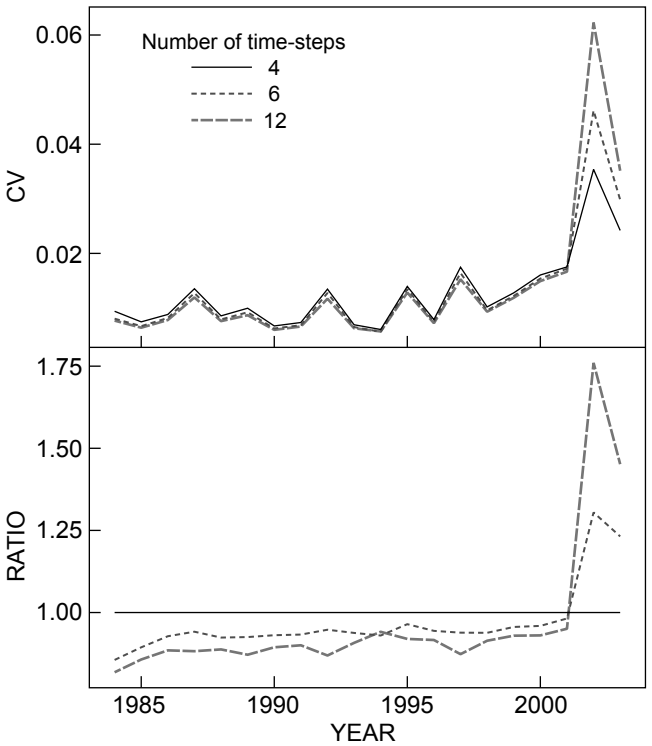

Figure 7: The CV of recruitment arising from the inverted Hessian by year as the number of intra-year time-steps is increased (upper panel). The bottom panel shows the ratio of the CV of the model with 3-month time-steps to the models with 2-month and 1-month steps

methods. For example, the MCMC evaluation of a Bayesian posterior involves a simulation of a correlated time-series whose stationary distribution is the posterior. This process is not trivially parallelisable over an arbitrary grid of computers (some of the difficulties are described in Wilkinson 2006). In comparison, the bootstrap approach described here is fairly trivially distributed onto a computer cluster.

To make the bootstrap proposed here more feasible, one could reduce the number of resampled datasets. Using 100 bootstrap replicates instead of 1 000 yields satisfactory results in terms of variance estimation, allowing a drastic reduction in the computing time needed. Conditioning on the weights from the original sample could further reduce the time needed but, judging by the results presented here, possible estimation biases may be harder to detect.

When compared to the bootstrap the Hessian-based approximation appears to underestimate the uncertainty by a factor of eight. This may seem contrary to previous results. Magnusson et al. (2013), using a simple catch-at-age simulation model, concluded that the MCMC method and the Hessian-based approach performed similarly. Recently, in Stewart et al. (2013), an MCMC and a Hessian-based approach performed similarly for real applications. The notable difference between the model described here and the aforementioned approaches is the objective function used here and the total number of data points (defined in Supplementary Appendix A, Section A.2.4; available online) used in the estimation process. The objective function consists of simple sums of squares that ignore potential correlations and tend to exaggerate the confidence level in the Hessian-based approach as the number of data points increases. This is illustrated in Figure 7 where it appears that the main factor in determining the size of the CV is the number of data points in the input files. Scale changes, such as aggregating data to larger length groups or increasing the size of the plus group by lowering the modelled maximum age, would in this case increase the size of the CV by simply reducing the number of data points. In contrast to the approach used here, a multinomial model, where the degrees of freedom are estimated, is often employed on catch-at-age (e.g. Trenkel et al. 2012) but length distributions, in the case of Icelandic cod, have serious distributional problems (Hrafnkelsson and Stefansson 2004). Future work on the model could potentially evaluate different distributional assumptions similar to those suggested above using the proposed bootstrap approach.

In this particular case study there were no discernible biases detected. Thus, the consequences of the Hessian-based approach appear to be restricted mostly to narrower confidence intervals. However, it is reasonable to assume that inconsistencies arising from conflicting data sources (e.g. in Schnute and Hilborn 1993; Stefansson 2003) would not be detected without analysing the effects of their relative weights. On the other hand, incorrect variance estimates may directly affect how annual catches are set. This occurs, for example, if a harvest control rule were to be based on a probabilistic measure such as that of a biomass not falling below a threshold, or a TAC not deviating too much from a target.

It is of considerable interest to compare the proposed bootstrap method to MCMC methods used in the Bayesian framework. This is, however, outside the scope of this study as it would require a considerable effort to adapt the Gadget framework to the Bayesian one. Future work could potentially focus on the evaluation of the two methodologies applied both on simulated datasets and for real applications in similar manner to Hannesson et al. (2009).

It is reassuring that the modelled years in which the greatest uncertainty is found are those where it is expected i.e. the initial year and then increasing towards the end of the modelled time period. The first year is the most data-poor, with no survey data or age–length compositions, and towards the end of the time period there are fewer cohorts with data available for most ages.

The method described here is designed to alleviate several known problems with other methods of uncertainty estimation. Several issues remain, however. For example, if a model is too 'stiff' through fixing parameters or other assumptions, then this may not be detected here except in special cases. These considerations could be explored by different models, e.g. split the commercial fleet component by gears, which can be implemented within the Gadget framework. On a related note there is also a balance to be found between estimation errors due to too-small size classes and distribution error caused by too-large size classes (Vandermeer 1978). It is therefore of interest to investigate the effects of the choice of scale such as size-class width but also time-step (Drouineau et al. 2009). The relative merits of these models can then be evaluated using an approach similar to the one proposed here. Similarly, different modelling

approaches, such as the different data weighting discussed in Francis (2011) or Hu and Zidek (2002), can be also be compared using the bootstrap technique presented here. Ultimately, each reweighting scheme is a different method for obtaining a point estimate and the bootstrap is a perfectly general method to obtain variance estimates.

When designing an aggregated database to be used for modelling, several issues need to be taken into account. The most important statistical condition on the choice of the 'data units' is that correlations between them should be minimal. On the other hand, there also needs to be a fair number of them within each model area if the bootstrap mechanism is to provide some variation in results. For a given measurement type one can, in many cases, investigate spatial correlation or variograms to determine the distances at which those become negligible (Petitgas 2001). This cannot easily be done for many data types, however (age–length tables, tagging experiments, etc.). In fact, the original reasoning for the areas used in this paper was ecological (Stefansson and Palsson 1997b; Taylor 2003) rather than based on spatial correlation, and it is likely that in most real situations data will be aggregated either according to such criteria or pragmatically into 'statistical rectangles' of some form.

Simple bootstrap resampling usually assumes that the elementary data units, $\{y_1,...,y_n\}$, behave like independently and identically distributed samples. Data in fisheries tend to be collected in a somewhat stratified manner, ranging from formal stratification to attempts to 'spread out' sampling, across gears, time and space. In the present setup this is simply ignored. This can be justified when the data are aggregated in a simple manner anyway (through sums or averages), since the bootstrap method then mimics the computation accordingly and/or when there is a large number of data units which can be viewed as representing a population of such units. In cases when one or a few of the subdivisions represent e.g. a spawning area, and the intended analysis is stratified accordingly, this approach can clearly not be used since then the bootstrap resampling does not reflect the computational method in use. When such issues arise, whether with respect to fishing gear, space or other units, an appropriate approach is to include these elements in the model. For example, the likelihood function can incorporate the various fishing gears, modelling each selectivity separately. The resampling then takes place separately for each gear.

Acknowledgements — Much of the work described here was undertaken while authors BÞE, LT, VK and GS were employed at the MRI (Marine Research Institute, Reykjavik) and uses data from the MRI databases. The Gadget code has been in development for more than a decade by many programmers at the MRI and IMR (Institute of Marine Research, Bergen). The work was supported in part by EU grants QLK5-CT1999-01609 and FP6 TP8.1 502482, as well as a grant from The Icelandic Centre for Research (Rannis). The authors would like to thank Dr S Gavaris for useful discussions, which have considerably improved the paper.

References

Babak O, Hrafnkelsson B, Palsson O. 2007. Estimation of the length distribution of marine populations in the Gaussian-multinomial setting using the method of moments. *Journal of Applied Statistics* 34: 985–996.

Begley J. 2004. Gadget user manual. Technical report 120. Marine Research Institute, Reykjavik.

Björnsson H, Sigurdsson T. 2003. Assessment of golden redfish (*Sebastes marinus* L.) in Icelandic waters. *Scientia Marina* 67(Suppl. 1): 301–314.

Bogstad B, Tjelmeland S, Tjelda T, Ulltang O. 1992. Description of a multispecies model for the Barents Sea (MULTSPEC) and a study of its sensitivity to assumptions on food preferences and stock sizes of minke whales and harp seals. Technical report SC/44/O 9. International Whaling Commission, Cambridge.

Chen Y, Breen P, Andrew N. 2000. Impacts of outliers and mis-specification of priors on Bayesian fisheries-stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences* 57: 2293–2305.

Davison A, Hinkley D. 1997. *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

De Croos M, Stefansson G. 2011. A case study of sampling strategies for estimating the length composition of commercial catches: The Sri Lankan shrimp trawl fishery. *Crustaceana* 84: 1581–1591.

Demyanov V, Wood SN, Kedwards TJ. 2006. Improving ecological impact assessment by statistical data synthesis using process-based models. *Journal of the Royal Statistical Society Series C* 55: 41–62.

Drouineau H, Mahévas S, Bertignac M, Fertin A. 2009. Assessing the impact of discretisation assumptions in a length-structured population growth model. *Fisheries Research* 91: 160–167.

Efron B. 1979. Bootstrap methods: another look at the jackknife. *The Annals of Statistics* 7: 1–26.

Efron B, Tibshirani RJ. 1994. *An introduction to the bootstrap*. London: Chapman & Hall/CRC.

Fournier DA, Skaug HJ, Ancheta J, Ianelli J, Magnusson A, Maunder MN, Nielsen A, Sibert J. 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* 27: 233–249.

Francis R. 2011. Data weighting in statistical fisheries stock assessment models. *Canadian Journal of Fisheries and Aquatic Sciences* 68: 1124–1138.

Gavaris S. 1988. An adaptive framework for the estimation of population size. *Canadian Atlantic Fisheries Scientific Advisory Committee Research Document* 88/29: 12.

Gavaris S, Ianelli JN. 2001. Statistical issues in fisheries stock assessment. *Scandinavian Journal of Statistics. Theory and Applications* 29: 245–272.

Gavaris S, Patterson KR, Darby CD, Lewy P, Mesnil B, Punt AE, Cook RM, Kell LT, O'Brien CM, Restrepo VR, Skagen DW, Stefansson G. 2000. Comparison of uncertainty estimates in the short term using real data. ICES Document CM 2000/V:03. International Council for the Exploration of the Sea, Copenhagen.

Gross S. 1980. Median estimation in sample surveys. In: *Proceedings of the Survey Research Methods Section.* Alexandria, Virginia: American Statistical Association. pp 181–184.

Gudmundsdóttir Á, Steinarsson BÆ, Stefánsson G. 1988. A simulation procedure to evaluate the efficiency of some otolith and length sampling schemes. ICES Document CM 1988/D:14, 500:14. International Council for the Exploration of the Sea, Copenhagen.

Hannesson S, Jakobsdottir A, Begley J, Taylor L, Stefansson G. 2009. On the use of tagging data in statistical multispecies multi-area models of marine populations. *ICES Journal of Marine Science* 65: 1762–1772.

Helle K, Pennington M. 2004. Survey design considerations for estimating the length composition of the commercial catch of some deep-water species in the northeast Atlantic. *Fisheries Research* 70: 55–60.

Hrafnkelsson B, Stefansson G. 2004. A model for categorical length data from groundfish surveys. *Canadian Journal of Fisheries and Aquatic Sciences* 61: 1135–1142.

Hu F, Zidek JV. 2002. The weighted likelihood. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 30: 347–371.

ICES (International Council for the Exploration of the Sea). 2011. Report of the North Western Working Group (NWWG), 26 April–3 May 2011. ICES CM 2011/ACOM:7. ICES headquarters, Copenhagen.

Jennrich RI. 1969. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* 40: 633–643.

Kupca V. 2004. A standardized database for fisheries data. ICES Document CM 2004/FF:15. International Council for the Exploration of the Sea, Copenhagen.

Kupca V, Sandbeck P. 2003. dst^2 Datawarehouse structure and data import. In: dst^2: Development of structurally detailed statistically testable models of marine populations. Technical Report 98. Marine Research Institute, Reykjavik. pp 37–46.

Lindstrøm U, Smout S, Howell D, Bogstad B. 2009. Modelling multi-species interactions in the Barents Sea ecosystem with special emphasis on minke whales and their interactions with cod, herring and capelin. *Deep Sea Research Part II* 56: 2068–2079.

Magnusson A, Punt AE, Hilborn R. 2013. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. *Fish and Fisheries* 14: 325–342.

Maunder MN, Punt AE. 2013. A review of integrated analysis in fisheries stock assessment. *Fisheries Research* 142: 61–74.

McCarthy PJ, Snowden CB. 1985. The bootstrap and finite population sampling. *Vital and Health Statistics,* Series 2, No. 95: DHHS Pub. No. (PHS) 85–1 369. Washington, DC: Public Health Service.

Methot RD. 1989. Synthetic estimates of historical abundance and mortality for northern anchovy. *American Fisheries Society Symposium* 6: 66–82.

Millar RB. 2002. Reference priors for Bayesian fisheries models. *Canadian Journal of Fisheries and Aquatic Sciences* 59: 1492–1502.

Myers RA, Cadigan NG. 1995. Statistical analysis of catch-at-age data with correlated errors. *Canadian Journal of Fisheries and Aquatic Sciences* 52: 1265–1273.

Oehlert G. 1992. A note on the delta method. *The American Statistician* 46: 27–29.

Palsson OK, Schopka SA, Stefansson G, Steinarsson BA. 1989. Icelandic groundfish survey data used to improve precision in stock assessments. *Journal of Northwest Atlantic Fishery Science* 9: 53–72.

Patterson K, Cook R, Darby C, Gavaris S, Kell L, Lewy P, Mesnil B, Punt A, Restrepo V, Skagen DW, Stefansson G. 2001. Estimating uncertainty in fish stock assessment and forecasting. *Fish and Fisheries* 2: 125–157.

Pennington M, Volstad JH. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. *Biometrics* 50: 1–8.

Petitgas P. 2001. Geostatistics in fisheries survey design and stock assessment: models, variances and applications. *Fish and Fisheries* 2: 231–249.

Punt AE, Hilborn R. 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. *Reviews in Fish Biology and Fisheries* 7: 35–63.

Rao JNK, Wu CFJ. 1988. Resampling inference with complex survey data. *Journal of the American Statistical Association* 83: 231–241.

Richards L. 1991. Use of contradictory data sources in stock assessments. *Fisheries Research* 11: 225–238.

Schnute JT, Hilborn R. 1993. Analysis of contradictory data sources in fish stock assessment. *Canadian Journal of Fisheries and Aquatic Sciences* 50: 1916–1923.

Sigurdsson T, Hjorleifsson E, Bjornsson H, Palsson OK. 1997. Fall groundfish survey in Icelandic waters (stofnmaeling botnfiska a Islandsmidum haustid 1996). Technical Report 61. Marine Research Institute, Reykjavik (in Icelandic).

Sitter RR. 1992. A resampling procedure for complex survey data. *Journal of the American Statistical Association* 87: 755–765.

Smith SJ. 1997. Bootstrap confidence limits for groundfish trawl survey estimates of mean abundance. *Canadian Journal of Fisheries and Aquatic Sciences* 54: 616–630.

Spiegelhalter D, Thomas A, Best N, Gilks W. 1996. Bugs 0.5: Bayesian inference using Gibbs sampling manual (version ii). Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Stefansson G. 1998. Comparing different information sources in a multispecies context. In: Funk F, II TQ, Heifetz J, Ianelli J, Powers J, Schweigert J, Sullivan P, Zhang C (eds), *Fishery stock assessment models: proceedings of 15th Lowell Wakefield Fisheries Symposium; Anchorage 1997*. pp 741–758.

Stefansson G. 2003. Issues in multispecies models. *Natural Resource Modeling* 16: 415–438.

Stefansson G, Palsson OK. 1997a. BORMICON. A boreal migration and consumption model. Technical Report 58. Marine Research Institute, Reykjavik.

Stefansson G, Palsson OK. 1997b. Statistical evaluation and modelling of the stomach contents of Icelandic cod (*Gadus morhua*). *Canadian Journal of Fisheries and Aquatic Sciences* 54: 169–181.

Stefansson G, Palsson OK. 1998. A framework for multispecies modelling of boreal systems. *Reviews in Fish Biology and Fisheries* 8: 101–104.

Stewart IJ, Hicks A, Taylor I, Thorson JT, Wetzel C, Kupschus S. 2013. A comparison of stock assessment uncertainty estimates using maximum likelihood and Bayesian methods implemented with the same model framework. *Fisheries Research* 142: 37–46.

Taylor L. 1999. Stock dynamics and long term assessment of haddock (*Melanogrammus aeglefinus*) in Icelandic waters. Paper presented at ICES North Western Working Group, 26 April 1999. Available at http://www.hi.is/~gunnar/papers/ltdyn.pdf [accessed 12 November 2013].

Taylor L. 2003. Definition of areas in Icelandic waters. In: dst^2: Development of structurally detailed statistically testable models of marine populations. Technical Report 98. Marine Research Institute, Reykjavik. pp 222–230.

Taylor L, Begley J, Kupca V, Stefansson G. 2007. A simple implementation of the statistical modelling framework Gadget for cod in Icelandic waters. *African Journal of Marine Science* 29: 223–245.

Tinker MT, Doak DF, Estes JA, Hatfield BB, Staedler MM, Bodkin JL. 2006. Incorporating diverse data and realistic complexity into demographic estimation procedures for sea otters. *Ecological Applications* 16: 2293–2312.

Tjelmeland S, Bogstad B. 1989. MULTSPEC: the manual. Bergen: Institute of Marine Research.

Trenkel VM, Bravington MV, Lorance P. 2012. A random effects population dynamics model based on proportions-at-age and removal data for estimating total mortality. *Canadian Journal of Fisheries and Aquatic Sciences* 69: 1881–1893.

Vandermeer J. 1978. Choosing category size in a stage projection matrix. *Oecologia* 32: 79–84.

Wilkinson DJ. 2006. Parallel Bayesian computation. *Statistics Textbooks and Monographs* 184: 477.

Received May 2013, revised November 2013, accepted November 2013