Estimating catch-at-age by combining data from different sources

David Hirst, Geir Storvik, Magne Aldrin, Sondre Aanes, and Ragnar Bang Huseby

Abstract: Estimating the catch-at-age of commercial fish species is an important part of the quota-setting process for many different species and almost all countries with a fishing fleet. Current procedures are usually very time-consuming and somewhat ad hoc, and the estimates have no measure of uncertainty. We previously developed a method for catch-at-age of Norwegian Atlantic cod (*Gadus morhua*), but this only considered aged fish sampled randomly from random hauls. In most countries, the sampling scheme is not so simple. There are usually a very large number of length-only samples from which the age must be estimated using an age–length relationship, and often some or all of the age samples are collected from data that are first stratified by length. This adds considerably to the difficulties in the estimation. In this paper, we model the three different kinds of data simultaneously using a development of our earlier Bayesian hierarchical model. This enables us to obtain estimates of the catch-at-age with appropriate uncertainty and also to provide advice on how best to sample data in the future. The data types are random samples of age, length, and weight; age and weight stratified by length; and length only.

Résumé : L'estimation de la récolte en fonction de l'âge est une étape importante du processus de définition des quotas pour plusieurs espèces et dans presque tous les pays qui possèdent une flotte de pêche. Les méthodes courantes exigent beaucoup de temps et elles sont ajustées à des situations particulières et elles ne comportent pas de mesure d'incertitude. Nous avons développé antérieurement une méthode pour estimer la capture en fonction de l'âge chez la morue franche (*Gadus morhua*) de Norvège, mais elle ne tient compte que des poissons d'âge connu échantillonnés au hasard dans des récoltes aléatoires. Dans la plupart des pays, le plan d'échantillonnage est loin d'être aussi simple. Il y a généralement un très grand nombre d'échantillons comportant seulement des longueurs, dont on doit estimer l'âge à l'aide d'une relation âge–longueur, et souvent quelques-uns ou même tous les échantillons contenant des déterminations d'âge ont été tirés de données préalablement stratifiées d'après la longueur. Une modification de notre modèle hiérarchique bayésien antérieur nous permet de traiter les trois types de données simultanément. Nous obtenons ainsi des estimations de la capture en fonction de l'âge assorties d'une mesure d'incertitude appropriée; nous proposons aussi comment mieux échantillonner les données dans le futur. Les types de données utilisées sont des échantillons aléatoires des âges, des longueurs et des masses; des âges et des masses stratifiées en fonction de la longueur; et des longueurs seules.

[Traduit par la Rédaction]

Introduction

As part of the process of setting fishing quotas, every country in Europe with a fishing fleet reports the total annual catch-at-age of various species to the International Council for the Exploration of the Seas. Strictly speaking, catch-at-age means the total number of fish caught at each age. However, it is common to group less frequent ages together to form a number of age groups. In our case, fish older than 12 years are considered one group. Also, an unknown number of fish are caught but discarded at sea. We do not take into account these discards. This kind of data is sometimes known as market sampling, although in Norway, a substantial part of the data is taken directly from the boat rather than from the market.

The weight of the total catch is usually considered to be known at a fairly fine resolution (in Norway, season by gear by area by year) and the aim of the analysis is (*i*) to estimate proportion-at-age and (*ii*) to estimate the mean weight of fish to convert the total to numbers from weight. A variety of different sampling schemes have been established for this purpose, and the data are analysed in a range of different ways. A common feature of most of these methods is that there is no statistical model for the sampling process. Ad

Received 5 January 2004. Accepted 4 December 2004. Published on the NRC Research Press Web site at http://cjfas.nrc.ca on 25 June 2005.
J17897

D. Hirst,[1] **M. Aldrin, and R.B. Huseby.** Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway.
G. Storvik. Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway, and University of Oslo, P.O. Box 1053 Blindern, N-0316 Oslo, Norway.
S. Aanes. Institute for Marine Research, P.O. Box 1870 Nordnes, 5817 Bergen, Norway.

[1]Corresponding author (e-mail: david.hirst@nr.no).

 doi: 10.1139/F05-026

hoc methods are used, which are very time-consuming and rely on individual judgement, which by definition is not repeatable. The Norwegian approach is outlined in Hirst et al. (2004). With such methods, it is very difficult to get a measure of the uncertainty in the reported results. To address this problem, Hirst et al. (2004) developed a Bayesian hierarchical model for the Norwegian catch of northeast Atlantic cod (*Gadus morhua*). This model, however, only addressed the strategy of sampling fish at random from random boats and estimating the age and measuring the weight of all of the fish in the sample. There was no modelling of length in this paper, and weight was modelled directly in terms of age.

The Norwegian sampling scheme is probably unique in Europe. Elsewhere, there is an emphasis on sampling large numbers of fish for which only length is measured and weighing and estimating the ages of only a few of these. These aged fish are usually stratified by length (e.g., one fish from each 5-cm length-class in a sample might be aged). This kind of sampling in fact takes place to a lesser degree in Norway as well, and additional length-only or age-given-length data (often from independent sources such as the Coast Guard) are utilized in the estimation. In this paper, we develop a common model for all of these kinds of data.

The difficulties in the analysis arise mostly because it is not possible to develop a proper sampling scheme for fishing vessels. In general, they are sampled when and if they are available. There are important differences in the catch between different seasons, fishing gears, and regions of the sea, and if we call each combination of these factors a cell, there are necessarily many cells with no samples. In addition, there is a large within-haul correlation in the ages and sizes of the caught fish. Thus, the effective sample size is very much smaller than the number of fish sampled (see Aanes and Pennington 2003). This leads to a larger uncertainty than would be apparent from a naïve assumption of independence of fish.

The aim of this paper is to establish a proper statistical framework within which market sampling data can be analysed. The hierarchical framework is very appropriate for this kind of modelling because it can easily accommodate the different sampling schemes and because it provides a full measure of uncertainty. Random effects are included into the model to take into account correlation between samples from the same haul.

Bayesian approaches are usually used for making inference in hierarchical models (Gelman et al. 1995) and are now slowly emerging into the fisheries community (e.g., see Millar and Meyer 2000 and references therein). They have also been used in stock assessments (Hilborn and Lierman 1998; Lewy and Nielsen 2003), estimation of depensation (Lierman and Hilborn 1997), and estimation of biological reference points (Prevost et al. 2003).

Materials and methods

Data

There are three main sources of data available to the Norwegian Institute of Marine Research (IMR), which is responsible for estimating the catch-at-age of cod in Norway.

(*i*) The *Amigo* is a research vessel hired by IMR that sails from port to port along the north Norwegian coast over a period of about 6 weeks four times a year (roughly corresponding to the four seasons). At each port, it takes a sample of about 80 fish from any boats available at the time. The fish sampled from a boat are intended to be a sample from a single haul, and this is usually achieved. In some cases, there may have been some mixing of hauls before the samples are taken, however. There is rarely more than one boat available. The fish are weighed, the length measured, and the otoliths extracted for estimating the age of the fish (Campana 2001). Each year, about 200 boats, and thus about 16 000 fish, are sampled. Note that the program only samples landings. There are an unknown number of small fish discarded at sea, although we refer to catch-at-age in this paper.

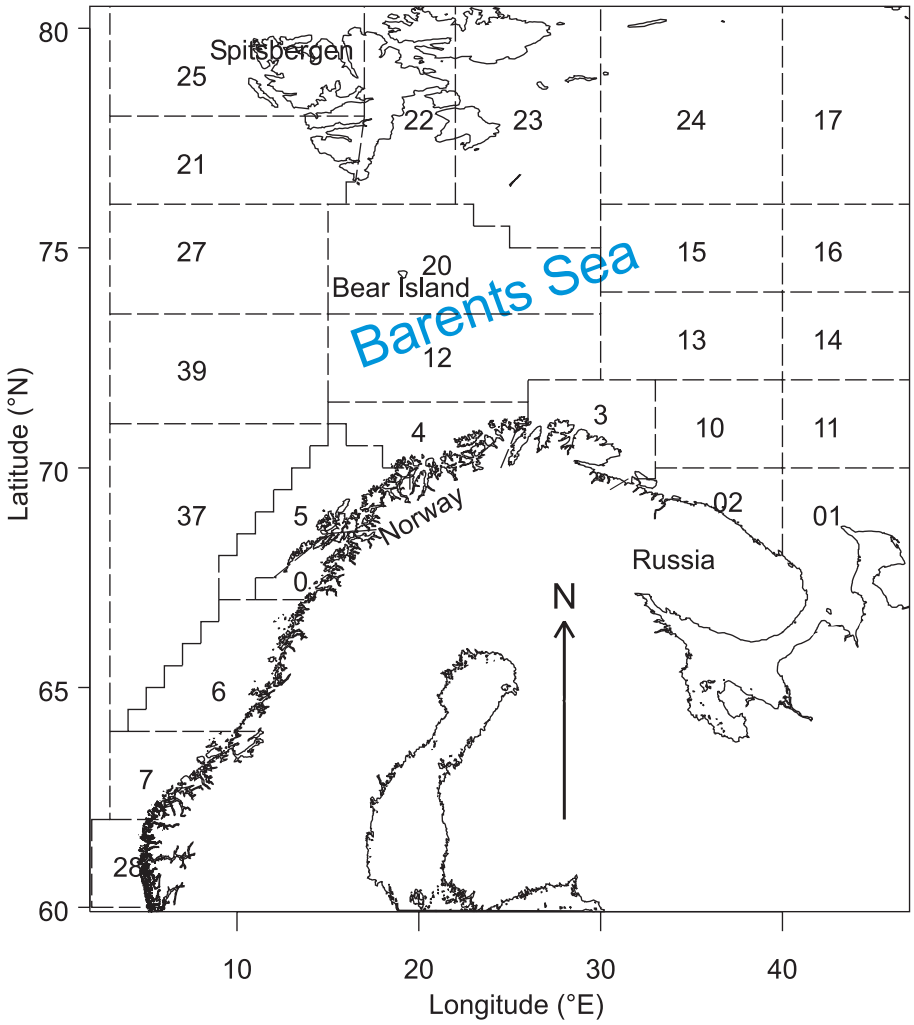
(*ii*) The Coast Guard, whose tasks include making sure that the Norwegian fishery laws and regulations are kept, have the right to inspect any vessel and to sample the catch. In most cases, the vessels sampled by the Coast Guard are a random sample of the vessels operating within an area, but in a few cases, the inspections may be based on suspicion of illegal fishing. Thus, it might be expected that some of the samples would be biased or unrepresentative for the total catch, although this does not appear to be the case. In general, these samples will only provide length measurements of the fish sampled, although occasionally, there are some ages and weights as well. The Coast Guard samples more of the trawlers than the *Amigo*. The number of fish sampled in each haul is very variable but averages about 100.

(*iii*) The reference fleet is a fleet of commercial fishing vessels that have agreed to provide IMR with data on their catch. The reference fleet was started in 2001 with six vessels and consists currently of eight vessels. The fleet targets several commercially important species including cod. This sampling program is developing and will expand in the years ahead. So far, it has consisted mostly of length-only data, but there are an increasing number of age samples. In 2002, this fleet sampled approximately 500 hauls of cod with around 90 fish sampled in each haul.

The "cells" that we consider in this paper are the individual combinations of the regions in Fig. 1, season (corresponding roughly to the quarters of the year), gear (bottom trawl, Danish seine, gill net, longline, and handline), and year (1995–2002, but the reference fleet only began in 2001). One cell therefore represents one gear in one region in one season of one year. Our sampling unit is the haul. The Coast Guard and reference fleet always provide data from individual hauls, and although we suspect that the *Amigo* data sometimes contain fish from mixed hauls (although always from the same cell), we do not believe that this is an important effect. We do not consider the actual boat that was sampled to be of interest. For the *Amigo* and Coast Guard data, it is very unlikely that the same boat would be sampled twice (at least in the same year), but clearly, the reference fleet provides many samples from the same few boats. Any boat effect, however, is largely due to the particular gear being used, and the remaining effect will be very small compared with the differences between hauls.

For the purposes of the analyses in this paper, we have formed super-regions by grouping the regions in the map. In

Fig. 1. Map of sampling area showing the Norwegian Directorate of Fisheries statistical regions in which the total catch is reported.



fact, we have used the eight standard IMR groups of (3, 2, 10, 11, 13, 14, 15, 16, 17, 24, 1), (12), (4), (5, 37, 39), (0), (6), (7, 28), and (20, 21, 22, 23, 25, 27). It is necessary to do some grouping because most regions have little or no data, although other groupings are possible. We have grouped ages over 12 together, and there are no fish younger than 2, giving us 12 age groups.

In the next sections, we develop the various components of the model: the proportion-at-age, length-given-age, and weight-given-length. The components are brought together in the likelihood for the whole data set. We then explain how to obtain samples from the posterior distribution of the parameters given the data using Markov Chain Monte Carlo (MCMC) (Gilks et al. 1996). Finally, we show some results and illustrate how these change when different data sources are included in the analysis. This also enables us to provide some guidance on how best to sample in the future.

Model for proportion-at-age

The samples from a boat are assumed to be randomly drawn from the total population of fish in that haul, and the hauls are themselves assumed to be randomly sampled from all of those within the appropriate cell. The numbers-at-age in a sample from haul h from cell c, $\mathbf{X}_{c,h}$, are therefore multinomial:

$$\mathbf{X}_{c,h} \sim \text{multinomial}(\boldsymbol{p}_{c,h}, n_{c,h})$$

The number of fish sampled from the haul, $n_{c,h}$, is assumed not to depend in any way on $\boldsymbol{p}_{c,h}$.

The vector of proportions-at-age in the haul, $\boldsymbol{p}_{c,h}$, has A elements, one for each age group. Let $p_{c,h}(a)$ be the ath element, where $0 \le p_{c,h}(a) \le 1$ and $\sum_{a'=1}^{A} p_{c,h}(a') = 1$. This is reparameterized as

$$p_{c,h}(a) = \frac{\exp(\alpha_{c,h}^{a})}{\sum_{a'=1}^{A} \exp(\alpha_{c,h}^{a'})}$$

We model $\alpha_{c,h}^{a}$ in terms of the various covariates as

$$\alpha_{c,h}^{a} = \alpha^{\text{base},a} + \alpha_{v(c)}^{\text{year},a} + \alpha_{s(c)}^{\text{season},a} + \alpha_{g(c)}^{\text{gear},a} + \zeta_{r(c)}^{\text{region},a} + \zeta_{c}^{\text{cell},a} + \zeta_{c,h}^{\text{haul},a}$$

Here, $y(c)$ means the year, $s(c)$ the season, $g(c)$ the gear, and $r(c)$ the region corresponding to cell c. From now on for clarity, we drop the c and just refer to $\alpha_y^{\text{year},a}$, etc.

The α terms and $\varsigma_r^{\text{region},a}$ are the main effects for year, season, gear, and region. The α terms are fixed effects and $\varsigma_r^{\text{region},a}$ is a spatially smoothed random effect. It is necessary to estimate the proportions in areas with no data, and our approach is to introduce some spatial smoothing. This is accomplished by assuming that $\varsigma_r^{\text{region},a}$ follows a Gaussian conditional autoregressive distribution (e.g., Carlin and Louis 1996). The alternative would be to group areas such that there were none with no data. This is unsatisfactory for several reasons, particularly because the grouping would have to be done differently in each analysis. It is assumed that there will always be some data for all levels of the fixed effects that are of interest. The $\varsigma_c^{\text{cell},a}$ terms are independent random effects modelling the interactions between the main effects (e.g., see Gelman et al. 1995). In other words, the differences between the fit from the main-effects-only model and the true cell means are modelled by the $\varsigma_c^{\text{cell},a}$ terms. The differences between hauls within a cell are modelled by the random effects $\varsigma_{c,h}^{\text{haul},a}$. These must be random (rather than fixed) effects because there are many cells and hauls with no data. We assume that all of the interactions (i.e., the $\varsigma_c^{\text{cell},a}$ terms) can be modelled by a single distribution. It would be plausible to assume that some interactions (e.g., between season and year) had a higher variance than others, but we have found no evidence for this in the data (for more details of the parameters, including identifiability constraints and the prior distributions, see Appendix A).

Models for length-given-age and weight-given-length

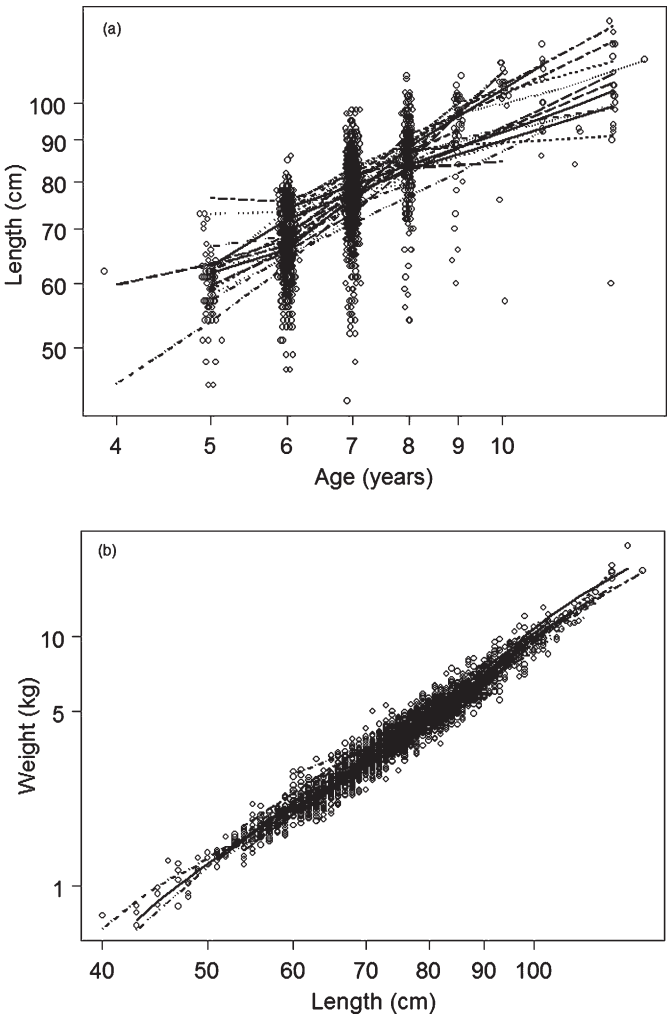
We plot log(length) against log(age) for all of the hauls in one cell (Fig. 2*a*). The lines are nonparametric fits for the individual hauls using the Splus function "supsmu". The equivalent plot of log(weight) against log(length) is also provided (Fig. 2*b*). It can be seen that there is a reasonably linear relationship between the variables, with some differences between hauls. We model both relationships as linear with a constant slope and variable intercept. This is unproblematic for the length–weight relationship, but the age–length model could be improved. There is some evidence of a nonlinear relationship, which may cause problems at high and low ages. Other models would be possible and may in particular be necessary for other species with different growth patterns. The von Bertalanffy (1938) growth curve is commonly used in fisheries science and could be used as an alternative. See Haddon (2001) for a discussion of possible growth curves.

Note that it would be possible to model weight given age directly but that modelling it via length enables us to get a better estimate of the mean weight-at-age in cells with length but no age data. Since the length–weight relationship is so strong, there would in any case be no advantage in modelling weight directly in terms of age.

We assume that length-given-age and weight-given-length are log-normal, with constant variances, and means linear in log(age) and log(length), respectively, in an individual haul. The slopes are constant, but the intercepts vary between cells and boats within a cell:

$$\log(\text{length}_{c,h,f}) = \beta_{0,c,h} + \beta_1 \log(\text{age}_{c,h,f}) + \varepsilon_{c,h,f}^{\text{fish}}$$

Fig. 2. Plots showing the relationship between (*a*) length and age and (*b*) weight and length for all hauls within a cell. The lines are nonparametric fits for individual hauls. Ages have been jittered for clarity.



$$\log(\text{weight}_{c,h,f}) = \delta_{0,c,h} + \delta_1 \log(\text{length}_{c,h,f}) + \nu_{c,h,f}^{\text{fish}}$$

Here, $\text{length}_{c,h,f}$ is the length of the fth fish from haul h in cell c, $\text{weight}_{c,h,f}$ its weight, and $\text{age}_{c,h,f}$ its age; $\varepsilon_{c,h,f}^{\text{fish}}$ and $\nu_{c,h,f}^{\text{fish}}$ are independent zero mean Gaussian random variables.

The slopes β_1 and δ_1 are common to all cells and hauls, and the intercepts $\beta_{0,c,h}$ and $\delta_{0,c,h}$ are given by

$$\beta_{0,c,h} = \beta^{\text{base}} + \beta_y^{\text{year}} + \beta_s^{\text{season}} + \beta_g^{\text{gear}} + \varepsilon_r^{\text{region}} + \varepsilon_c^{\text{cell}} + \varepsilon_{c,h}^{\text{haul}}$$

$$\delta_{0,c,h} = \delta^{\text{base}} + \delta_y^{\text{year}} + \delta_s^{\text{season}} + \delta_g^{\text{gear}} + \nu_r^{\text{region}} + \nu_c^{\text{cell}} + \nu_{c,h}^{\text{haul}}$$

$\varepsilon_r^{\text{region}}$ and ν_r^{region} are conditional autoregressive distribution parameters with properties to those of $\varsigma_r^{\text{region},a}$ in the age model (see Appendix A). $\varepsilon_c^{\text{cell}}$ and ν_c^{cell} and are random "all interactions" effects equivalent to $\varsigma_c^{\text{cell},a}$. $\varepsilon_c^{\text{haul}}$ and ν_c^{haul} are between-haul random terms. The β and δ terms are fixed effects similar to the α terms in the model for proportions-at-age (for more details, see Appendix A).

Inference on unknown parameters

Parts of the model are standard, and ordinary methods such as maximum likelihood could have been applied. This is certainly true for the length-given-age model and the weight-given-length model. With all ages known and with no random effects involved in the age model also, the parameters in the multinomial model describing the age proportions could easily be found by maximum likelihood. Both the inclusion of random effects in the multinomial model and missing ages make maximum likelihood estimation much more complicated. Further, a frequentist approach to estimation makes it difficult to take the uncertainty in the parameters into account. It is probably possible to use a frequentist model, perhaps by using the Expectation Maximization (EM) algorithm to maximize the likelihood in the presence of missing data, combined with a parametric bootstrap to obtain the uncertainty. This does not seem to us to be the best approach to the problem, however.

Our approach has been the Bayesian one. The full information about the parameters (including random effects) is described through the posterior distribution. This distribution is difficult to calculate, but approximations can be obtained through Monte Carlo sampling. The actual sampling is performed through an MCMC algorithm using a combination of Gibbs sampling and Metropolis–Hastings steps. The details are given in Appendix B, but in outline, the approach is as follows. (*i*) If the ages and lengths of all of the sampled fish were known, it would be simple to simulate the parameters of the length-given-age model (since this is just a linear model). It would also be relatively simple to simulate the parameters of the proportion-at-age model (although the inclusion of random effects complicates the simulations somewhat). Also, parameters from the different submodels (age model, length-given-age model, and weight-given-length model) are independent in this case. (*ii*) If the parameters of the length-given-age model and the proportion-at-age model are known, it is simple to simulate the ages of the fish with only length data (since age-given-length is multinomial). (*iii*) We therefore treat the missing ages as parameters and use Gibbs sampling to alternate between simulating the missing ages and simulating the other model parameters. It is also possible to use block updating for most of the parameters apart from the precisions.

Using this approach, it is possible to find the joint posterior distribution of all of the parameters very quickly. On a reasonably powerful computer, 1 year's data can easily be analysed in less than 5 min. Obviously, the time increases with the number of years of data, but even 8 years worth takes under an hour. Convergence of the MCMC chains is fast because of the block-updating. Research is currently underway to make this even more efficient.

Estimating catch-at-age

We need to estimate the total catch-at-age a in cell c, T_{ca}. To take the uncertainty of the parameters into account, parameters (including random effects for cell and for those hauls that are observed) are samples from the joint posterior distribution (as described in the previous section). Given the parameters in the model, it can be shown that this is a simple function of those parameters (see below). Denoting one set of such parameters by θ, the total catch-at-age given the parameters can be written as a function $T_{ca}(\theta)$ of θ. A Monte Carlo estimate of the catch-at-age is then given by the mean of the $T_{ca}(\theta)$s. Uncertainty measures can be calculated simultaneously, i.e., standard errors are estimated by empirical standard errors of the $T_{ca}(\theta)$s.

Consider now the calculation of $T_{ca}(\theta)$ for a given set of parameters θ. We have $T_{ca} = T_c \times \text{mean}_c(p(a))$, where T_c is the total catch in the cell in numbers of fish and $\text{mean}_c(p(a))$ is the mean proportion-at-age over all hauls in the cell (i.e., the mean over all hauls taken by all boats fishing in that cell, not just those observed). We assume that there are a large number of these hauls so that the mean is equal to the expected value, giving us $T_{ca} = T_c E_c(p(a))$.

The total catch in a cell, W_c, is given in weight rather than numbers. We therefore need the mean weight of fish caught in the cell, $\overline{w}_c$, to calculate $T_c = W_c / \overline{w}_c$. We have

$$\log(\text{weight}_{c,h,f}) = \delta_{0,c,h} + \delta_1 \log(\text{length}_{c,h,f}) + \nu^{\text{fish}}_{c,h,f}$$

$$\log(\text{length}_{c,h,f}) = \beta_{0,c,h} + \beta_1 \log(\text{age}_{c,h,f}) + \varepsilon^{\text{fish}}_{c,h,f}$$

Thus:

$$\begin{aligned}\log(\text{weight}_{c,h,f}) &= \delta_{0,c,h} + \delta_1(\beta_{0,c,h} + \beta_1 \log(a_{c,h,f}) \\ &\quad + \varepsilon^{\text{fish}}_{c,h,f}) + \nu^{\text{fish}}_{c,h,f} \\ &= \delta_{0,c,h} + \delta_1\beta_{0,c,h} + \delta_1\beta_1 \log(a_{c,h,f}) \\ &\quad + (\delta_1 \varepsilon^{\text{fish}}_{c,h,f} + \nu^{\text{fish}}_{c,h,f}) \\ &= A_c + A_h + \beta_c \log(a_{c,h,f}) + C_f\end{aligned}$$

Here, A_c and B_c are constant for all hauls in a cell and given from the simulations, A_h is an unknown random haul dependent intercept, and C_f is an unknown random fish effect. From the earlier equations:

$$\begin{aligned}A_c &= (\delta^{\text{base}} + \delta^{\text{year}}_y + \delta^{\text{season}}_s + \delta^{\text{gear}}_g + \nu^{\text{region}}_r + \nu^{\text{cell}}_c) \\ &\quad + \delta_1(\beta^{\text{base}} + \beta^{\text{year}}_y + \beta^{\text{season}}_s + \beta^{\text{gear}}_g + \varepsilon^{\text{region}}_r + \varepsilon^{\text{cell}}_c)\end{aligned}$$

$$B_c = \delta_1\beta_1$$

A_h is a random haul dependent intercept:

$$A_h = \nu^{\text{haul}}_{c,h} + \delta_1 \varepsilon^{\text{haul}}_{c,h}$$

This is constant in a haul but a Gaussian random variable within a cell:

$$A_h \sim N\left(0, \frac{1}{\tau^{\text{haul}}_{\text{weight}}} + \frac{\delta_1^2}{\tau^{\text{haul}}_{\text{length}}}\right)$$

C_f is a random fish effect:

$$C_f = \delta_1 \varepsilon^{\text{fish}}_{c,h,f} + \nu^{\text{fish}}_{c,h,f}$$

This is a random fish effect, with a constant distribution:

$$C_f \sim N\left(0, \frac{1}{\tau^{\text{fish}}_{\text{weight}}} + \frac{\delta_1^2}{\tau^{\text{fish}}_{\text{length}}}\right)$$

Hence, the weight of a random fish f of age $a_{c,h,f}$ in haul h, cell c, is lognormal:

$$\log(\text{weight}_{c,h,f} \mid a_{c,h,f}) \sim N\left(A_c + A_h + B_{\text{cell}} \log(a_{c,h,f}), \frac{1}{\tau^{\text{fish}}_{\text{weight}}} + \frac{\delta_l^2}{\tau^{\text{fish}}_{\text{length}}}\right)$$

Its expectation (over all fish in the haul) is

$$E_{c,h}(\text{weight}_{c,h,f} \mid a_{c,h,f}) = \exp[A_c + B_c \log(a_{c,h,f})] \times \exp(A_h) \exp\left[\tfrac{1}{2}\left(\frac{1}{\tau^{\text{fish}}_{\text{weight}}} + \frac{\delta_l^2}{\tau^{\text{fish}}_{\text{length}}}\right)\right]$$

Taken over all hauls in a cell, this expectation is itself a random variable, also lognormal, since $\exp(A_h)$ is lognormal:

$$\log(E_{c,h}(\text{weight}_{c,h,f} \mid a_{c,h,f})) \sim N\left(A_c + B_c \log(a_{c,h,f}) + \tfrac{1}{2}\left(\frac{1}{\tau^{\text{fish}}_{\text{weight}}} + \frac{\delta_l^2}{\tau^{\text{fish}}_{\text{length}}}\right), \frac{1}{\tau^{\text{haul}}_{\text{weight}}} + \frac{\delta_l^2}{\tau^{\text{haul}}_{\text{length}}}\right)$$

The expected weight of a fish of age a in a cell is therefore

$$E_c(\text{weight} \mid a) = \exp\left[A_c + B_c \log(a) + \tfrac{1}{2}\left(\frac{1}{\tau^{\text{fish}}_{\text{weight}}} + \frac{\delta_l^2}{\tau^{\text{fish}}_{\text{length}}}\right) + \tfrac{1}{2}\left(\frac{1}{\tau^{\text{haul}}_{\text{weight}}} + \frac{\delta_l^2}{\tau^{\text{haul}}_{\text{length}}}\right)\right]$$

Again assuming a large number of hauls in a cell so that the mean weight-at-age is equal to its expected value, the mean weight of a fish in the cell is given by

$$\overline{w}_c = \sum_a E_c(p(a))\, E_c(\text{weight} \mid a)$$

There is no explicit formula for $E_c(p(a))$, but the estimator

$$\hat{E}(p(a)) = \frac{\exp[E_c(\alpha^a_{c,h})]}{\sum_{a'=1}^{A} \exp[E_c(\alpha^{a'}_{c,h})]}$$

is almost unbiased, so long as the within-cell variance of $\alpha^a_{c,h}$ for each age is small compared with the difference between the $E(\alpha^a_{c,h})$ for the different ages within the cell. This is certainly the case for our data.

Results

The model can be fitted in about 30 min for 8 years of data on a modestly powerful computer. We take every tenth sample from an MCMC run of 10 000, after a burn-in of 1000 samples. There are no problems with convergence, and the thinning removes the serial correlation in the samples. We obtain as output the joint posterior distribution of total catch for each age group for any combination of cells as well as the (joint) posteriors of all the individual parameters. We expect that interest will usually be mostly in the catch-at-age results, along with the uncertainty in the estimates, which is directly available from the posteriors. One example of these results is shown (Fig. 3) where the posterior means for each age, along with 95% credible intervals, are plotted for 1995 and 2002 using data from 1995 to 2002. Obviously, the 12 age groups are not independent, but the error bars give a good indication of the uncertainty. If, for example,

Fig. 3. Estimates of catch-at-age in (*a*) 1995 and (*b*) 2002 using only *Amigo* data (crosses) and all data (circles). Bars are 95% credible intervals.

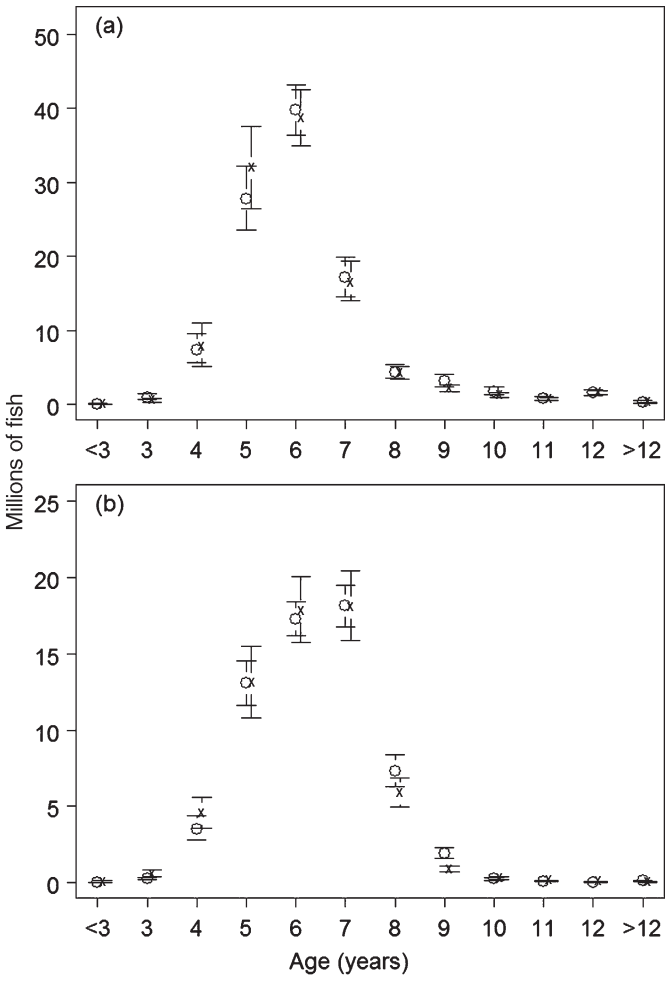


Fig. 4. Numbers of *Amigo* hauls (crosses) and extra hauls (circles) sampled in each year.

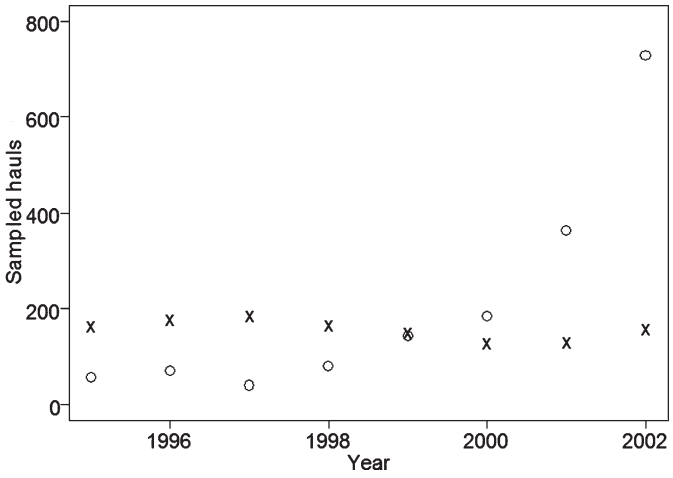
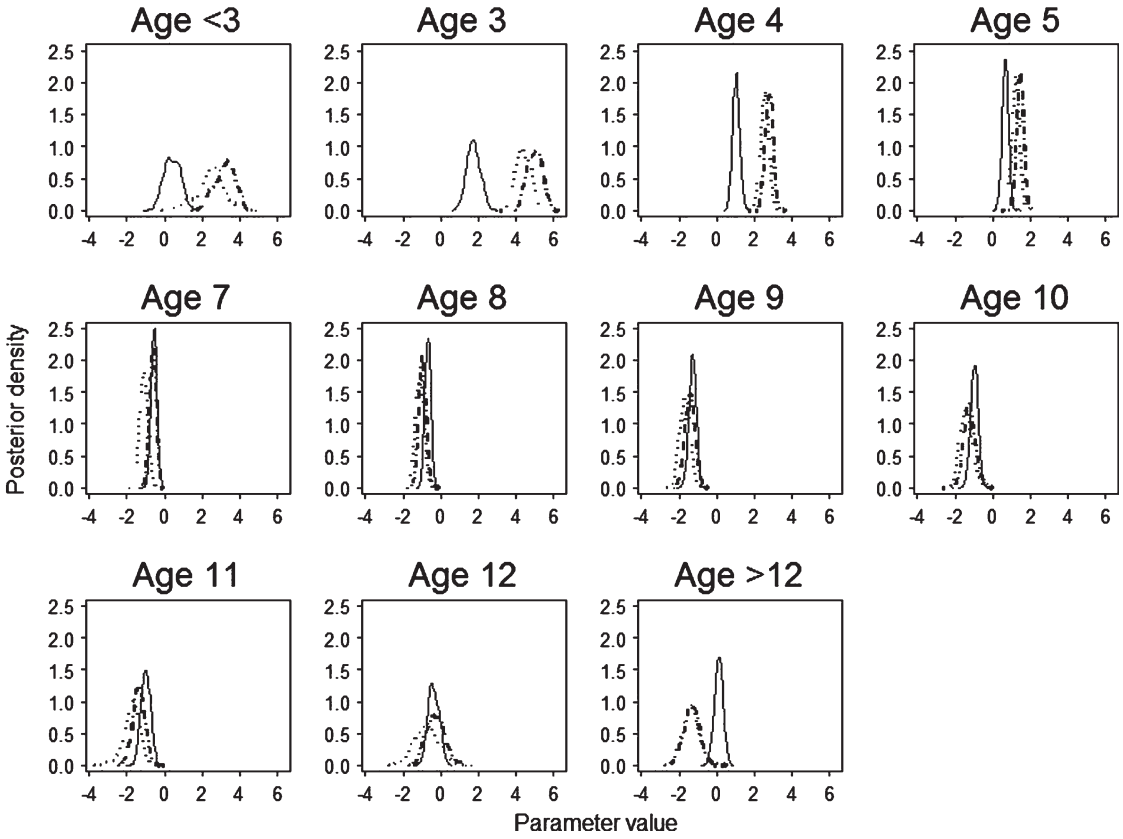


Fig. 5. Posterior distributions of the standard deviations of the random effects in the proportion-at-age model. Solid line, area standard deviation; dotted line, haul standard deviation; dot-dashed line, cell standard deviation.

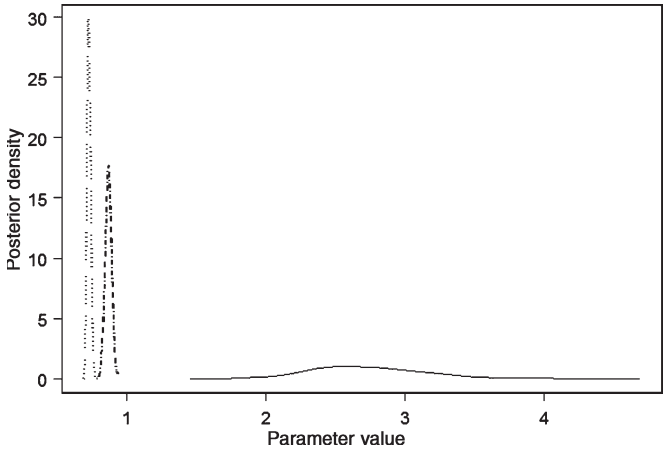


the results were to be used for a virtual population analysis, samples from the full joint posteriors for all required years would be available. These could be used to obtain the posterior distributions of the parameters calculated in the virtual population analysis.

These plots also illustrate the effect of using the length-only data in addition to the length and age data from the *Amigo*. The numbers of *Amigo* and extra (i.e., reference fleet and Coast Guard) hauls sampled per year are shown in Fig. 4. The extra samples are virtually all length only, although it is expected that in the future, there will be more age-given-length samples. In 1995, there were only about 50% as many extra samples as *Amigo* samples, and there is almost no difference in the results. By 2002, however, there were about four times as many extra as *Amigo* samples, and there is a useful reduction in the size of the error bars. Also, some of the point estimates have moved outside the *Amigo*-only intervals.

The model parameters are also of interest, and we show some examples in Figs. 5 and 6. The first (Fig. 5) shows the posterior distributions of standard deviations of the random effects in the age model, i.e., $\sqrt{1/\tau_{\text{age}}^{\text{cell}}}$, $\sqrt{1/\tau_{\text{age}}^{\text{haul}}}$, and $\sqrt{1/\tau_{\text{age}}^{\text{region}}}$. The region standard deviation is the most uncertain, not surprisingly, since we only use eight regions in its estimation. The haul standard deviation is the most precisely estimated. Note that the region standard deviation should be scaled by the number of neighbours of a region in the distribution of $\varsigma_r^{\text{region},a}$, and so, for a region with several neighbours, the posterior mean of the standard deviation would in fact be smaller than the posterior means of the other two standard deviations. It is difficult to make any useful interpretation of these parameters, except to see which are well or badly estimated.

Fig. 6. Posterior distributions of the season effects in the proportion-at-age model. Solid line, season 2; dotted line, season 3; dot-dashed line, season 4. Season 1 is defined to be zero.



The second example (Fig. 6) shows the posterior means of the season effects in the age model (i.e., $\alpha_s^{\text{season},a}$). Note that the values for age 6 and season 1 are defined to be zero. A high value for this parameter means a higher probability of catching fish from age group a in season s. Take, for example, the first plot (age < 3). The value for season 1 is defined to be zero. The mean for season 2 (solid line) is slightly greater than zero; the mean for season 3 (dotted line) is big-

ger again, and for season 4 (dot-dashed line), it is even bigger. Thus, the proportion of this age group in the catch increases with season. This may be because the fish get significantly bigger. This effect can be seen to a lesser degree for ages 3 and 4, but it disappears, or maybe even reverses, for older fish.

Discussion

The model described is as far as we know the first comprehensive approach to analysing multiple sources of catch-at-age data, in a way that can include all types of sampling schemes that we know of. It is explicit in its assumptions and given these assumptions properly accounts for the uncertainty in the estimation. It is very fast (at least compared with traditional methods) and in addition to the catch-at-age estimates can also give information on the model parameters, which may be interesting biologically.

We have shown herein that adding length-only samples to the age samples improves the precision of the catch-at-age estimates. The relatively small improvement even with a large sample of lengths reflects the relative lack of information in these samples. One very useful function of our model would be to explore the effect of adding more length samples or changing the sampling strategy in some other way. It would be possible to simulate from the model and investigate any desired changes in numbers and locations of samples. We hope to do this in a later paper.

There are a number of additions and improvements that could be made. Perhaps the most interesting would be to include errors in the age reading. This was done for the simpler sampling scheme of Hirst et al. (2004) and with some development could be included in this model. This may be important because unpublished work from the IMR suggests that on average, about 10% of the ages of cod may be wrong by 1 year, increasing up to 40% for older fish. It would also be possible to include different length-given-age or weight-given-length models, which may be appropriate for different fish species, or a different spatial model that may suit different fisheries.

Acknowledgements

This work was funded by the Norwegian Institute of Marine Research and the Norwegian Research Council, project 154079/420.

References

Aanes, S., and Pennington, M. 2003. On estimating the age composition of the commercial catch of Northeast Arctic cod from a sample of clusters. ICES J. Mar. Sci. **60**: 297–303.

Campana, S.E. 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. J. Fish Biol. **59**: 197–242.

Carlin, B.P., and Louis, T.A. 1996. Bayes and empirical Bayes methods for data analysis. Chapman and Hall, London, UK.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 1995. Bayesian data analysis. Chapman and Hall, London, UK.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. 1996. Markov chain Monte Carlo in practice. Chapman and Hall, London, UK.

Haddon, M. 2001. Modelling and quantative methods in fisheries. Chapman and Hall, London, UK.

Hilborn, R., and Lierman, M. 1998. Standing on the shoulders of giants: learning from experience in fisheries. Rev. Fish Biol. Fish. **8**: 273–283.

Hirst, D.J., Aanes, S., Storvik, G., Huseby, R.B., and Tvete, I.F. 2004. Estimating catch-at-age from market sampling data using a Bayesian hierarchical model. Appl. Stat. **53**: 1–14.

Lewy, P., and Nielsen, A. 2003. Modelling stochastic fish stock dynamics using Markov Chain Monte Carlo. ICES J. Mar. Sci. **60**: 743–752.

Lierman, M., and Hilborn, R. 1997. Depensation in fish stocks: a hierarchic Bayesian meta-analysis. Can. J. Fish. Aquat. Sci. **54**: 1976–1984.

Millar, R.B., and Meyer, R. 2000. Non-linear state space modelling of fisheries biomass dynamics by using Metropolis–Hastings within-Gibbs sampling. Appl. Stat. **49**: 327–342.

Prevost, E., Parent, E., Crozier, W., Davidson, I., Dumas, J., Gudbergsson, G., Hindar, K., McGinnity, P., MacLean, J., and Saettem, L.A. 2003. Setting biological reference points for Atlantic salmon stocks: transfer of information from data-rich to sparse-data situations by Bayesian hierarchical modelling. ICES J. Mar. Sci. **60**: 1177–1193.

von Bertalanffy, L. 1938. A quantitative theory of organic growth. Hum. Biol. **10**: 181–213.

Appendix A. Details of the model.

In all three models (proportion-at-age, length-given-age, and weight-given-length), the fixed effect parameter values are relative to the baseline terms $\alpha^{\text{base},a}$, β^{base}, and δ^{base}, and it is necessary to set one level of each fixed effect to zero for identifiability:

$$\alpha_{y^*}^{\text{year},a} = \alpha_{s^*}^{\text{season},a} = \alpha_{g^*}^{\text{gear},a} = \beta_{y^*}^{\text{year}} = \beta_{s^*}^{\text{season}} = \beta_{g^*}^{\text{gear}} = \delta_{y^*}^{\text{year}} = \delta_{s^*}^{\text{season}} = \delta_{g^*}^{\text{gear}} = 0$$

For the proportion-at-age model, the proportions must sum to 1, and so we have the additional restriction that all parameters for one age group a^* are set to zero:

$$\alpha^{\text{base},\,a^*} = \alpha_y^{\text{year},a^*} = \alpha_s^{\text{season},a^*} = \alpha_g^{\text{gear},a^*} = 0$$

We use $a^* = 6$ (usually the most common age group) and $y^* = s^* = g^* = 1$. The choice of y^*, s^*, and g^* is unimportant, but convergence is fastest if a^* is one of the most common age groups. Setting the parameters to zero for some value of a^* has the undesirable effect of giving the catch-at-age for this age group a smaller posterior variance than for the other age groups. A better restriction might be to make the mean over all age groups constant.

We give all nonzero fixed effects noninformative prior distributions:

$$\alpha^{\text{base},a} \sim N(0,1/0.001)\ \forall\, a \neq a^*$$

$$\alpha_y^{\text{year},a} \sim N(0,1/0.001)\ \forall\, y \neq y^*, a \neq a^*$$

$$\alpha_s^{\text{season},a} \sim N(0,1/0.001)\ \forall\, s \neq s^*, a \neq a^*$$

$$\alpha_g^{\text{gear},a} \sim N(0,1/0.001)\ \forall\, g \neq g^*, a \neq a^*$$

and so on for the β and δ terms.

The spatial terms $\varsigma_r^{\text{region},a}$, $\varepsilon_r^{\text{region}}$, and ν_r^{region} have Gaussian conditional autoregressive prior distributions (e.g., see Carlin and Louis 1996):

$$\varsigma_r^{\text{region},a^*} = 0$$

$$\varsigma_r^{\text{region},a^*} \left| \varsigma_{j\neq r}^{\text{region},a} \sim N\left(\overline{\varsigma}_r^{\text{region},a}, \frac{1}{\tau_{\text{age}}^{\text{region}} n_r} \right), \quad a \neq a^* \right.$$

$$\overline{\varsigma}_r^{\text{region},a} = n_r^{-1} \sum_{j\in\partial(r)} \varsigma_j^{\text{region},a}$$

where $\partial(r)$ is the set of neighbours of region r and n_r is the number of neighbours of region r.

The priors for $\varepsilon_r^{\text{region}}$ and ν_r^{region} are similar.

The ς, ε, and ν terms are independent random effects, again set to zero for $a = a^*$, with the following priors:

$$\varsigma_c^{\text{cell},a} \sim N(0,1/\tau_{\text{age}}^{\text{cell}}) \forall\, c, a \neq a^*$$

$$\varsigma_h^{\text{haul},a} \sim N(0,1/\tau_{\text{age}}^{\text{haul}}) \forall\, h$$

$$\varsigma_c^{\text{cell},a^*} = \varsigma_h^{\text{haul},a^*} = 0 \,\forall\, c, h$$

The ε and ν priors are similar. All precision terms τ are given vague Gamma(0.01,0.01) priors.

A sensitivity analysis showed no effect of varying the priors. It is known that in some models, the choice of prior for the precision terms can be very influential on the results, but in our case, this is not true. The reason for this is probably that there are only a relatively small number of precisions to estimate, and there is a large amount of data. The only noticeable effect of varying the precision priors between Gamma(0.1,0.1) and Gamma(0.001,0.001) was to slow the sampling routine down somewhat for the vaguest priors.

Note that throughout the paper, we have parameterized the Gaussian distributions in terms of the precision (i.e., the reciprocal of the variance), i.e., we write $N(\mu,1/\tau)$ rather than, for example, $N(\mu,\sigma^2)$. This is the usual Bayesian notation used for algebraic simplicity because we usually give τ a Gamma prior distribution.

References

Carlin, B.P., and Louis, T.A. 1996. Bayes and empirical Bayes methods for data analysis. Chapman and Hall, London, UK.

Appendix B. Markov Chain Monte Carlo (MCMC) algorithm.

Denote by θ_{age}, θ_{lga}, and θ_{wgl} the set of parameters involved in the age, length-given-age, and weight-given-length model, respectively. Our goal is to simulate from the posterior distribution $p(\theta_{\text{age}},\theta_{lga},\theta_{wgl}|\text{data})$. Note first that

$$p(\theta_{\text{age}},\theta_{lga},\theta_{wgl}|\text{data}) = p(\theta_{\text{age}},\theta_{lga}|\text{data})p(\theta_{wgl}|\text{data})$$

so that simulation of θ_{wgl} can be performed separately from $(\theta_{\text{age}},\theta_{lga})$. In the current implementation, the simulation of θ_{wgl} is performed using Gibbs sampling by sequentially altering between the following steps:

(*i*) Simulate fixed effects conditional on random effects, precision parameters, and data.

(*ii*) Simulate random effects conditional on fixed effects, precision parameters, and data.

(*iii*) Simulate precision parameters conditional on fixed effects, random effects, and data.

Since this part of the model is an ordinary regression model (with random effects), each step can be performed by standard methods (e.g., see Gelman et al. 1995, chap. 8).

Concerning the simulation of $(\theta_{\text{age}},\theta_{lga})$, note that these two sets of variables become dependent because of the length-only data. Denote by a_{miss} the set of missing ages corresponding to the length-only data. Simulation of $(\theta_{\text{age}},\theta_{lga})$ is performed through the following three main steps:

(*i*) Simulate the missing ages a_{miss} conditional on $(\theta_{\text{age}},\theta_{lga})$ and data.

(*ii*) Simulate θ_{age} conditional on θ_{lga}, a_{miss}, and data.

(*iii*) Simulate θ_{lga} conditional on θ_{age}, a_{miss}, and data.

Given $(\theta_{\text{age}},\theta_{lga})$, all missing ages are independent multinomial variables, making them easy to draw.

For given a_{miss}, the length-given-age model is a standard regression model similar to the weight-given-length model and simulation of θ_{lga} can be performed as for θ_{wgl}. The simulation of θ_{age} is more difficult owing to the nonlinear dependence between data and parameters. In the current implementation, simulation of θ_{age} is performed through the following two main steps:

(*i*) Simulate the $\alpha_{c,h}^a$ given the rest of θ_{age} and data.

(*ii*) Simulate the rest of θ_{age} given the $\alpha_{c,h}^a$ and data.

Given the rest of θ_{age}, all $\alpha_{c,h}^a$ are independent and can be simulated through Metropolis–Hastings steps. Given the $\alpha_{c,h}^a$, the rest of the age model can be considered as a standard regression model and simulation of the remaining parameters can be performed similar to the simulation of θ_{lga} and θ_{wgl}.

References

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 1995. Bayesian data analysis. Chapman and Hall, London, UK.