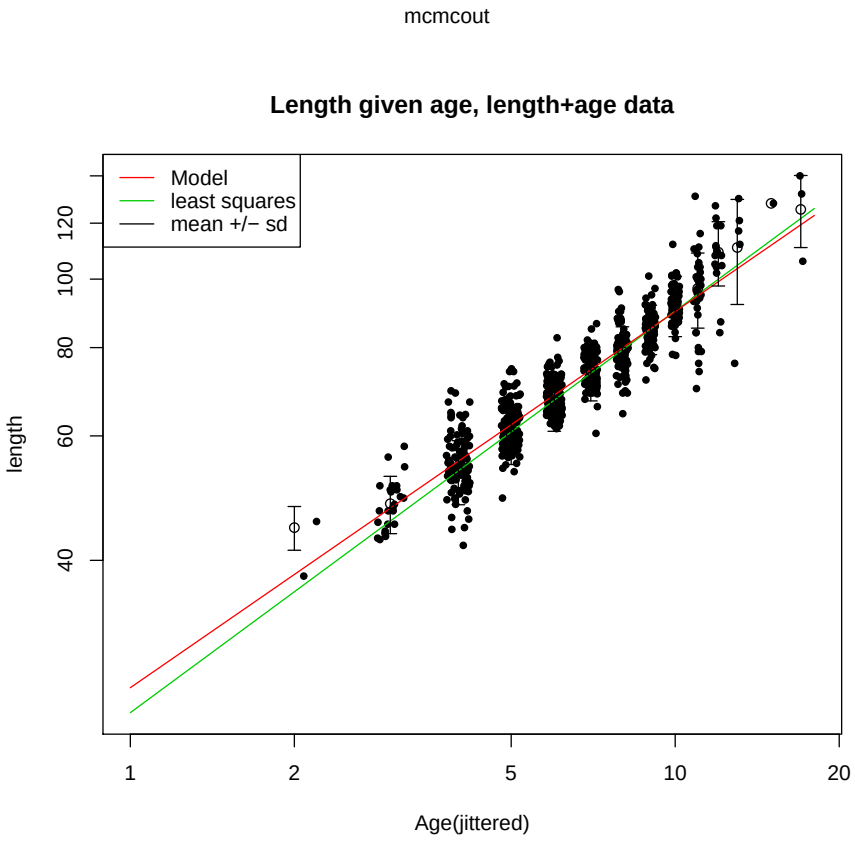
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Catch-at-age – Version 4.0:

Technical Report



Note no SAMBA/54/16

Authors Hanne Rognebakke
David Hirst
Sondre Aanes
Geir Storvik

Date December 2016

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Catch-at-age – Version 4.0: Technical Report**
Authors	**Hanne Rognebakke** <hanne.rognebakke@nr.no> **David Hirst** <david.hirst@nr.no> **Sondre Aanes** <sondre.aanes@nr.no> **Geir Storvik** <geirs@math.uio.no>
Date	December 2016
Publication number	SAMBA/54/16

Abstract

The Norwegian Computing Center and the Institute of Marine Research have over years developed a Bayesian hierarchical model to estimate the catch-at-age of fish. Such a model enables us to obtain estimates of the catch-at-age with appropriate uncertainty. This is considered as essential input in most age structured stock assessment processes.

Recent improvements of the model include modelling a haulsize effect in the proportion-at-age model. The model is implemented in C with an R interface. This note gives a thorough description of the model and the simulation algorithm corresponding to version 4.0 of the program.

Keywords	Bayesian hierarchical models, MCMC
Target group	Institute of Marine Research
Availability	Open
Project	
Project number	220305
Research field	Statistics for Climate, Environment, Marine Resources and Health
Number of pages	28

Contents

1 Introduction

Predicting the catch-at-age, i.e. the number of fish caught within each age group, of commercial fish species is an important part of the quota-setting process for many different species. The Norwegian Computing Center (NR) and the Institute of Marine Research (IMR) have over years developed a Bayesian hierarchical model to estimate the catch-at-age of fish, see Hirst et al. (2004, 2005, 2012). The model aims to estimate both the proportion-at-age and the mean weight of fish, in order to convert total landings statistics to numbers at age.

The model is implemented in C with an R interface. The model has been revised and extended over the years, and modifications have been made to the simulation algorithm in order to make it more robust to different kind of data sets. This note gives a thorough description of the model and the simulation algorithm for Version 4.0.

1.1 Overview

Chapter 2 describes the model. The main parts of the model are given in Section 2.1 (proportion-at-age), Section 2.2 (length-given-age) and Section 2.3 (weight-given-length). There can be errors in the age readings, which is described in Section 2.4. The model can handle multiple species, and it is possible to include uncertainty in classification into the different species. This is described in Section 2.5. A recent improvement of the model is including a haulsize effect in the proportion-at-age model, which is described in Section 2.6. Details of the simulation algorithm is given in Chapter 3. Chapter 4 describes estimating the catch-at-age. Finally, there is a short description of the connection between the C program and the R interface in Chapter 5.

2 Model description

In order to predict the catch-at-age, i.e. the number of fish caught within each age group, we need to estimate the proportion-at-age and the mean weight-at-age of fish caught. Assuming that we know the total weight landed, we can then convert the proportions-at-age to numbers-at-age.

The data are sampled in different years, regions and seasons, and with different gears. Typically, predictions are performed on cell levels, where one cell that we consider represents one gear in one region in one season of one year. We assume that the total weight landed is known for each cell. Depending on the observation scheme, modelling can be made either on haul level or on trip level. The index u will be used to describe the modelling unit of choice. It is well known that between-unit variation in catch composition can be very large and it is crucial to take this into account. This is built into the model by cell- and unit-specific random effects.

It would be possible to model weight-given-age directly, but modelling it via length enables us to get a better estimate of the mean weight-at-age in cells with length but no age and weight data. In cells with age data there are usually also a large amount of length-only data available. Hence, we have developed models for proportion-at-age, length-given-age and weight-given-length. Let $a_{c,u,f}$, $L_{c,u,f}$ and $W_{c,u,f}$ denote the age, length and weight, respectively, of the fth fish from unit u in cell c. In all cases, the models are described including all the covariates, but alternatives can be obtained by removing different terms.

2.1 Model for proportion-at-age

The samples from a boat are assumed to be randomly drawn from the total population of fish in that unit. Further, the units are assumed to be randomly sampled from all of those within the appropriate cell. Hence, the numbers-at-age in a sample from a given unit and cell, $\boldsymbol{N}_{c,u}$, is given by the multinomal distribution

$$\boldsymbol{N}_{c,u} \sim \text{multinomial}(\boldsymbol{p}_{c,u}, n_{c,u}),$$

where $\boldsymbol{p}_{c,u}$ is a vector of proportions-at-age in the unit, and $n_{c,u}$ is the number of fish sampled from the unit. We assume that $n_{c,u}$ is independent of $\boldsymbol{p}_{c,u}$. Age groups are indexed by $a \in \{1, ..., A\}$. Let $p_{c,u}(a)$ be the ath element in $\boldsymbol{p}_{c,u}$. Then, $0 \leq p_{c,u}(a) \leq 1$ and $\sum_a p_{c,u}(a) = 1$. We assume a multinomial logistic-type model (Dey et al., 2000) and reparametrise this as

$$\Pr(a_{c,u,f} = a) = p_{c,u}(a) = \frac{\exp(\alpha^a_{c,u})}{\sum_{a'} \exp(\alpha^{a'}_{c,u})}.$$

The parameters $\alpha^a_{c,u}$ are modelled in terms of various covariates as

$$\begin{aligned} \alpha^a_{c,u} =& \alpha^{const,a} + \alpha^{year,a}_{y(c)} + \alpha^{season,a}_{s(c)} + \alpha^{gear,a}_{g(c)} \\ &+ \zeta^{region,a}_{r(c)} + \zeta^{boat,a}_{b(c)} + \zeta^{cell,a}_{c} + \zeta^{unit,a}_{c,u}, \end{aligned} \tag{1}$$

where $y(c)$ 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 c and just refer to $p_u(a)$, $\alpha_y^{year,a}$, etc.

The main effects are the following terms: constant $\{\alpha^{const,a}\}$, year $\{\alpha_y^{year,a}\}$, season $\{\alpha_s^{season,a}\}$, gear $\{\alpha_g^{gear,a}\}$, region $\{\zeta_r^{region,a}\}$ and boat $\{\zeta_b^{boat,a}\}$. $\{\zeta_c^{cell,a}\}$ are cell-specific effects. $\{\zeta_{c,u}^{unit,a}\}$ are unit-specific effects, and are always included in the model. The α-terms are fixed effects and the ζ-terms are random effects. We assume that there will always be some data for all levels of the fixed effects that are of interest.

For identifiability, we assume

$$\begin{aligned}\sum_a \alpha^{const,a} &= 0,\\ \sum_a \alpha_y^{year,a} &= 0,\ \forall y,\\ \sum_y \alpha_y^{year,a} &= 0,\ \forall a,\end{aligned}$$

and likewise for the season and gear effects. We have deliberately avoided the "treatment" type constraint where one category is put to zero, since this gives an asymmetry in the priors in our Bayesian framework.

If region is included in the model, we model this by a spatially smoothed surface, allowing estimation of proportions-at-age in regions with no data. 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. We assume a conditional autoregressive (CAR) model (Besag, 1974)

$$[\zeta_{r_i}^{region,a}|\zeta_{r_{i'\neq i}}^{region,a}] = \mathcal{N}\left(\phi_{age,r} n_i^{-1} \sum_{i'\in\delta(i)} \zeta_{r_i}^{region,a}, \left(\tau_{age}^{region}[\phi_{age,r} n_i + 1 - \phi_{age,r}]\right)^{-1}\right),$$

where n_i is the number of neighbours of region i, while $\delta(i)$ is the set of neighbours of region i. Further, $\phi_{age,r}$ is the AR-parameter, which is assumed uniformly distributed between 0 and 1, and τ_{age}^{region} is the precision parameter. This model for the variance corresponds to making the precision matrix a sum of a spatial and an independent part.

The terms $\{\zeta_c^{cell,a}\}$ are independent random effects modelling the interactions between the main effects (see e.g. Gelman et al. (1995)). In other words, it models the differences between the fit from the main-effects-only model and the true cell means. The terms $\{\zeta_{c,u}^{unit,a}\}$ are independent random effects modelling the differences between units within a cell. These must be random effects because there are many cells and units with no data. We assume

$$\begin{aligned}\zeta_c^{cell,a} &\overset{iid}{\sim} \mathcal{N}\left(0, \tau_{age}^{cell\ -1}\right)\\ \zeta_{c,u}^{unit,a} &\overset{iid}{\sim} \mathcal{N}\left(0, \tau_{age}^{unit\ -1}\right),\end{aligned}$$

again with a sum-constraint over ages.

The full Bayesian hierarchical model is completed through specification of priors for the hyperparameters involved. All the parameters involved are assumed independent a priori. For the main effects, Gaussian priors are assumed. For some of these parameters, prior information is available and can be incorporated into the the prior. For precision parameters, Gamma distributions are used.

2.2 Model for length-given-age

There is a reasonable linear relationship between log(length) and log(age). From data analysis it appears that the slope is constant, but the intercepts vary between cells, and boats within a cell. However, for some data sets there is a nonlinear relationship, which may cause problems at high and low ages. Hence, the model has been extended to include both a linear relationship and a nonlinear relationship. The resulting model is

$$l_{u,f} = \log(L_{u,f}) = \beta_{0,u} + \beta_1 g(a_{u,f};\theta_g) + \epsilon_{u,f}^{fish},$$

where $\epsilon_{u,f}^{fish} \overset{iid}{\sim} \mathcal{N}\left(0, \tau_{lga}^{fish\ -1}\right)$ is the random within-unit variation in length-given-age.

The intercept is given by

$$\beta_{0,u} = \beta^{const} + \beta_y^{year} + \beta_s^{season} + \beta_g^{gear} + \epsilon_r^{region} + \epsilon_b^{boat} + \epsilon_c^{cell} + \epsilon_{c,u}^{unit},$$

where the β-terms are fixed effects similar to the α-terms in the proportion-at-age model. ϵ_r^{region} and ϵ_b^{boat} are the main effects for region and boat, respectively, and ϵ_c^{cell} and $\epsilon_{c,u}^{unit}$ are independent random effects similar to $\zeta_c^{cell,a}$ and $\zeta_{c,u}^{unit,a}$, respectively.

For the linear model we have

$$g(a_{u,f};\theta_g) = \log(a_{u,f}).$$

Otherwise we use a non-linear age-length model given by

$$g(a_{u,f};\theta_g) = \log\left[1 - \theta\exp(-\gamma a_{u,f}^c)\right], \tag{2}$$

where θ, γ and c are parameters to be estimated. In order to avoid identifiability problems with respect to $\beta_{0,u}$ and β_1, we have linearly transformed $g(\cdot)$ such that $g(1) = 0$ and $g(A) = 1$.

One should note that the age in the length-given-age model, $a_{u,f}$, should be as close as possible to the actual age of the fish f rather than the age-class a modelled in the age model. We use $a_{u,f} = a + season/S$, if there are S seasons in the data set numbered from 1 to S, and age-class $a = 1, 2, ..$ refer to fish of ages 1, 2 etc.

2.3 Model for weight-given-length

There is a strong linear relationship between log(length) and log(weight). Similarly to the length-given-age model, we use a constant slope while the intercepts vary between cells, and between boats within a cell. The resulting model is

$$w_{u,f} = \log(W_{u,f}) = \delta_{0,u} + \delta_1 \log(L_{u,f}) + \nu_{u,f}^{fish},$$

where $\nu_{u,f}^{fish} \overset{iid}{\sim} \mathcal{N}\left(0, \tau_{wgl}^{fish\ -1}\right)$ is the random within-unit variation in weight-given-length.

The intercept is given by

$$\delta_{0,u} = \delta^{const} + \delta_y^{year} + \delta_s^{season} + \delta_g^{gear} + \nu_r^{region} + \nu_b^{boat} + \nu_c^{cell} + \nu_{c,u}^{unit},$$

where the terms are similar to the terms in the length-given-age model.

2.4 Age uncertainty

If ages are read by errors, we assume the knowledge of an $A \times A$ transition matrix $\boldsymbol{E}$, where the columns give the conditional probability of the observed fish age, given the true age. Hence,

$$E_{i,j} = \Pr(a^{obs} = i | a = j).$$

2.5 Multiple stocks and classification error

Some stocks are regarded as consisting of multiple stocks, like the Norwegian cod stocks that consists of two stocks: Atlantic cod (skrei) found in deep water, and coastal cod found nearer the shore. The model above is described for one stock, but can easily be extended to handle multiple stocks.

In case of modelling coastal cod and Atlantic cod, the same model is used for proportion-at-age, but with $a \in \{1, ..., A, A+1, ..., 2A\}$. The first A age groups then corresponds to coastal cod, and the last A age groups corresponds to Atlantic cod. The parameters therefore have separate fixed effect terms, but common precisions for the random effect terms.

In the length-given-age model and weight-given-length model separate age-length and weight-length relationships are considered for each species.

It is difficult to distinguish between fish from the two stocks visually, and classification, which is usually based on reading the otoliths, divides the fish into certain or uncertain coastal cod or Atlantic cod. Uncertainty in classification is mostly due to the shape and pattern of the otoliths, rather than to the person who interpreted them. The majority of fish only have length measurements, giving no indication of which stock they come from. The total landings statistics also do not distinguish between the two stocks. Hence, it is of importance to be able to model these stocks simultaneously.

The uncertainty in classification is included in the model by regarding the classification to species as equivalent to classification into age groups. The only difference is that there are two different types of classification for both stocks, type 1 (which is "certain" and easy to classify) and type 2 (which is "uncertain" and harder to classify). If we make the assumption that a type 1 fish is never confused with a type 2 fish, then the new error

matrix takes the form

$$
\begin{array}{c}
 & \multicolumn{2}{c}{\text{Coastal cod (C)}} & \multicolumn{2}{c}{\text{Atlantic cod (A)}} \\
 & \text{Type1C} & \text{Type2C} & \text{Type1A} & \text{Type2A} \\
\text{Type1C} & pclass_1^C E_1^{CC} & 0 & (1-pclass_1^A)E_1^{AC} & 0 \\
\text{Type2C} & 0 & pclass_2^C E_2^{CC} & 0 & (1-pclass_2^A)E_2^{AC} \\
\text{Type1A} & (1-pclass_1^C)E_1^{CA} & 0 & pclass_1^A E_1^{AA} & 0 \\
\text{Type2A} & 0 & (1-pclass_2^C)E_2^{CA} & 0 & pclass_2^A E_2^{AA}
\end{array}
\Bigg).
$$

The probabilities $pclass_1^C$ and $pclass_2^C$ are the probabilities that a type 1 and type 2 coastal cod, respectively, will be correctly classified. Similarly, $pclass_1^A$ and $pclass_2^A$ are the probabilities that a type 1 and type 2 Atlantic cod, respectively, will be correctly classified. E_1^{CC} and E_2^{CC} are the age error matrices for coastal cod that are classified as type 1 and type 2 coastal cod, respectively, while E_1^{CA} and E_2^{CA} are the age error matrices for coastal cod that are misclassified as type 1 and type 2 Atlantic cod, respectively, and so on. Hence, the columns give the conditional probability of the observed type, given the true species. We can allow the age error matrices to be different for the certain and uncertain types.

For a thorough discussion of this model and example of results, see Rognebakke et al. (2011).

2.6 Haulsize effect in proportion-at-age model

If haulsize is included in the model, it is modelled as a continuous covariate. Let hsz_u be the haulsize in unit u that can be measured in either numbers or weight. The haulsize is modelled separately as

$$
\begin{aligned}
\log hsz_u =& \delta^{const} + \delta_y^{year} + \delta_s^{season} + \delta_g^{gear} + \nu_r^{region} \\
&+ \nu_c^{cell} + \epsilon_u^{haul}.
\end{aligned}
$$

The terms ϵ_u^{haul} are then included in the model for proportion-at-age as

$$
\begin{aligned}
\alpha_u^a =& \alpha^{const,a} + \alpha_y^{year,a} + \alpha_s^{season,a} + \alpha_g^{gear,a} + \zeta_r^{region,a} + \zeta_b^{boat,a} + \zeta_c^{cell,a} \\
&+ \alpha_u^{hsz,a}\epsilon_u^{haul} + \zeta_{c,u}^{unit,a}.
\end{aligned}
$$

3 Simulation algorithm

Inference on the unknown parameters are obtained by using Bayesian hierarchical modeling (Gelman et al., 1995). The inference is performed through an MCMC algorithm (Gilks et al., 1996) using block-updating (Knorr-Held and Rue, 2002; Roberts and Sahu, 1997). A main challenge has been to develop an algorithm that is robust with respect to different kinds of data. Let $\boldsymbol{\theta}_a$, $\boldsymbol{\theta}_l$, $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_w$ denote the parameters in the proportion-at-age model, the length-given-age model except the g-function, the g-function and the weight-given-length model, respectively. We are interested in simulating from the posterior distribution $p(\boldsymbol{\theta}_a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g, \boldsymbol{\theta}_w|\text{data})$. Since

$$p(\boldsymbol{\theta}_a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g, \boldsymbol{\theta}_w|\text{data}) = p(\boldsymbol{\theta}_a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g|\text{data})p(\boldsymbol{\theta}_w|\text{data}), \tag{3}$$

we can perform the simulation of $\boldsymbol{\theta}_w$ separately from the other parameters. Due to the length-only data, the parameter sets $\boldsymbol{\theta}_a$ and $(\boldsymbol{\theta}_l, \boldsymbol{\theta}_g)$ become dependent.

There are three types of data available for estimating the age model and the length-given-age model; age-length data, age-given-length data and length-only data. In version 1.0 all data were used when estimating all parameters. In version 2.0 only data from units where there are some observed ages are included when estimating the parameters in the length-given-age model (including the g-function). The exception is when estimating the unit effect, where all the data is used.

The simulation of the different parameter sets are described in Section 3.1 ($\boldsymbol{\theta}_a$, $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_l$) and Section 3.2 ($\boldsymbol{\theta}_w$). Sampling from the model when including age errors or classification error if multiple stocks is described separately in Section 3.1.1. Simulation of the haulsize effect, if included in the age model, is described in Section 3.1.5.

3.1 Simulation of age model and length-given-age model

If the ages and lengths of all the sampled fish were known, the parameter sets of the age model and length-given-age model would be independent. However, since there are missing ages, these become dependent. Let $\boldsymbol{a}_{miss}$ denote the set of missing ages corresponding to the length-only data. The posterior distribution of interest is then

$$p(\boldsymbol{a}_{miss}, \boldsymbol{\theta}_a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g | \text{data}).$$

Using a Gibbs sampling scheme we alternate between

- Sample missing ages $\boldsymbol{a}_{miss}$ conditional on $(\boldsymbol{\theta}_a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g)$ and data.
- Sample $\boldsymbol{\theta}_a$ conditional on $(\boldsymbol{\theta}_l, \boldsymbol{\theta}_g)$, $\boldsymbol{a}_{miss}$, and data.
- If non-linear model, sample $\boldsymbol{\theta}_g$ conditional on $(\boldsymbol{\theta}_l, \boldsymbol{\theta}_a)$, $\boldsymbol{a}_{miss}$, and data.
- Sample $\boldsymbol{\theta}_l$ conditional on $(\boldsymbol{\theta}_a, \boldsymbol{\theta}_g)$, $\boldsymbol{a}_{miss}$, and data.

3.1.1 Sample missing ages, $\boldsymbol{a_{miss}}$, if any.

Many fish contain only length-observations. Given all other parameters and observations, ages are independent. Sampling missing ages is then easy to perform by multinomial sampling on each fish. However, this part becomes time-consuming if there are many missing ages. We simulate the missing age from

$$p(a_{u,f}|l_{u,f}, \boldsymbol{\theta}_a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g) \propto p(a_{u,f}|\boldsymbol{\theta}_a) p(l_{u,f}|a_{u,f}, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g). \tag{4}$$

In practice, the length observations are given in length intervals, i.e. $l_{u,f} \in (l^l_{u,f}, l^u_{u,f})$. The probability $p(l_{u,f}|a_{u,f}, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g)$ is given by

$$p(l_{u,f}|a_{u,f}, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g) \approx \Phi(l^u_{u,f}|a_{u,f}, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g) - \Phi(l^l_{u,f}|a_{u,f}, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g),$$

where $\Phi(\cdot)$ is the cumulative Gaussian distribution.

For starting values, we construct a frequency matrix $P(a|l, season)$ for a finite number of lengths using the observed ages, and simulate the missing ages from this.

• Age uncertainty

If there are errors in age-readings, we do not only sample the missing ages a_{miss}. We also resample the observed ages taking the age error matrix into account. Hence, we sample from

$$p(a_{u,f}|a^{obs}_{u,f}, l_{u,f}, \boldsymbol{\theta}_a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g) \propto p(a_{u,f}|\boldsymbol{\theta}_a) p(a^{obs}_{u,f}|a_{u,f}) p(l_{u,f}|a_{u,f}, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g),$$

where $a^{obs}_{u,f}$ is the observed age and $p(a^{obs}_{u,f}|a_{u,f})$ is given by the age error matrix $\boldsymbol{E}$.

One should note that we still use only data from units where there are some observed ages when estimating the parameters in the length-given-age model (except the unit effect).

• Multiple stocks and classification error

In the case when we have both coastal cod and Atlantic cod, we simulate missing ages from (4). If ages are observed we resample the type using

$$p(type|type^{obs}, a_{u,f}, l_{u,f}) = \frac{p(type^{obs}|type)p(type|a_{u,f})}{\sum_{i\in\{1C,2C,1A,2A\}} p(type^{obs}|type=i)p(type=i|a_{u,f})} p(l_{u,f}|a_{u,f}),$$

where $type$ is the true type and $type^{obs}$ is the observed type. Further, $p(type^{obs}|type)$ is given from the classification error matrix, e.g., $p(type^{obs} = 1C|type = 1C) = pclass_1^C$. From the proportion-at-age parameters, we can only find the probability that the true type is either coastal cod or Atlantic cod, and not differ between type 1 and type 2 fish. Hence, we have introduced two parameters, k_C and k_A, that estimates the proportion of type 1 coastal cod to all coastal cod, and proportion of type 1 Atlantic cod to all Atlantic cod, respectively. This defines the following probabilities

$$\begin{aligned}
p(type = 1C|a_{u,f}) &= k_C p(a_{u,f}) \\
p(type = 2C|a_{u,f}) &= (1-k_C)p(a_{u,f}) \\
p(type = 1A|a_{u,f}) &= k_A p(a_{u,f})) \\
p(type = 2A|a_{u,f}) &= (1-k_A)p(a_{u,f})).
\end{aligned}$$

Since we have made the assumption that a type 1 fish is never confused with a type 2 fish, then $p(type = 1C|type^{obs} = 2C, a_{u,f}) = 0$ etc.

After all missing ages are sampled and ages with classification error are resampled, we sample the parameters k_C and k_A. We assume a Beta(1.0,1.0) prior for the parameters. Hence, the posterior distributions are given by

$$\begin{aligned}
k_C|\text{data} &\sim \text{Beta}(1.0 + N_1^C, 1.0 + N^C - N_1^C) \\
k_A|\text{data} &\sim \text{Beta}(1.0 + N_1^A, 1.0 + N^A - N_1^A),
\end{aligned}$$

where N_1^C and N_1^A are the number of type 1 coastal cod and Atlantic cod, respectively, and N^C and N^A are the total number (type 1 and type2) of coastal cod and Atlantic cod, respectively.

3.1.2 Sample parameters in age model (θ_a).

The parameters θ_a consists of linear regression parameters, random effects and precision parameters. We have divided these variables into three sub-blocks, the first containing all the main effects and cell effects in (1), the second containing the terms $\{\alpha_u^a\}$ and the third containing all the precision parameters and the AR-parameter. We update these sub-blocks sequentially. One problem with this approach is that there is a large dependency between these blocks, which gives slow convergence.

Note that the likelihood for the age model is given by

$$\prod_{u=1}^{U}\prod_{f=1}^{N_u} p(a_{u,f};\boldsymbol{\theta_a}) = \prod_{u=1}^{U}\prod_{a=1}^{A} p_u(a;\boldsymbol{\theta_a})^{N_u(a)},$$

where N_u is the number of fish in unit u and $N_u(a)$ is the number of fish in age group a from unit u and U is the total number of units. This shows that $\{N_u(a)\}$ are sufficient statistics for updating $\boldsymbol{\theta_a}$. However, due to sampling of missing ages (or resampling of ages if age-reading errors, see 3.1.1), the sufficient statistics must be updated for each iteration.

If there is a large number of age groups with hardly any fish, so that for some levels of the covariates there may be no fish of that age at all, it can be difficult to estimate the covariates. In order to improved the estimation we have added a small amount, δ_{age}, to the probability of each age group, in each units where there are no missing ages. An appropriate value is 0.005, which must be specified by the user. Otherwise, the default value is 0. This amount is then subtracted when the units are simulated to estimate the catch-at-age.

• Sample fixed and random effects, $\boldsymbol{\alpha^{const,a}}$, $\boldsymbol{\alpha_y^{year,a}}$, $\boldsymbol{\alpha_s^{season,a}}$, $\boldsymbol{\alpha_g^{gear,a}}$, $\boldsymbol{\alpha_u^{hsz,a}}$, $\boldsymbol{\zeta_r^{region,a}}$, $\boldsymbol{\zeta_b^{boat}}$ and $\boldsymbol{\zeta_c^{cell,a}}$.

All nonzero fixed effects are given non-informative Gaussian prior distributions with zero mean. The precisions in the prior distribution are input to the C-program, and are currently 0.0001 for the constant term and 0.001 for the other fixed effects.

Given $\{\alpha_u^a\}$, the main effects and the cell effects are ordinary linear Gaussian regression parameters. Hence, the conditional distribution will also be multivariate Gaussian. Note that the observation precision is the unit precision, τ_{age}^{unit}. The parameter vector can be large, especially if there are many cell effects. However, the precision matrix in the conditional distribution is sparse, which makes the computation involved fast (Rue and Held, 2005, Chapter 2). We only simulate those cell effects for which we have observations. We use the `GMRFLib`[1] library (Rue and Held, 2005, Appendix B) to sample from the conditional distribution. Using this library, it is also easy to incorporate the sum-constraints involved, see Appendix A.1.

• Sample unit effects, $\boldsymbol{\zeta_u^{unit,a}}$.

Given all the other parameters, dependency only occur within units for α_u^a, making it possible to perform simulation on units independently. Further, $\{\alpha_u^a, a = 0, ..., A\}$ has a multivariate Gaussian prior distribution, with the sum constrained to be zero. The observations are numbers at age $\{N_u(a), a = 0, ..., A\}$, which follow a multinomial distribution with probabilities $p_u(a)$. Updating is performed through Metropolis-Hastings independence sampling steps. Within unit, the parameters have the sum constrained to be zero. Hence, only the first $A - 1$ variables are updated, and the last is calculated from these.

A multivariate Gaussian proposal distribution is constructed by searching for the conditional mode running a few Newton-Raphson iterations (maximum 4 iterations) with the prior means as initial values. After finding the mode, it is fast to generate proposal samples. Hence, several Metropolis-Hastings steps (currently 10 steps) are performed

1. A C-library freely available from `http://www.math.ntnu.no/~hrue/GMRFLib/doc/html/`

to obtain a sample of the unit effects. Details on the Metropolis-Hastings algorithm are given in Appendix A.2.

• Sample precision parameters, τ_{age}^{unit}, τ_{age}^{cell}, τ_{age}^{region} and τ_{age}^{boat}, and AR-parameter, $\phi_{age,r}$.

For all the precision parameters we use Gamma priors, mainly for computational convenience. The parameters in the Gamma priors are input to the C-program, and currently Gamma(0.01,0.01) are used for all precision parameters.

Given all fixed and random effects and with independent Gamma priors on the precision parameters, the conditional distributions for the precision parameters become independent Gamma distributions.

Sample τ_{age}^{unit}.

The posterior distribution of interest is given by

$$p(\tau_{age}^{unit}|\boldsymbol{\theta}_a, \text{constraints}) \propto p(\tau_{age}^{unit}) \prod_{u=1}^{U}\prod_{a=1}^{A} p(\alpha_u^a|\tau_{age}^{unit}, \text{constraints})$$
$$\propto (\tau_{age}^{unit})^{n/2+0.01-1} \exp\left\{-\left(\frac{1}{2}\sum_{u=1}^{U}\sum_{a=1}^{A}(\alpha_u^a - \mu_\alpha^a)^2 + 0.01\right)\tau_{age}^{unit}\right\},$$

where $n = U(A-1)$ and μ_u^a is the sum of the right hand side of (1), except the unit effect $\zeta_u^{unit,a}$. See Appendix A.3 for details regarding n due to the constraints.

Sample $\phi_{age,r}$.

$\phi_{age,r}$ is the spatial AR-parameter. It is assumed uniformly distributed between 0 and 1. Sampling is performed by calculating probabilities on a discrete number of outcomes between 0 and 1. Gibbs sampling is then performed in order to obtain a sample.

3.1.3 Sample parameters in g-function (θ_g), if non-linear model.

In this section we describe the sampling of the parameters in the non-linear g-function in (2), θ, γ and c.

In theory all parameters could be estimated from data. We would then perform Metropolis-Hastings steps on each parameter involved at a time. However, there do not seem to be almost any information regarding θ in the data, and simulations indicate that there are many parameter sets that define the same function. Hence, we have fixed two parameters; $\theta = 0.5$ and $c = 1$.

Details of sampling the γ parameter is given in Appendix A.4. Because each step is computationally fast compared to updating the other variables involved, a fixed number of iterations (currently 1000) are performed, giving in practice samples from the conditional posterior.

3.1.4 Sampling of parameters in length-given-age model (θ_l).

The parameters in the length-given-age model (apart from the g-function) are the linear regression parameters and the precision parameters. The likelihood contribution can

be written as a function of a few sufficient statistics, which must be calculated in each iteration. This speeds up the computation considerably.

We divide the parameters into three sub-blocks, the first containing all the main effects and cell effects in (2.2) together with the slope parameter, the second containing the unit effects $\{\epsilon_{c,u}^{unit}\}$ and the third containing all the precision parameters and the AR-parameter. We update these blocks sequentially. Remember that we only use data from units where there are some observed ages when estimating the parameters in the length-given-age model. The only exception is when estimating the unit effects, where all the data is used.

• Sample fixed and random effects, $\boldsymbol{\beta^{const}}$, $\boldsymbol{\beta_y^{year}}$, $\boldsymbol{\beta_s^{season}}$, $\boldsymbol{\beta_g^{gear}}$, $\boldsymbol{\beta_u^{hsz}}$ $\boldsymbol{\epsilon_r^{region}}$, $\boldsymbol{\epsilon_b^{boat}}$, $\boldsymbol{\epsilon_c^{cell}}$, $\boldsymbol{\epsilon_{c,u}^{unit}}$ and $\boldsymbol{\beta_1}$.

All nonzero fixed effects are given non-informative Gaussian prior distributions with zero mean, similar to the parameters in the proportion-at-age model. This makes the conditional posterior a multivariate Gaussian distribution. The parameters are updated using the `GMRFLib` library. Details of the sufficient statistics involved are given in Appendix A.5.

• Sample unit effects, $\boldsymbol{\epsilon_{c,u}^{unit}}$

We have assumed a Gaussian prior for the unit effects. Hence, given all the other variables, the posterior distribution is Gaussian, and the parameters are easily updated. Details are given in Appendix A.6.

• Sample precision parameters, $\boldsymbol{\tau_{lga}^{fish}}$, $\boldsymbol{\tau_{lga}^{unit}}$, $\boldsymbol{\tau_{lga}^{cell}}$, $\boldsymbol{\tau_{lga}^{region}}$ and $\boldsymbol{\tau_{lga}^{boat}}$, and AR-parameter, $\boldsymbol{\phi_{lga,r}}$

We assume independent Gamma priors, currently Gamma(0.01,0.01), on the precision parameters. Given all fixed and random effects and with independent Gamma priors on the precision parameters, the conditional distributions for the precision parameters become independent Gamma distributions. The precision parameters τ_{lga}^{cell}, τ_{lga}^{region} and τ_{lga}^{boat} and the AR-parameter $\phi_{lga,r}$ are simulated similarly to the parameters in the age model. The conditional posterior distribution of τ_{lga}^{fish} is given by

$$
\begin{aligned}
p(\tau_{lga}^{fish}|\text{data}) \propto & p(\tau_{lga}^{fish}) \prod_{u=1}^{U} \prod_{f=1}^{N_u} p(l_{u,f}|a_{u,f}, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g) \\
\propto & (\tau_{lga}^{fish})^{N/2+0.01-1} \\
& \cdot \exp\left\{ - \left(\frac{1}{2} \sum_{u=1}^{U} \sum_{f=1}^{N_u} [l_{u,f} - \beta_{0,u} - \beta_1 g(a_{u,f}; \boldsymbol{\theta}_g)]^2 + 0.01 \right) \tau_{lga}^{fish} \right\}.
\end{aligned}
$$

The terms involved here can be calculated using the sufficient statistics, ssq, described in Appendix A.5.

3.1.5 Sample parameters for haulsize effect

If haulsize is included in the model, the separate model is estimated before doing the MCMC simulation from the age and length-given-age model. This model is similar to the weight-given-length model, except that the slope is zero. Hence, the same simulation algorithm is used here. The mcmc samples from the simulation are stored, and are later used when estimating the age model.

3.2 Simulation of weight-given-length model

This part of the model only involves linear regression parameters and random effects in addition to precision parameters, and is simulated similarly to the length-given-age model.

The fixed and random effects and the linear regression parameters are simulated using the `GMRFLib` library. Since this part is not affected by missing ages, the sufficient statistics can be calculated without need for recalculation. Hence, the calculations for each iteration are fast.

The precision parameters have independent Gamma priors, so the conditional distributions are also independent Gamma and are easily updated.

This part of the model is typically not affected by missing ages or classification errors. Simulation of θ_w can therefore be performed independently of the other parts of the model.

4 Estimating catch-at-age

The main procedure for predicting the catch-at-age is based on two simple relations:

- The number of fish, T, is equal to the total weight, W, divided by the average weight $\bar{w}$.
- The number of fish in an age group a, $T(a)$, is equal to the number of fish multiplied by the proportion of fish in age group a, $p(a)$.

The total catch is known through fisheries reports, and is given in weight for each cell, W_c. The number of fish is usually so large that average weight can be replaced by expected weight. Let $\mathrm{E}_c[w]$ denote the expected weight of fish caught in cell c. Then, the number at age in a cell is given as

$$T_c(a) \propto \frac{W_c}{\mathrm{E}_c[w]}\mathrm{E}_c[p(a)].$$

We can also estimate expected length-at-age, $\mathrm{E}_c[l|a]$, and expected weight-at-age, $\mathrm{E}_c[w|a]$.

Given samples $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_S^*$ from the posterior distribution (3), samples of catch-at-age are obtained by

$$T_{s,c}^*(a) = \frac{W_c}{\mathrm{E}_c[w|\boldsymbol{\theta}_s^*]}\mathrm{E}_c[p(a|\boldsymbol{\theta}_s^*)].$$

These samples can be used to construct predictions together with credibility intervals.

The cells for which we want to predict the catch-at-age, can be a combination of cells that were used in the fitting, for which we have estimated values for the different effects, and new cells. For new cells, the values of the fixed effects are zero. We simulate values for the unobserved cell effects, and we assume there are no unobserved areas. For the age model we include the unit effect by simulation over a given number of units. For the length-given-age model and weight-given-length model, the unit effect is included in the variance, and not when calculating the intercept.

Details on estimating $\mathrm{E}_c[p(a)]$ and $\mathrm{E}_c[w]$ are given in Appendix A.7.

5 Program structure

The program is available in an R-package `caa`.

The model is implemented in C. Recent developments have resulted in a C-program that can be run separately, where all the input and output are through binary files. The R interface (see Hirst et al. (2016)) handles all the input data and parameter selections, and creates the correct input files to the C-program. It also post-processes the resulting binary files and creates summary tables and different plots.

The C-code depends on several libraries (`taucs`, `metis`, `blas` and `lapack`). However, the latest version now uses a static version of these libraries, so there is no need for extra installation of these libraries.

The C-code consists of three main routines; estimation of the age model and length-given-age model, estimation of the weight-given-length model and the prediction routine.

The C-code is also documented in `http://files.nr.no/samba/eca/`.

References

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):192–236.

Dey, D., Ghosh, S., and Mallick, B. (2000). *Generalized linear models: a Bayesian perspective*. CRC Press.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. (1995). *Bayesian Data analysis*. Chapman & Hall, New York.

Gilks, W., Richardson, S., and Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. London: Chapman & Hall.

Hirst, D., Aanes, S., and Rognebakke, H. (2016). User manual for eca – version 4.0. *NR Note SAMBA/53/16*.

Hirst, D., 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. *Applied Statistics*, 53(1):1–14.

Hirst, D., Storvik, G., Aldrin, M. Aanes, S., and Huseby, R. B. (2005). Estimating catch-at-age by combining data from different sources. *Canadian J. of Fisheries and Aquatic Sciences*, 62(6):1377–1385.

Hirst, D., Storvik, G., Rognebakke, H., Aldrin, M., Aanes, S., and Vølstad, J. H. (2012). A Bayesian modelling framework for the estimation of catch-at-age of commercially harvested fish species. *Canadian J. of Fisheries and Aquatic Sciences*, 69(12):2064–2076.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, pages 597–614.

Roberts, G. and Sahu, S. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 291–317.

Rognebakke, H., Hirst, D., Storvik, G., and Aldrin, M. (2011). Catch-at-age for multiple stocks: Modelling skrei and coastal cod simultaneously. *NR Note SAMBA/46/11*.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman & Hall/CRC.

A Appendix

A.1 Conditional simulation of a Gaussian Markov random field (GMRF)

We briefly discuss the case where we want to sample from a GMRF under an additional linear constraint. Let $\boldsymbol{x}$ be a vector of length n of all the effects to be simulated. Hence, we want to sample from $p(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0})$, where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$ and $\boldsymbol{A}$ is a $k \times n$ matrix of constraints. If $k \ll n$, samples can be produced by first sampling from the unconstrained GMRF $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}^{-1})$, and then correct for the constraints by computing

$$\boldsymbol{x}^* = \boldsymbol{x} - \boldsymbol{Q}^{-1}\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\boldsymbol{x}.$$

Then $\boldsymbol{x}^*$ has the correct conditional distribution. Note that $\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^T$ is a dense $k \times k$ matrix, and its factorization is fast to compute for small k (Rue and Held, 2005, Chapter 2).

A.2 Sample unit effects in age model

A Metropolis-Hastings algorithm is used to sample unit effects in the age model. Given all the other parameters, dependency only occur within units for α_u^a, making it possible to perform simulation on units independently. The multivariate Gaussian proposal distribution is constructed by searching for the conditional mode running a few Newton-Raphson iterations. Let $\alpha_u^{a,opt}$ denote the alpha-values at the mode. The Hessian matrix H at the mode is also calculated. Proposed values for α_u^a, denoted $\tilde{\alpha}_u^a$, are constructed from

$$\begin{aligned} \epsilon &= (L^T)^{-1}\tilde{\epsilon} \\ \tilde{\alpha}_u^a &= \alpha_u^{a,opt} + \epsilon, \end{aligned}$$

where $H = LL^T$ and $\tilde{\epsilon} \sim \mathcal{N}(0, 1)$. Within unit, the parameters have the sum constrained to be zero. Hence, only proposals for the first $A - 1$ variables are constructed, and the last is calculated from these.

The proposed values for all age groups are accepted with probability $\exp(\log l_{new} - \log l_{old})$, where

$$\log l_{new} = \sum_a N_u(a) \log \frac{\exp \tilde{\alpha}_u^a}{\sum_{a'} \exp \tilde{\alpha}_u^{a'}} - \sum_a (\tilde{\alpha}_u^a - \mu_u^a)^2 + \frac{1}{2}(\tilde{\epsilon}^a)^2,$$

and similarily for $\log l_{old}$. Further, μ_u^a is the sum of the right hand side of (1), except the unit effect $\zeta_u^{unit,a}$.

A.3 Sample unit precision in age model

Assume $\boldsymbol{x} \in \mathcal{R}^n$ and $\boldsymbol{x} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{S})$. Assume we have available $\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ where $\boldsymbol{A}$ is of dimension $n_c \times n$. For the unit precision in the age model, we have $n = U \times A$ and $n_c = U$. Now

$$\begin{aligned}
\log[p(\boldsymbol{x}|\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0})] =& \log[p(\boldsymbol{x})] - \log[p(\boldsymbol{A}\boldsymbol{x})] \\
=& \text{Const.} - \frac{1}{2}\log(|\sigma^2 \boldsymbol{S}|) - \frac{1}{2\sigma^2}\boldsymbol{x}^t \boldsymbol{S}^{-1}\boldsymbol{x} + \\
& \frac{1}{2}\log(|\sigma^2 \boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T|) + \frac{1}{2\sigma^2}\boldsymbol{x}^t \boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T]^{-1}\boldsymbol{A}\boldsymbol{x} \\
=& \text{Const.} - \frac{n - n_c}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{x}^t[\boldsymbol{S}^{-1} - \boldsymbol{A}^T[\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}^T]^{-1}\boldsymbol{A}]\boldsymbol{x} \\
=& \text{Const.} - \frac{n - n_c}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\boldsymbol{x}^t \boldsymbol{S}^{-1}\boldsymbol{x}
\end{aligned}$$

where we in the last equality have used that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$.

This shows that under the constraint the only correction we need to do is to subtract n_c from n in the $\log(\sigma^2)$ term. This is similar to the REML situation.

A.4 Sample parameters in non-linear g-function

A Metropolis-Hastings algorithm is used to sample the parameters in the non-linear g-function in (2). In theory all parameters could be estimated from data. But we have fixed the parameters θ and c, and only simulate γ. We use Gamma(0.01,0.01) as the prior distribution. The algorithm for simulating γ is as follows:

- Let γ be the current value.
- Draw independent proposal γ' from

  $$\gamma' = f\gamma$$

  where f has density

  $$p(f) \propto 1 + \frac{1}{f}$$

  on the interval $[1/F, F]$, where $F > 1$ (currently 1.01). F is a "step-length" for the proposal for γ. The reason for using this is that

  $$\frac{p(\gamma|\gamma')}{p(\gamma'|\gamma)} = 1,$$

- Calculate the acceptance probability, which is $\min\{1, R\}$, where

  $$R = \frac{p(\boldsymbol{\theta}_g'|\boldsymbol{\theta}_a, \boldsymbol{\theta}_l, a, l)}{p(\boldsymbol{\theta}_g|\boldsymbol{\theta}_a, \boldsymbol{\theta}_l, a, l)} \frac{p(\boldsymbol{\theta}_g|\boldsymbol{\theta}_g')}{p(\boldsymbol{\theta}_g'|\boldsymbol{\theta}_g)} = \frac{p(l|a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g')p(\boldsymbol{\theta}_g')}{p(l|a, \boldsymbol{\theta}_l, \boldsymbol{\theta}_g)p(\boldsymbol{\theta}_g)}.$$

We have

$$\prod_{u,f} p(l_{u,f}|a_{u,f},\boldsymbol{\theta}_l,\boldsymbol{\theta}_g) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{u,f}[l_{u,f}-\beta_0-\beta_1 g(a;\boldsymbol{\theta}_g)]^2\right\}$$
$$= \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\left[N\beta_1^2 g(a;\boldsymbol{\theta}_g)^2 - 2\beta_1\left(\sum_{u,f} l_{u,f} - N\beta_0\right) g(a;\boldsymbol{\theta}_g)\right.\right.$$
$$\left.\left. + \sum_{u,f}(l_{u,f}^2) - 2\beta_0\sum_{u,f} l + N\beta_0^2\right]\right\},$$

where N is the total number of fish. This means that only a few sufficient observators must be calculated. The acceptance probability is then found from

$$\begin{aligned}\log R = &-\frac{\tau}{2}\left[N\beta_1^2 g(a;\boldsymbol{\theta}_g')^2 - 2\beta_1\left(\sum_{u,f} l_{u,f} - N\beta_0\right) g(a;\boldsymbol{\theta}_g')\right.\\ &\left. -N\beta_1^2 g(a;\boldsymbol{\theta}_g)^2 - 2\beta_1\left(\sum_{u,f} l_{u,f} - N\beta_0\right) g(a;\boldsymbol{\theta}_g)\right]\\ &+ (\alpha_\gamma - 1)(\log\gamma' - \log\gamma) - \beta_\gamma(\gamma'-\gamma).\end{aligned}$$

Because each step is computationally fast compared to updating the other variables involved, a fixed number of iterations (currently 1000) are performed, giving in practice samples from the conditional posterior.

A.5 Sample fixed and random effects in length-given-age model

In this section we will first show how to sample fixed and random effects in the length-given-age model when we don't include haulsize in the model. The unit effects are sampled separately, and described in Appendix A.6.

Let $\boldsymbol{x} = [\beta^{const}, \beta_y^{year}, \beta_s^{season}, \beta_g^{gear}, \epsilon_r^{region}, \epsilon_c^{cell}, \beta_1]^T$. The conditional posterior distribution can be written as

$$\begin{aligned}p(\boldsymbol{x}|\text{data}) \propto& p(\boldsymbol{x})p(\text{data}|\boldsymbol{x})\\ \propto& p(\boldsymbol{x})\exp\left\{-\frac{1}{2}\tau_{lga}^{fish}\sum_{u=1}^{U}\sum_{f=1}^{N_u}[l_{u,f}-\beta_{0,u}-\beta_1 g(a_{u,f};\theta_g)]^2\right\}\\ \propto& \exp\left\{-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right\}.\end{aligned}$$

This is a multivariate Gaussian distribution with precision matrix $\boldsymbol{Q}$ and mean $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}$.

We can write

$$
\begin{aligned}
&\sum_{u,f}[l_{u,f}-\beta_{0,u}-\beta_{1,u}g(a_{u,f};\theta_g)]^2\\
&=\sum_{u,f}\left[l_{u,f}-\hat\beta_{0,u}-\hat\beta_{1,u}g(a_{u,f};\theta_g)+\hat\beta_{0,u}-\beta_{0,u}+\hat\beta_{1,u}g(a_{u,f};\theta_g)-\beta_{1,u}g(a_{u,f};\theta_g)\right]^2\\
&=\sum_{u}\sum_{f=1}^{N_u}\left[l_{u,f}-\hat\beta_{0,u}-\hat\beta_{1,u}g(a_{u,f};\theta_g)\right]^2\\
&\quad+\sum_{u}\sum_{a}N_u(a)(\hat\beta_{0,u}-\beta_{0,u})^2+\sum_{u}\sum_{a}N_u(a)g(a;\theta_g)^2(\hat\beta_{1,u}-\beta_{1,u})^2\\
&\quad+\sum_{u}2\sum_{a}N_u(a)g(a;\theta_g)(\hat\beta_{0,u}-\beta_{0,u})(\hat\beta_{1,u}-\beta_{1,u}),
\end{aligned}
$$

where $N_u(a)$ is the number of fish at age a in unit u. Further, $\hat\beta_{0,u}$ and $\hat\beta_{1,u}$ are the least squares estimates from the data, given by

$$
\begin{aligned}
\hat\beta_{0,u} &= \frac{\sum_{f=1}^{N_u}l_{u,f}-\hat\beta_{1,u}\sum_{a=1}^{A}N_u(a)g(a;\theta_g)}{N_u}\\
\hat\beta_{1,u} &= \frac{\sum_{f=1}^{N_u}l_{u,f}\cdot\sum_{a=1}^{A}N_u(a)g(a;\theta_g)-N_u\sum_{a=1}^{A}[g(a;\theta_g)\sum_{f=1}^{N_u(a)}l_{u,f}]}{(\sum_{a=1}^{A}N_u(a)g(a;\theta_g))^2-N_u\sum_{a=1}^{A}g(a;\theta_g)^2},
\end{aligned}
$$

where $N_u=\sum_a N_u(a)$. Then, we find that the precision matrix is given by

$$
\boldsymbol{Q}_u=\begin{bmatrix}
\tau_{lga}^{const}+\tau_{lga}^{fish}N_u & \tau_{lga}^{fish}N_u & \dots & \tau_{lga}^{fish}N_u & \tau_{lga}^{fish}\sum_a N_h(a)g(a;\theta_g)\\
& \tau_{lga}^{year}+\tau_{lga}^{fish}N_u & & & \\
& & \ddots & & \\
& & & \tau_{lga}^{cell}+\tau_{lga}^{fish}N_u & \tau_{lga}^{slope}+\tau_{lga}^{fish}\sum_a N_u(a)g(a;\theta_g)^2
\end{bmatrix},
$$

and

$$
\boldsymbol{b}_u=\tau_{lga}^{fish}\begin{bmatrix}
\hat\beta_{0,h}N_u+\hat\beta_{1,u}\sum_a N_h(a)g(a;\theta_g)\\
\vdots\\
\hat\beta_{0,h}N_u+\hat\beta_{1,u}\sum_a N_h(a)g(a;\theta_g)\\
\hat\beta_{0,h}\sum_a N_h(a)g(a;\theta_g)+\hat\beta_{1,u}\sum_a N_u(a)g(a;\theta_g)^2
\end{bmatrix}.
$$

In addition we have that

$$
\begin{aligned}
ssq&=\sum_{f=1}^{N_u}\left[l_{u,f}-\hat\beta_{0,u}-\hat\beta_{1,u}g(a_{u,f};\theta_g)\right]^2\\
&=\sum_{f=1}^{N_u}l_{u,f}^2+\hat\beta_{0,u}^2N_u+\hat\beta_{1,u}^2\sum_{a=1}^{A}N_u(a)g(a;\theta_g)^2\\
&\quad-2\left(\hat\beta_{0,u}\sum_{f=1}^{N_u}l_{u,f}+\hat\beta_{1,u}\sum_{a=1}^{A}[g(a;\theta_g)\sum_{f=1}^{N_u(a)}l_{u,f}]-\hat\beta_{0,u}\hat\beta_{1,u}\sum_{a=1}^{A}N_u(a)g(a;\theta_g)\right),
\end{aligned}
$$

which is needed when sampling the observation precision, τ_{lga}^{fish}.

If we include haulsize in the model, the likelihood contribution is calculated from

$$\sum_{u,f}\left[l_{u,f}-\tilde{\beta}_{0,u}-\beta_{1,u}g(a_{u,f};\theta_g)-\beta_u^{hsz}x_u^{hsz}\right]^2,$$

where $\tilde{\beta}_{0,u}$ is the intercepth without the haulsize term. Similarily to above, we can find that the precision matrix and $\boldsymbol{b}$-vector are given by

$$\boldsymbol{Q}_u=\begin{bmatrix} \tau_{lga}^{const}+\tau_{lga}^{fish}N_u & \dots & \tau_{lga}^{f}\sum_a N_u(a)g(a;\theta_g) & \tau_{lga}^{fish}N_u x_u^{hsz} \\ & & & \tau_{lga}^{fish}N_u x_u^{hsz} \\ & \ddots & & \\ & & \tau_{lga}^{slope}+\tau_{lga}^{fish}\sum_a N_u(a)g(a;\theta_g)^2 & \tau_{lga}^{fish}\sum_a N_u(a)g(a;\theta_g)x_u^{hsz} \\ & & & \tau_{lga}^{hsz}+\tau_{lga}^{fish}N_u x_u^{hsz\,2} \end{bmatrix},$$

and

$$\boldsymbol{b}_u=\tau_{lga}^{fish}\begin{bmatrix} \hat{\beta}_{0,u}N_u+\hat{\beta}_{1,u}\sum_a N_u(a)g(a;\theta_g) \\ \vdots \\ \hat{\beta}_{0,u}N_u+\hat{\beta}_{1,u}\sum_a N_u(a)g(a;\theta_g) \\ \hat{\beta}_{0,u}\sum_a N_u(a)g(a;\theta_g)+\hat{\beta}_{1,u}\sum_a N_u(a)g(a;\theta_g)^2 \\ \hat{\beta}_{0,u}N_u x_u^{hsz}+\hat{\beta}_{1,u}\sum_a N_u(a)g(a;\theta_g)x_u^{hsz} \end{bmatrix}.$$

The last column and row corresponds to the haulsize effect. The rest of the matrix and vector is the same as before.

A.6 Sample unit effects in length-given-age model

The length-given-age model can be written as

$$l_{u,f}=\tilde{\beta}_{0,u}+\epsilon_{c,u}^{unit}+\beta_1 g(a_{u,f};\theta_g)+\epsilon_{u,f}^{fish},$$

where $\tilde{\beta}_{0,u}$ contains all the terms in the intercept except $\epsilon_{c,u}^{unit}$. We have assumed a Gaussian prior for the unit effects. Hence, given all the other variables, the posterior distribution is Gaussian, and the parameters are easily updated. We have

$$\sum_{u=1}^{U}\sum_{f=1}^{N_u}\left[l_{u,f}-(\tilde{\beta}_{0,h}+\epsilon_{c,u}^{unit}+\beta_1 g(a_{u,f};\theta_g))\right]^2 =$$

$$\sum_u N_u\epsilon_{c,u}^{unit\,2}-2\sum_u\epsilon_{c,u}^{unit\,2}\left[N_u(\hat{\beta}_{0,u}-\beta_{0,u})+\sum_a N_u(a)g(a)(\hat{\beta}_1-\beta_1)\right]+const,$$

where $N_u(a)$ is the number of fish at age a in unit u, and $N_u=\sum_a N_u(a)$.

The precision matrix and the b-vector in the canonical representation become

$$\begin{aligned} Q &= \tau_{lga}^{unit}+\tau^{fish}N_u \\ b &= \tau_{lga}^{fish}\left[N_u(\hat{\beta}_{0,u}-\beta_{0,u})+\sum_a N_u(a)g(a)(\hat{\beta}_1-\beta_1)\right]. \end{aligned}$$

A.7 Estimating catch-at-age

In order to calculate the number at age in a cell we need to estimate $\mathrm{E}_c[p(a)]$ and $\mathrm{E}_c[w]$.

First we estimate $\mathrm{E}_c[p(a)]$ from Monte Carlo simulation over a given number of units, where the number of units, $N_{MC,c}$, can be different for each cell. In each unit we do the following:

- First, we sample the precision terms

$$\begin{aligned}\epsilon_u^{age} &\sim \mathcal{N}(0, \tau_{age}^{boat\ -1} + \tau_{age}^{unit\ -1}),\\ \epsilon_u^{lga} &\sim \mathcal{N}(0, \tau_{lga}^{unit\ -1}),\\ \nu_u^{wgl} &\sim \mathcal{N}(0, \tau_{wgl}^{unit\ -1}).\end{aligned}$$

- If haulsize is included in the model we also sample $\epsilon_u^{haul} \sim \mathcal{N}(0, \tau_{hsz}^{haul\ -1})$ and calculate new haulsize using the values from the estimation for the other covariates

$$\log hsz_u = \delta^{const} + \delta_y^{year} + \delta_s^{season} + \delta_g^{gear} + \nu_r^{region} + \nu_c^{cell} + \epsilon_u^{haul}.$$

- Then we calculate $p_u(a)$ using values from the estimation, and the simulated precision terms

$$\begin{aligned}\alpha_u^a =& \alpha^{const,a} + \alpha_y^{year,a} + \alpha_s^{season,a} + \alpha_g^{gear,a} + \zeta_r^{region,a} + \zeta_c^{cell,a}\\ &+ \alpha_u^{hsz,a}\epsilon_u^{haul} + \epsilon_u^{age},\end{aligned}$$

  The terms $\zeta_c^{cell,a}$ can be missing and are then simulated. The resulting estimate of $p_u(a)$ is then

$$p_u(a) = \frac{\exp(\alpha_u^a)}{\sum_{a'}\exp(\alpha_u^{a'})}$$

Then, in a given cell we have

$$\begin{aligned}\mathrm{E}_c[p(a)] =& \frac{1}{\sum_{u=1}^{N_{MC,c}} hsz_u}\sum_{u=1}^{N_{MC,c}} p_u(a) hsz_u\\ \mathrm{E}_c[l|a] =& \exp\left\{\beta_{0,u} + \beta_1 g(a;\theta_g) + \frac{1}{2}\left(\frac{1}{\tau_{lga}^{fish}} + \frac{1}{\tau_{lga}^{unit}}\right)\right\}\\ \mathrm{E}_c[w|a] =& \exp\left\{A + Bg(a;\theta_g) + \frac{1}{2}\sigma^2\right\}\\ P_c(l|a) =& \Phi(l^u|\mu_l(a), \tau_{lga}^{fish}) - \Phi(l^l|\mu_l(a), \tau_{lga}^{fish}); \qquad l \in (l^l, l^u)\\ \mathrm{E}[l]_c =& \sum_a \mathrm{E}_c[p(a)]\mathrm{E}_c[l|a]\\ \mathrm{E}[w]_c =& \sum_a \mathrm{E}_c[p(a)]\mathrm{E}_c[w|a],\end{aligned}$$

where

$$
\begin{aligned}
A =& \delta_{0,u} + \delta_1 \beta_{0,u} \\
B =& \delta_1 \beta_1 \\
\sigma^2 =& \delta_1^2 \left(\frac{1}{\tau_{lga}^{fish}} + \frac{1}{\tau_{lga}^{unit}} \right) + \left(\frac{1}{\tau_{wgl}^{fish}} + \frac{1}{\tau_{wgl}^{unit}} \right) \\
\mu_l(a) =& \beta_{0,u} + \beta_1 g(a; \theta_g).
\end{aligned}
$$

The catch-at-age in cell c for a given age and length interval is given by

$$T_c(a,l) = \frac{W_c}{\mathrm{E}[w]_c} E_c[p(a)] P_c(l|a).$$

Hence, the total for all cells become

$$
\begin{aligned}
T(a,l) &= \sum_c T_c(a,l) \\
\mathrm{E}[l|a] &= \frac{\sum_c (E_c[l|a] \sum_l T_c(a,l))}{\sum_l T(a,l)} \\
\mathrm{E}[w|a] &= \frac{\sum_c (E_c[w|a] \sum_l T_c(a,l))}{\sum_l T(a,l)}.
\end{aligned}
$$