ICES CM 2009/N:02

A Bayesian hierarchical approach to estimating landings- and discards-at-age.

David Hirst, Hanne Rognebakke and Geir Storvik, Norwegian Computing Centre.

Contact author David Hirst: Strathweltie, Tarland,Aboyne AB344YS, tel +44 13398 81448. Email david.hirst@nr.no

Keywords: Bayesian hierarchical model, COST, catch-at-age

Abstract

A Bayesian modeling approach has been implemented in the COST project to estimate landings and discards at age, along with length and weight at age. The method takes the various different kinds of data commonly available from sampling of commercial fisheries, ie length only samples, length and age samples and length stratified age samples, both for discards and landings, and combines them into a single analysis. Because length-given-age is modeled explicitly, there is no need to create age-length keys, and modeling of the covariates means that empty “cells” (ie combinations of covariates) do not present a problem. The hierarchical nature of the analysis allows data at different stratifications (hauls or trips) to be included, and the uncertainty is fully described in the posterior distributions of the parameters. The modeling approach also allows virtual populations to be simulated, which mean that the results can be compared with known true values and a realistic assessment of the performance made.

Introduction

A modeling approach is taken to estimate landings and discards at age, along with weight and length at age. The model is implemented in a program written as part of the COST project. It can analyse data from most kinds of sampling programs within ICES. The model includes submodels for proportion at age and length-given age, including in both cases the covariates season, gear and area. This approach means that there is no need to create an age-length-key, and so no need to collect age data in every length class, and empty “cells” (combinations of the covariates) do not present a problem. Discards and landings can be included in the same analysis, making possible an estimate of the total landings at age, with the correct uncertainty. The approach is outlines in this paper, but full details will be published elsewhere.

Types of data

The data that can be included in the analysis are any combination of the following:
Random length samples
Random length and age samples
Length stratified age samples

Weight samples can be included with any kind of sampling so long as the length is also measured.

The data can be stratified at either trip or haul level, or both.
Covariates can include gear, season, area and year. They must be recorded for all data if they are to be included in the model.

The only data that cannot be used are age samples collected for an age-length-key, where either the origin (ie sampling unit) of the samples is not known, or there is no corresponding length distribution. This kind of data is unfortunately quite common in some parts of Europe, and it is not obvious that is possible to make a correct analysis of it using any method at all.

Details of the models

The models are fitted using MCMC. Their details are as follows:

1) The model for proportion at age

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

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

$$\alpha_{c,h}^{a} = \alpha^{base,a} + \alpha_{y(c)}^{year,a} + \alpha_{s(c)}^{season,a} + \alpha_{g(c)}^{gear,a} + \varsigma_{r(c)}^{region,a} + \varsigma_{c}^{cell,a} + \varsigma_{c,h}^{trip,a}.$$

In the above *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 $\alpha_{y}^{year,a}$ etc.

The α terms and $\varsigma_{r}^{region,a}$ are the main effects for year, season, gear and region. The α terms are fixed effects and $\varsigma_{r}^{region,a}$ 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 $\varsigma_{r}^{region,a}$ follows a Gaussian conditional autoregressive distribution (CAR) (e.g. Carlin and Louis, 1996). It is assumed that there will always be some data for all levels of the fixed effects that are of interest,

though not for all combinations of the covariates. The $\varsigma_c^{cell,a}$ terms are independent random effects modelling the interactions between the main effects (see e.g. 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 $\varsigma_c^{cell,a}$ terms. The differences between trips within a cell are modelled by the random effects $\varsigma_{c,h}^{trip,a}$. These must be random (rather than fixed) effects because there are many cells and trips with no data. We assume that all the interactions (ie the $\varsigma_c^{cell,a}$ terms) can be modelled by a single distribution.

2) The models for length-given-age and weight-given-length

Length given age is modelled using a log-linear or Schnute-Richards relationship:

$$\log(length_{c,h,f}) = \beta_{0,c,h} + \beta_1 \log(age_{c,h,f}) + \beta_2 hz_h + \varepsilon_{c,h,f}^{fish}$$

or

$$\log(length_{c,h,f}) = \beta_{0,c,h} + \beta_1 g(age_{c,h,f}) + \beta_2 hz_h + \varepsilon_{c,h,f}^{fish}$$

where

$$g(age) = \log\left(1 + \theta \exp\left(-\gamma \exp\left(c \log\left(age\right)\right)\right)\right)$$

and θ, γ and c are parameters to be estimated. The function is standardised to lie between 0 and 1. Note that age here is a continuous variable, ie age in years and months or years and seasons. The age in the age model is the year class.

The β_0 parameters are modelled in a similar way to the α parameters.

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

The slope β_1 is common to all cells and trips.

The weight-given-length model is similar to the log-linear length-given-age model.

$$\log(weight_{c,h,f}) = \delta_{0,c,h} + \delta_1 \log(length_{c,h,f}) + \nu_{c,h,f}^{fish}$$

Here $length_{c,h,f}$ is the length of the f^{th} fish from trip h in cell c, $weight_{c,h,f}$ its weight and $age_{c,h,f}$ its age. $\varepsilon_{c,h,f}^{fish}$ and $\nu_{c,h,f}^{fish}$ are independent zero mean Gaussian random variables. ε_r^{region} and ν_r^{region} are CAR parameters with similar properties to $\varsigma_r^{region,a}$ in the age model (see appendix 1). ε_c^{cell} and ν_c^{cell} are random 'all interactions' effects equivalent to $\varsigma_c^{cell,a}$. ε_c^{haul} and ν_c^{trip} are between trip random terms. The β and δ terms are fixed effects similar to the α terms in the model for proportions at age.

3) The discard function

We assume that a fish is landed (i.e. not discarded) with a probability depending only on its length:

$$p(landed \mid length) = k_h \Phi\left(m_h(\log(length) - r_h)\right)$$

Here Φ is the cumulative Gaussian distribution function. Further, r_h is a parameter specifying the mean log(length) at which fish are discarded, m_h describes the rate of change in the probability around r_h while k_h is a factor taking into account the possibility that large fish might be discarded. We allow the discard function parameters to vary with trip.

Size class and haul sizes

We assume the sampling unit is the trip, and this is usually the case for age data, but the length data are often collected from a sub-unit, eg the haul or the size class. The length data are then combined and stratified to collect the age data. It would be much more satisfactory if the age data were collected from the same sampling units as the length data, and since age-length-keys are not necessary in the modelling approach this only involves noting which length sample the age data came from. However, in the absence of this information our approach is to resample the length data to create a new sample of the same size as the (estimate of) the total stratum size. For example, if there are only 3 hauls, with 100, 300 and 500 fish in each, and length samples of size 10 have been taken from each, then we resample the first length sample up to 100 fish, the second up to 300 and the third up to 500. These samples are then combined and assumed to be similar to the true combined length sample. This resampling is done at every step of the MCMC algorithm, but the uncertainty in the true haul or size class size is ignored.

Catch-at-age

Once the parameters of the model have been estimated from the data, it is possible to estimate the landings and discards at age. Note that these are not parameters in the model so they are not estimated directly. Instead, given the posterior distributions of the parameters in the model, we can simulate many trips in each cell, and calculate the catch-at-age from the simulated trips.

The method is as follows:

1) Fit the model to obtain the joint posterior distribution of all the parameters in the age, length-given-age and weight-given-length models, and the discard function model.
2) Simulate a large number of trips in each cell using one set of parameters.
3) Find the mean length-given-age and weight-given-age for the simulated trips.
4) Find the numbers at age for landings and discards, and scale by the desired raising factor, eg if the numbers are to be raised to total weight of catch, find the weight of the simulated fish and scale accordingly.
5) Go back to step 2 and repeat for a different set of parameters from the joint posterior.

This results in a posterior distribution of the fishery parameters of landings-at-age etc.

Simulations

Given the parameters from a fitted model it is possible to simulate data sets using any desired sampling strategy. This enables both comparison of strategies, and comparison of estimation methods. An example is given below.

Example of estimation by different methods:
As an illustration of what can potentially be done, 20 small data sets were simulated and the three methods implemented in COST tested on them. The data sets consisted of the following:
20 trips, from each of which one length sample of size 100 was taken from each size class on each trip. One age sample was taken from each length class in the combined length sample for each trip. Thus a full ALK can be constructed for each trip.
The only covariate is season (at 4 levels), and there is at least one trip sampled from each season.

The results for one typical simulation are shown in figure 1. The true value of the landings at age is shown in red, and 90% intervals for these values created by the three different methods are also shown. The overall means (over the 20 simulations) are shown in figure 2. The mean widths of the 90% intervals are shown in figure 3, and the coverage of the intervals in figure 4. These simulations are intended as an illustration of what can be done with the model and programs in COST, and it is not possible to draw any meaningful conclusion from them. It will be possible in the future to do a more comprehensive simulation study and to fully investigate the properties of the three methods.

Conclusions

A model has been developed which can combine most kinds of data to get estimates of landings and discards at age. It removes the need for age-length-keys and the problem of empty cells. It can also estimate discards and landings simultaneously. It can also be used to simulate data in order to test both alternative estimation methods, and different sampling schemes.

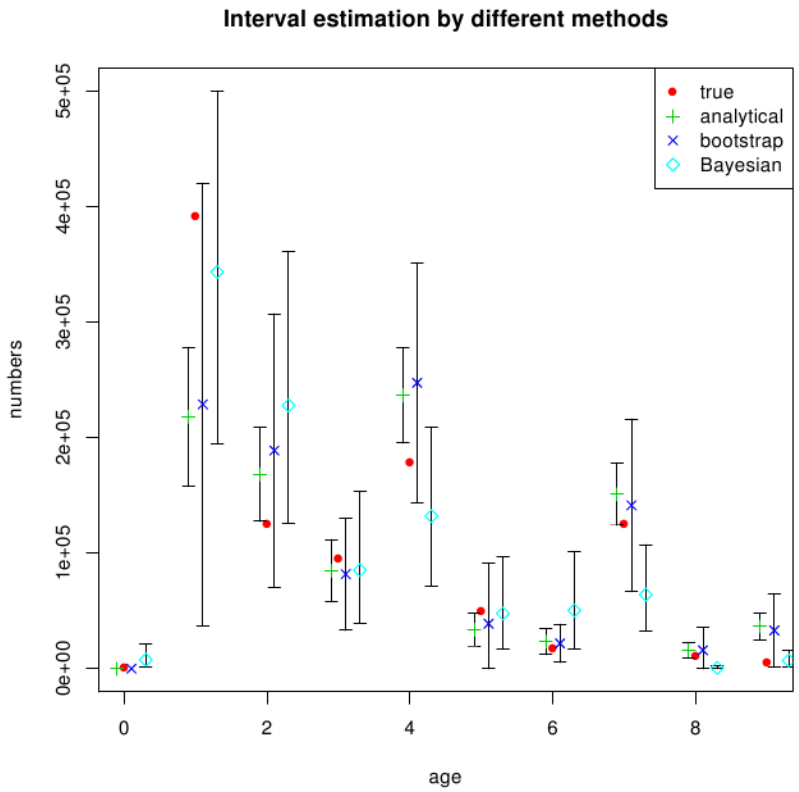

Figure 1. Interval estimation by three different methods on a simulated data set with known true landings-at-age.

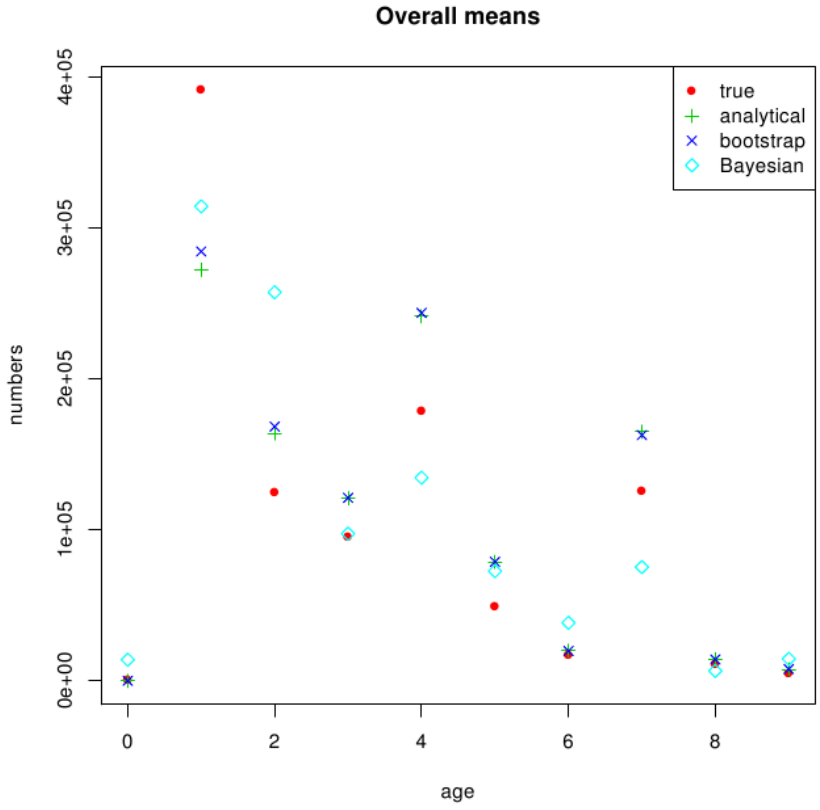

Figure 2. Estimates of landings at age averaged over 20 simulations for the three different methods.

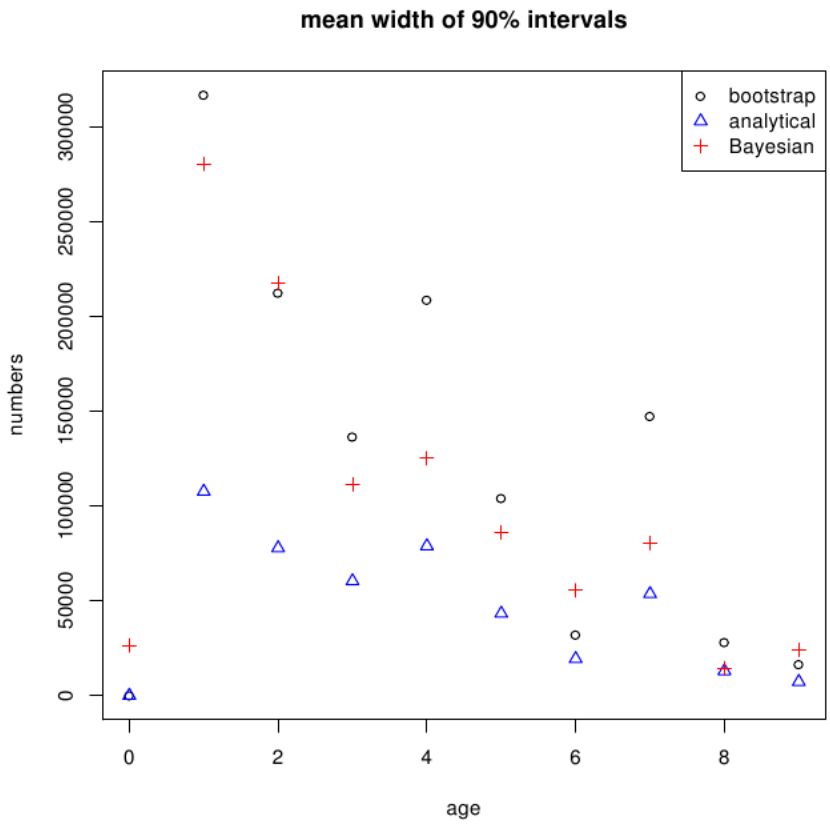

Figure 3. Mean width of 90% intervals produced by the three methods.

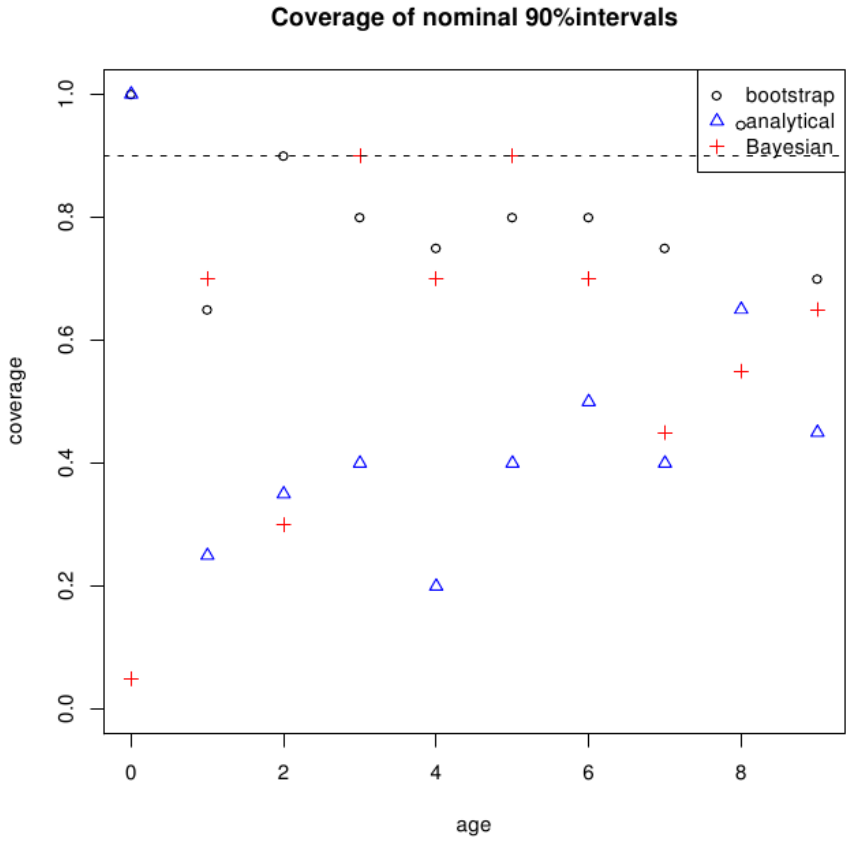

Figure 4. Coverage of the 90% intervals over 20 simulations.