CM 2001/P:05

Estimating catch-at-age from market sampling data using a Bayesian hierarchical model

David Hirst, Ingunn Fride Tvete, Geir Storvik,
Norwegian Computing Center, Oslo
and Sondre Aanes,
Institute for Marine Research, Bergen.

Introduction

The Norwegian Institute for Marine Research has the task of estimating the total catch-at-age of cod by the commercial fishing fleet in the Barents Sea. The procedure is the following: A boat (the 'Amigo') sails from port to port along the north Norwegian coast over a period of about 6 weeks, 4 times a year (roughly corresponding to the 4 seasons). At each port it takes a sample of about 100 fish from whichever boats are available at the time. The fish are weighed, the length measured, and the otoliths extracted. These otoliths are then used to estimate the age of the fish. Each year about 300 boats, and thus about 30,000 fish are sampled. Note that the program only samples landings, though we refer to catch-at-age in this paper.

The total catch (in weight) is known, and is available for each year and season, for each area and gear. There are a large number of areas, though only about 6 have significant catches. In total over the 5 years from 1996 to 2000 16 areas were sampled. There are 5 gear types that have significant catches. The Amigo program aims to provide an estimate of the proportion of the catch at each age, and a mean weight per fish, for each combination of gear, season and area (a 'cell'). These are then raised to total catches using the mean weight and the total catch for that cell.

Clearly not every cell is sampled, and where a cell is missing, the age distribution is estimated by an ad-hoc procedure involving finding a 'similar' cell that has been sampled. This is a time consuming and somewhat unreliable method. The uncertainty in the estimates has not often been addressed, though a bootstrapping approach has been used on occasion. The Amigo samples boats very approximately in proportion to the number fishing in each cell, though since boats are necessarily only sampled when they are available, it is inevitable that some cells are missed, and some are over-sampled. This makes the bootstrap extremely difficult or impossible to use properly.

The aim of this paper is to introduce a modelling strategy for the data which can provide reliable estimates of the total catch-at-age, for whichever cells or combination of cells are required. We also want to provide a realistic measure of the uncertainty in our estimates. We approach this problem by using a Bayesian hierarchical model.

The modelling approach

Assumptions:
(1) Boats are sampled randomly within a cell.
(2) Fish are sampled randomly from a boat.
(3) Age is measured without error.
(4) The total catch is given without error.

The model for proportion-at-age:

1) The fish on a boat are assumed to be drawn from a multinomial distribution:

$X_j \sim \text{multinomial}(p_{j.}, n_j)$

Here j references the boat. Xj is the vector of numbers at age sampled from boat j. In our case we use the following 9 age groups; less than 4, single ages from 4 to 10, and greater than 10. Therefore X_j is a vector of length 9. p_j is the vector of probabilities for these age groups, on boat j. Since we assume the sample is taken randomly from the boat, p_j is equal to the proportion of the total catch on boat j in each age group. n_j is the sample size, usually about 100. It is assumed that this is determined in advance, or at least that it is independent of the age and weight distribution of the fish on the boat.

2) The p_j are assumed to vary from boat to boat, even within the same cell. We are not directly interested in the values from the boats that are actually sampled, since they comprise a very small proportion of the total fishing effort. Rather we are interested in the underlying population of 'catchable fish'. We therefore assume that the p_j are themselves random variables, drawn from the population of interest.

If p_{jk} is the element of p_j corresponding to the k^{th} age class, then p_{jk} must be positive and sum to 1 over k, so we can write:

$$p_{jk} = \frac{\exp(\gamma_{jk})}{\sum_{k=1}^{9} \exp(\gamma_{jk})}$$

We regard γ_{jk} as itself a random variable, ie there is a boat effect:

$$\gamma_{jk} \sim N(\mu_{jk}, \sigma^2)$$

The μ_{jk} are the underlying population parameters, functions of location, gear, year and season. It is these parameters that we are interested in. They are modelled as follows:

$$\mu_{jk} = \alpha_k + \beta_{s(j)k} + \chi_{g(j)k} + \delta_{y(j)k} + \varepsilon_{l(j)k}$$

Here $s(j)$ means the season corresponding to the j^{th} boat. For each age k we have an overall mean, α_k, a season effect, β_{sk}, a gear effect χ_{gk} ,a year effect δ_{yk}, and a spatial effect ε_{lk}. Note that the effects are additive on this scale, but multiplicative on the

probability scale. We have not fitted any interactions between the effects, since they were not justified by the data, though in principle they could be included.

The prior distribution for the spatial term is modelled as a Gaussian conditional autoregressive (CAR) variable.

$$\varepsilon_{k.} \sim \tau^{R/2} \exp\left(-\frac{\tau}{2} \sum_{i} \sum_{b \in Neigh[i]} (\varepsilon_{ki} - \varepsilon_{kb})^2 \right)$$

$$R = \dim(\varepsilon_{k.})$$

Here $\varepsilon_{k.}$ is the vector of effects for age k, for all regions.

This model implies that the mean for each area is a Gaussian variable with variance τ^2, and mean equal to the mean of the ε_k over all of its neighbours. This model enables us to estimate the age distribution for any area, even if there are no samples. Clearly the estimate will have less uncertainty if there are samples however.

We require the following standardisation for identifiability:
$\alpha_1 = \beta_{1k} = \beta_{s1} = \chi_{1k} = \chi_{g1} = \delta_{1k} = \delta_{y1} = 0$

All α, β, χ and δ (except those defined as zero) are given independent vague Gaussian priors.

The model for weight-given-age:

In order to raise proportion at age to total catch at age, it is necessary to estimate the mean weight per fish in each cell (since the total catch is given as a total weight). This could be done directly by modelling the mean weight of fish on each boat, but since weight is highly dependent on age, it is more efficient to model weight-given-age, and to combine this model with that for the proportion at age.

We assume that log(weight) is linear in log(age), with the slope and intercept of the regression depending on gear, season and year. In principle it would depend on location but this effect was very small and was therefore excluded from the model. A regression was done separately for each boat, and the slope and intercept were then modelled over all boats, ie we assume that on boat j
$\log(weight_{ij}) = intercept_j + slope_j \left(\log(age_{ij}) - 2\right) + \varepsilon_{ij}$.
Here $weight_{ij}$ is the weight of the i^{th} fish on boat j, and ε_{ij} is a normal random variable. The variance of ε_{ij} is assumed to be constant over all boats within all cells. 2 is subtracted from $\log(age_{ij})$ as this is approximately the mean of the log(age) over all fish sampled, and thus subtracting it removes most of the correlation between the intercept and slope parameters.

The slope and intercept parameters are now modelled in a similar way to the γ parameters in the model for the age distribution, ie we assume that the values on a given boat are randomly drawn from a distribution for the appropriate cell, and this distribution is modelled.

$$slope_j \sim N(\omega_j, \upsilon^2)$$

ω_j is the underlying population parameter, a function of gear, year and season, but not in this case location. It is modelled as follows:

$$\omega_j = \alpha + \beta_{s(j)} + \chi_{g(j)} + \delta_{y(j)}$$

The mean (α), season (β_s), gear (χ_g) and year (δ_y) effects are given independent vague Normal priors, and the first level for season, gear and year is set to zero for identifiability.

Simulation

Using the BUGS software (Spiegelhalter et al, 1996) we are able to estimate the parameters in the above models for age distribution and weight-given-age, along with their uncertainty. We then simulate from these parameter distributions (assuming they are Gaussian), in order to simulate age distributions for each cell. Note that we are simulating at the underlying population level and not at the boat level. Thus we are simulating the 'catchable' population, which is assumed to be very similar to the fish actually caught. For each simulation of an age distribution, we then simulate a slope and intercept for the weight given age, and thus simulate a mean weight. (Note that since we have modelled log(weight), it is necessary to adjust the mean by a parameter dependent on the variance of the residuals from the original regressions). Using the total catch for that cell, we can thus get a simulation of the total catch-at-age. This is repeated a large number of times and the mean and standard deviation over the simulations are found.

Results

The estimates of proportion at age are shown in Figure 1 for 8 cells where there is sufficient data to make a comparison. The dotted lines are the observed proportions on each of around 10 boats. The bars show the means plus and minus 2 standard deviations for the estimates from our model. The means show a very similar pattern to the data. The uncertainty in the estimates is much less than the variability in the data. This is to be expected, because the data includes 2 extra sources of variation – the multinomial variability in the sample from the boat, and the boat variability within the cell. Also our estimates use all the available information (including that from different cells).

The estimates of the total catch-at-age for one of the main fishing regions and for all regions combined are plotted in Figure 2, for all 5 years. Again the bars show the mean plus and minus 2 standard deviations. As would be expected, the uncertainty is much less for the combined areas than for an individual area.

Discussion

The modelling approach taken in this paper allows the estimation of catch-at-age in a coherent and repeatable way. It also gives a realistic measure of uncertainty. The estimation can be done for cells even where there are no samples. In addition is it reasonably fast. The BUGS program is the slowest part of the procedure, but on a reasonably powerful pc it can easily be run overnight. Where it is possible to compare the results with the data, they agree very well. We believe that this kind of modelling could be very beneficial for stock assessment both in Norway and elsewhere. The Norwegian data have some advantages for this kind of modelling, because we have been able to assume that samples are randomly taken from the boats. This means the likelihood is very simple. In other countries there is often stratification by size class. This makes the likelihood more complicated, especially if the size class is determined by length rather than weight. This may necessitate joint modelling of age, weight and length. This is however only a technical problem, which should be possible to overcome.

This kind of modelling could also be used to investigate the effectiveness of different sampling schemes. It is possible to simulate samples from any number of boats, in any combination of cells, and to estimate parameters and parameter uncertainty from the simulations. This would allow optimisation (or at least improvement) of sampling strategies. The difficulty with this kind of work at present is that the BUGS program would have to be run a large number of times in different scenarios, which would be extremely time consuming. However, there is no technical reason why a specialised program could not be written in C++ for example.

References:

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W (1996), BUGS 0.5: {Bayesian} inference using {Gibbs} sampling, manual (version ii), MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR.
http://www.iph.cam.ac.uk/bugs/

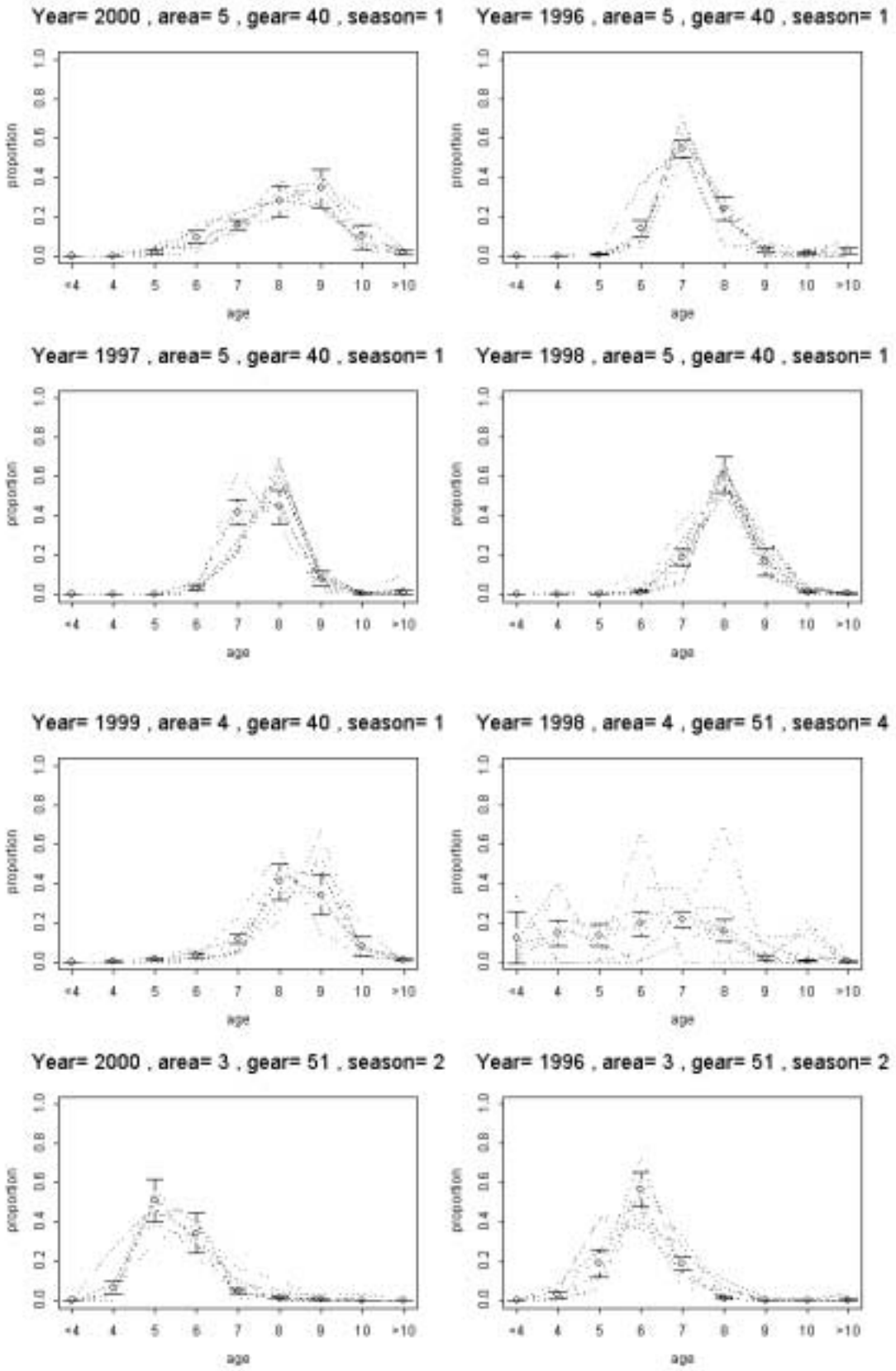

Figure 1. Estimated and observed age distributions for 8 cells. Dotted lines are observed proportions at age for all boats sampled in that cell, bars are mean estimates plus and minus 2 standard deviations.

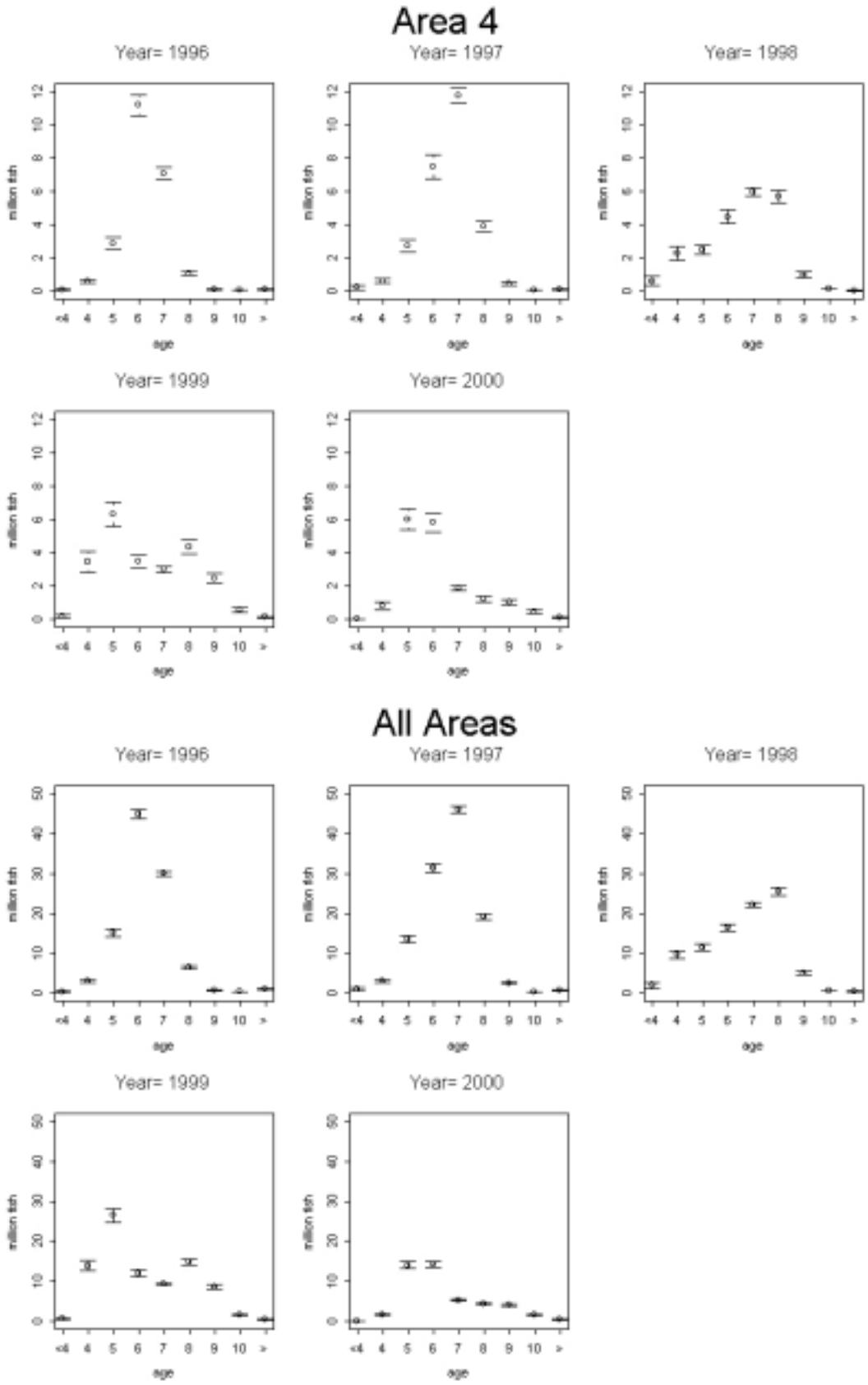

Figure 2. Estimated total catch at age, for area 4 (top) and all areas combined (bottom). The bars show means plus and minus 2 standard deviations.