Northwest Atlantic Fisheries Organization

Serial No. N4155 **NAFO SCR Doc. 99/83**

SCIENTIFIC COUNCIL MEETING – SEPTEMBER 1999
(Joint NAFO/ICES/PICES Symposium on Pandalid Shrimp Fisheries)

Use of Subjective Prediction in Optimal Stratified Sampling with Application to Shrimp Surveys in the Barents Sea

by

Alf Harbitz

Norwegian Institute of Fisheries and Aquaculture Ltd., 9291 Tromsø, Norway

Abstract

We consider stratified sampling and the task of applying subjective knowledge in predicting the number of trawl samples per stratum that minimises the cv of the abundance estimator. The constraint is a given vessel time available. It is assumed that the strata biomass means, arbitrarily scaled, are the only unknown parameters needed to find the optimal solution. The concept of a subjective prediction distribution of the unknown stratum means is introduced. The distribution is person-dependent and is determined based on intervals [L,U] for the minimum and maximum subjectively predicted biomass values compared with the true measured values found after the predictions. The approach assumes a constant subjective confidence level defined as the probability of covering the true value in a random interval. A pilot subjective prediction experiment was conducted during the 1998 shrimp survey in the Barents Sea. Based on 62 [L,U] predictions of shrimp biomass in the next trawl haul combined with the true biomass, the subjective prediction distribution for the cruise leader was estimated. The distribution was applied to her stratum predictions for the next survey. 10000 random predictions of true strata means were simulated from the distribution. For each simulation cv-values of the abundance estimator were estimated based on relative strata means predicted from historical data as well as the subjective predictions. A significant cv-reduction was obtained based on a combination of subjective prediction and historical data, compared to the use of historical data alone.

Kew words: subjective prediction, optimal stratified sampling, adaptive sampling, Bayesian analysis.

Introduction

To get a most reliable stock assessment of shrimp resources in large areas, effective sampling strategies are of vital importance due to limited vessel time. We consider the case where the study area is divided in predetermined geographical areas (strata). The problem studied is to predict the number of samples in each stratum that minimises the cv of the abundance estimator of biomass in the entire study area, combining historical data and subjective prediction. In each stratum we assume that the trawl samples are independent observations which provide unbiased estimates of the unknown sample stratum mean and variance. The terms stratum mean and variance are here used for the expected biomass and the variance of the biomass from a random trawl haul in the stratum.

If the cost of providing a sample is neglected, a well-known result in stratified sampling is that the optimal number of samples in a stratum is proportional to stratum area and stratum standard deviation (Cochran 1977, Thompson 1992). This solution is often denoted Neyman allocation, after Neyman (1934). For trawl surveys in large areas, the stratum-dependent sailing time between trawl stations is the main cost. In this case the optimal solution must be found numerically, as shown in Harbitz *et al.* (1998). It turns out, however, that the optimal solution in this case deviates negligibly from the Neyman allocation unless there are extreme relative differences between strata areas and/or between stratum standard deviations (ibid). As a reaonsable approach we apply the Neyman allocation as the optimal solution in this paper.

Supported by data (Fig. 2) we assume that the stratum standard deviations are proportional to the strata means with a proportionality constant independent of strata and time (survey). In this case the sufficient, but unknown, parameters needed to determine the Neyman allocation are the relative strata means defined as the stratum mean in each stratum divided by the sum of strata means in all strata.

If the estimated relative stratum means appear to be quite similar from survey to survey, this is an indication that it is wise to predict the relative stratum means in a future survey from historical data. Note that this might be the case even if the abundance varies from survey to survey.

If scientists or others have qualified reasons to doubt that we can predict the relative strata means from past data, due to environmental changes or other reasons, the challenge faced is how to utilise such subjective knowledge to improve these predictions. A major goal of this paper is to outline a method to quantify the person-dependent ability of subjective prediction in statistical terms.

A core in our approach is to establish a link between subjective data and real measurements of biomass, based on subjective prediction experiments. Each subjective observation simply consists of a minimum, a most likely and a maximum predicted biomass value for a future parameter, e.g. the shrimp biomass in the next trawl haul or the stratum mean. A major assumption is that the subjective confidence level, i.e. the expected relative proportion of the data where the true parameter value falls within the interval limits, is a person-dependent constant independent of stratum and survey. The better subjective knowledge, the smaller interval width is assumed, maintaining the subjective confidence level.

Based on data from experiments as described above, the concept of a subjective prediction distribution of the unknown parameter is introduced, which has a frequentistic interpretation. Based on subjective predictions for the strata means in the next survey, along with historical trawl data, a combined prediction for the unknown relative strata means is constructed. The weights balancing historical data versus subjective prediction are a linear combination of the inverse variances involved. When the combined predictions are established, the corresponding predicted optimal number of trawl stations in each stratum are calculated.

In order to assess the effect of applying subjective prediction, simulations based on a case study with real historical trawl data and subjective prediction data from shrimp surveys in the Barents Sea are performed. Based on subjective predictions for a future survey, along with an established subjective distribution, the future values of true relative strata means are simulated. For each simulation, the cv of the abundance estimator is calculated based on: 1) proportional allocation (non-stratified sampling), 2) optimal sampling based on historical data, 3) optimal sampling based on historical data and subjective prediction, and 4) optimal allocation based on the true relative strata means. A comparison of the different cv-distributions from the simulations are used in order to study the effect of the 3 first approaches, which are all applicable in practice. Then they are compared to the cv-distribution obtained from optimal allocation based on true strata means.

The main focus of the paper is optimal allocation, but the concepts developed can easily be extended to a Bayesian framework for inference about the unknown abundance, by applying the subjective prediction distribution as a prior. After the data from the survey are available, we can construct the aposteriori distribution of abundance conditional on trawl data, and e.g. calculate a credibility interval of specified level for the unknown abundance. Because the prior now has a frequentistic interpretation, it is meaningful to define the bias of the predicted parameter with respect to the prior. If it is unbiased, the posterior mean will be unbiased as well. Bayesian inference (Carlin and Louis 1996) is getting increased popularity within fisheries science (McAllister and Kirkwood 1998), not least due to the accelerating development of e.g. computer-efficient Markov chain Monte Carlo techniques (Gilks 1996). This enables our approach to be attractive also in case the subjective prediction distribution is used as a non-conjugate prior.

The subjective prediction approach can also be applied in an adaptive setting (Thompson and Seber 1996), where new subjective predictions are made before each new stratum is reached, and an optimal reallocation of effort within the remaining strata are performed.

Material and Methods

Notations

i	=	subscript for stratum
m	=	number of strata
j	=	subscript for survey
n_{yr}	=	number of surveys
l	=	summation index over strata
Y	=	biomass in kg from trawl haul
$\overline{Y}$	=	stratum mean of Y-values
μ	=	stratum mean, EY
σ	=	stratum standard deviation, std(Y)
s	=	1) empirical standard deviation, 2) subscript for subjective
$[L,U]$	=	subjective biomass prediction interval
M	=	modal subjective predicted value
α	=	lower fractile of f_s corresponding to L
β	=	upper fractile of f_s corresponding to U
ε	=	interval factor in the relationship $(U–L) = \varepsilon\cdot(U+L)/2$
r	=	subscript for relative
k	=	proportionality factor in relation $\sigma = k\mu$
$\gamma(b,c)$	=	gamma distribution (pdf) with scale parameter b and shape parameter c
Q	=	chi-square sum
Z	=	standardised subjective prediction variable
f_{s0}	=	standardised subjective distribution
f_s	=	subjective prediction distribution
N	=	number of samples in a stratum
A_i	=	area of stratum i
A	=	ΣA_i = total area of all strata
A_0	=	area covered by one trawl haul
μ_A	=	true abundance in total area
h	=	subscript for historical
*	=	superscript for estimator

The survey data

The shrimp biomass data to be used are from annual trawl surveys in the Barents Sea in the period 1992-1999. The area is divided in $m = 6$ fixed geographical areas, or strata, see Fig. 1 and lower row of Tab. 1. The total number of trawl stations (biomass observations) varies between a minimum of 92 (1999) to a maximum of 139 (1996). The minimum number of observations at one single stratum is 7, while the maximum is 59, see Tab. 1. The estimated stratum means and standard deviations are given in Tab. 2. Aschan and Sunnanå (MS 1997) give a description of survey design and trawl technology.

In 1998 a pilot subjective prediction experiment was conducted where each of 4 persons reported a minimum value, L, a most probable (modal) value, M, and a maximum value, U, for the biomass they predicted in the next trawl haul. The 4 persons included two scientists (the cruise leader and another biologist), the captain and the chief mate. All persons predicted independent of each other. As a basis for their prediction the biomasses from the previous haul and the previous year were easily available. Of particular interest was the cruise leader data, because this person has the authority to determine the effort allocation. A sample size of n_s = 62 (L,M,U)-data was reported by the cruise leader, and a similar number for the other persons.

Statistical population model

The statistical population model is a major basis for most of the statistical analysis. A population in statistical sense must not be confused with the biological meaning of population. We therefore find a precise definition of statistical population model appropriate here.

Let Y denote the biomass found from a trawl haul covering a standardised area, dA, synonymous with a fixed trawled distance. We define a population as the set of all N_y possible y-values under specified conditions. The conditions considered here are location and time, i.e. stratum and time period of survey. As a consequence there will be different populations for different strata and surveys. The population distribution is then defined as the relative frequency histogram of the N_y possible y-values. Here N_y is so large that it is reasonable to model the population distribution as a continuous probability density function, pdf.

We assume that the actual sample consisting of N biomass values, one from each trawl station, is a random sample of independent observations from the population distribution. It is further assumed that the shape of the population distribution is identical to the one we would have found if the y-values could have been replaced by the corresponding true biomass of shrimp in the whole water column. These two distributions will then deviate from each other essentially through a scaling factor taking account of e.g. vertical diurnal migration and trawl catchability. The factor is assumed to be independent of stratum and survey period.

Let $f(y)$ denote the population distribution with mean μ and standard deviation σ. Supported by our data (Fig. 2) it is reasonable to assume that f belongs to the class of scale-parameter distributions, i.e. $f(y) \equiv (1/b) \cdot f_0(y/b)$. Here $f_0(\cdot)$ is a fixed function independent of b and independent of stratum and time. If all our y-values from different strata and surveys were divided by their respective stratum means, μ, the resulting values would then follow the same distribution. We therefore divide the y-values from all surveys and strata with their empirical stratum means and use this rich database in order to determine an appropriate choice of f_0 among known distributions. We also assign equal probability to each of the standardised y-values, y_0, to provide a parameter-free approximation to f_0 with application to e.g. simulation experiments.

The population model is further simplified by assuming that the population standard deviation is proportional to the population mean, $\sigma = k\mu$, where the proportionality factor k is independent of stratum and time. Our data (Fig. 2) support this assumption. We estimate k based on $(\bar{y}, s)$-values where $\bar{y}$ denotes empirical stratum mean and s denotes empirical stratum standard deviation. Two different least square estimators for k are considered. The first estimator is

$$(1) \qquad k_1^* = \frac{\sum_{j=1}^{n_{yr}} \sum_{i=1}^{m} s_{ij} \cdot N_{ij}}{\sum_{j=1}^{n_{yr}} \sum_{i=1}^{m} \bar{y}_{ij} \cdot N_{ij}}$$

which motivated by theoretical and empirical considerations is synonymous with a weight inversely proportional to $\bar{y}_{ij}$ in each term of the least square sum. The second estimator is analogous to the one above, but is based on log-transformed variables:

$$(2) \qquad k_2^* = \exp\left(\sum_{j=1}^{n_{yr}} \sum_{i=1}^{m} \left(\frac{N_{ij}}{\sum_{ij} N_{ij}} \right) \log\left(\bar{y}_{ij} / s_{ij} \right) \right)$$

The properties of the estimators for k are completely determined by the population distribution. We limit our attention to study bias(k^*) = Ek^*–k, and the standard deviation of k^*. Note that though S^2 is an unbiased estimator for σ^2, S is in general not an unbiased estimator for σ.

The study of standard deviation is based on bootstrapping (Efron and Tibshirani 1993) from the standardised *y*-value database as well as simulations from the parametric pdf fit to these *y*-values, while we use the latter approach to study bias. In both cases we choose the $n_{yr}\cdot m$ estimated $\bar{y}_{ij}$ -values as "true" values of Q_{ij}. The procedure for bootstrapping n_{boot} resamples of the $n_{yr}\cdot m$ $(\bar{y}_{ij}, s_{ij})$ -data then goes as follows:

1. Draw N_{ij} random y_0-values (with replacement) from the y_0 database.
2. Scale the y_0-values by multiplying with $\bar{y}_{ij}$.
3. Calculate the bootstrapped mean value, $\bar{y}_{ij}^{B}$, and standard deviation, s_{ij}^{B} , of the N_{ij} values provided in 2.
4. Repeat 1-3 for each value of *i* (stratum) and *j* (survey) providing a random sample of the $n_{yr}\cdot m$ $(\bar{y}_{ij}, s_{ij})$ -data.
5. Repeat 1-4 n_{boot} times.
6.

Based on each of the n_{boot} samples we calculate the corresponding bootstrapped values k_1^{*B} and k_2^{*B} for k_1^{*} and k_2^{*} , respectively, and then estimate std(k^*) for both estimators.

For the study of bias we apply the gamma distribution $\gamma(1/c,c)$ with scale parameter $1/c$ and shape parameter $c = 1/k^2$, for the standardised biomasses Y/μ, which corresponds to a $\gamma(\mu/c,c)$ population distribution. In this case the strata means become gamma-distributed as well:

$$(3) \qquad Y \sim g\,(m/c, c) \Rightarrow \bar{Y} \sim g\,(m/(cn), cn)\,,$$

which mathematically is a very convenient result for simulations. We use an estimator k^* for k based on data in order to estimate the value for c to be used in the simulations. This time we do not need to sample from the population distribution, but we sample random $\bar{Y}_{ij}$ -values directly from the distribution above. Based on the known k-value in the simulations, we estimate the bias based on e.g. $n_{\mathrm{MC}} = 1000$ Monte Carlo simulations of k. We also, as before, estimate the standard deviation of k^*. If both estimators appear to be approximately unbiased, the one with the lowest estimated standard deviation is preferred.

The optimal number, N_i, of trawl stations per stratum

We apply the results in Harbitz et al. (1998) in order to determine the number, N_i, of trawl stations in each of m predetermined strata that minimise the cv of the abundance estimator for the entire area under the constraint of a given vessel time, t, available. This is synonymous with minimising the variance:

$$(4) \qquad \mathrm{var}\left(m_A^*\right) = \sum_{i=1}^{m} s_i^{\,2} \cdot \left(A_i / A_0\right)^2 / N_i\ ,$$

where $s_i^{\,2}$ is the variance of biomass of a random trawl haus in stratum i (stratum variance), A_i is the corresponding area, and A_0 is the area covered by a trawl station. The general problem in stratified sampling is that the stratum variances, σ_i^2, are not known. We assume (*ibid*) that the optimal solution deviate negligibly (*ibid*) from the Neyman-allocation $N_i \propto A_i\sigma_i$. Applying the model $\sigma = k\mu$, the optimal solution then depends on the relative stratum means, $m_{ri} = m_i / \sum m_l,\ i = 1,\mathrm{K}\ \ ,m$, as the only unknown parameters:

$$(5) \qquad N_i = \left(\frac{\sum A_i \sqrt{m_{rl}}}{2vt_0 \sum A_l m_{rl}}\right)^2 \cdot \left(\sqrt{1 + \frac{4v^2 tt_0 \sum m_{rl} A_l}{\left(\sum A_l \sqrt{m_{rl}}\right)^2}} - 1\right)^2 \cdot m_{ri} A_i\,,\ \ i = 1,\mathrm{K}\ \ ,m$$

where the sums are over strata, v is the sailing speed between successive trawl stations and t_0 is the trawl and handling time at each station. v and t_0 are assumed to be constants independent of strata, and the trawl stations within each stratum are assumed to be located at the points of intersection in a square grid. As a measure of precision we apply the coefficient of variation, cv(μ_A^*):

$$(6) \qquad cv\left(\mu_A^*\right) = \frac{std\left(\mu_A^*\right)}{\mu_A}$$

where $\mu_A = \Sigma\mu_i \cdot (A_i/A_0)$. We use the equation above combined with simulations to study the effect of different approaches to predict the μ_{ri}-values, to be compared with the minimum cv-value obtained by using true values for the μ_{ri}'s.

The relative stratum mean vector, $\underline{\mu}_r = [\mu_1,\ldots,\mu_m]/\Sigma\mu_l$.

As noted previously we assume that it is sufficient to know the relative proportion of strata means pr. stratum, $\mu_{ri} = \mu_i/\Sigma\mu_l$, in order to determine an optimal number of trawl stations within each stratum. Let $\underline{\mu}_r = [\mu_{r1},\ldots,\mu_{rm}]$ denote what we call the relative stratum mean vector, with estimator $\underline{\mu}_r^*$. The main focus in this paper is to outline how we may construct reasonable predictors, $\underline{\mu}_r^*$, for $\underline{\mu}_r$ based on a combination of historical data and subjective prediction.

It is generally complicated to determine the multivariate distribution of $\underline{\mu}_r^*$, even in the mathematical convenient case of the gamma distribution as a population distribution. An exception occurs when all strata have the same area, $A_1 = \ldots = A_m$, and N_i is proportional to μ_i. In this case $\underline{\mu}_r^*$ follows a Dirichlet distribution (Aitchison 1986).

We do not aim at finding an appropriate $\underline{\mu}_r^*$-distribution here, but restrict our attention to estimate $\mathrm{var}(\mu_{ri}^*)$, $i = 1,\ldots,m$. The following approximate expression is applied:

$$(7) \qquad \mathrm{var}\left(\mu_{ri}^*\right) \approx \mu_{ri}^2 \cdot \mathrm{var}\left(\frac{\overline{Y}_i}{\mu_i} - \frac{\Sigma \overline{Y}_l}{\Sigma \mu_l}\right),\ i = 1,\ldots,m,$$

where the sums are over strata. Based on the gamma population distribution the variance above is estimated by

$$(8) \qquad \mathrm{var}^*\left(\mu_{ri}^*\right) = \mu_{ri}^{*2} \cdot \frac{k^2}{N_i} \cdot \left(1 - 2\mu_{ri}^* + N_i \cdot \Sigma \mu_{rl}^{*2} / N_l\right),\ \ i = 1,\ldots,m,$$

where the sums are over strata. We now face the question of utilising historical data to predict $\underline{\mu}_r$ in a future survey. Let us first assume as a working hypothesis that $\underline{\mu}_r$ has not changed during the period for which we have data:

$$(9) \qquad H_0\text{: } \underline{\mu}_{rj} = \underline{\mu}_{rh},\ \ j = 1,\ldots,n_{yr},$$

where $\underline{\mu}_{rh}$ is constant with subscript h denoting "historical". We assume that the number of observations is so large that

$$(10) \qquad Q_j = \sum_{i=1}^{m} \frac{\left(\mu_{rij}^* - \mu_{rih}\right)^2}{\mathrm{var}\left(\mu_{rij}^*\right)},\ \ j = 1,\ldots,n_{yr}$$

is approximately chi-square distributed with about m degrees of freedom. It is not straightforward to determine the appropriate number of degrees of freedom in this distribution, because the μ_{ri}^*-values are negatively correlated due to the restriction $\Sigma\mu_{ri}^* = 1$. We simulate the chi-square approximation based on our assumption of a gamma population distribution, analogous to the simulation of the k^*-distribution.

In order to test H_0 we first use an "ad hoc" procedure in order to estimate $\underline{\mu}_{rh}$, which weights each stratum mean inversely proportional to the total number, N_j, of observations from survey j:

$$(11) \qquad \mu^*_{rih} = \sum_j \frac{(1/N_j)\cdot \mu^*_{rij}}{\sum_j (1/N_j)}, \quad i = 1,\ldots,m,$$

where the sums are over surveys. To test for possible differences between years we apply the test statistic Q in eq.(10) replacing μ_{rih} with μ^*_{rih} and perform one test per survey. H_0 is rejected if Q is larger than e.g. the upper .05/m fractile in the chi2(m)-distribution with for any of the years. If H_0 is rejected a more detailed analysis is required providing another predictor for $\underline{\mu}_r$, e.g. $\underline{\mu}_r^*$ from the last survey(s).

The subjective prediction distribution for μ, $f_s(\mu)$

In order to illustrate concepts and terminology, we restrict the attention to an unknown stratum expectation value μ to be predicted based on subjective prediction. In order to quantify the prediction ability of a given person, the following simple and practical measures (data) are used: the minimum, L, the most probable (modal), M, and the maximum, U, predicted values for μ are reported. A basic assumption for the subjective prediction distribution to be defined, $f_s(\mu)$, is that L and U correspond to specific, person-dependent, fractiles in the distribution. These fractiles are denoted $\alpha_L = \alpha$ and $\alpha_U = 1 - \beta$ and are estimated based on subjective prediction experiments to be described below. We define $\alpha_U - \alpha_L = 1 - \alpha - \beta$ to be the corresponding *subjective confidence level*, which is assumed to be rather constant for one and the same person.

Imagine now that we have a large number of historical (L,M,U)-values (data) for μ with corresponding unbiased estimates μ^* based on trawl data. We let the proportion of cases where $\mu^* < L$ and $\mu^* > U$ be the estimators for α and β, respectively. Let Z be a standardised variable defined as follows:

$$(12) \qquad Z = \frac{\mu^* - (L+U)/2}{(U-L)},$$

i.e., the difference between the unbiased estimator based on data, μ^*, and the mid point in the subjective prediction interval for μ. We assume that Z will follow a standardised subjective distribution, $f_{s0}(z)$, dependent on person, but independent of strata and survey (time).

The reason for choosing the mid value before the modal value, M, in the definition of Z is that the mid value is believed to be a more reasonable estimate for the mean value than M. When f_{s0} is established we may in principle calculate the mean, E(Z), and the variance, var(Z). If $\alpha = \beta$, a positive mean indicates that the mid value of the subjective prediction interval tends to underestimate the true μ-value and vice versa. The bias is considered negligible if the magnitude | E(Z) | is small compared to std(Z).

In practical life it will take a long time before sufficient data are available to determine an appropriate $f_{s0}(z)$-distribution. To overcome this, we may perform experiments providing subjective prediction values (L,M,U) for each single observation, y, i.e. the true biomass in the next trawl haul. The assumption is then that the $f_{s0}(y)$ -distribution established this way is approximately equal to the f_{s0}-distribution for prediction of μ.

Note that there is in general no one-to-one correspondence between Z and a given [L,U]-interval, in principal many different [L,U]-values may correspond to the same z-value. Motivated by data, however, we assume that there is a proportional relationship between the mid value (L+U)/2 and interval width, U–L.

$$(13) \qquad U - L = e \cdot (L+U)/2$$

where ε is the interval factor. Also motivated by data we assume that $\alpha = \beta$ and that f_{s0} equals a normal distribution $N(0,\sigma_z)$. In this case the subjective distribution for μ, $f_s(\mu)$, which is deduced directly from the definition of Z and $f_{s0}(z)$, is normal with expectation value $(L+U)/2$ and variance

$$(14) \qquad \sigma_s^2 = \mathrm{var}_s(\mu) = \left(\frac{\varepsilon(U+L)/2}{2\Phi^{-1}(\beta)} \right)^2$$

where an appropriate estimator σ_s^{2*} for σ_s^2 is found by replacing ε and β by estimators based on experimental data. Note that for a given interval $[L,U]$ it is now meaningful to consider $f_s(\mu)$ as a probability density function of μ in a classical frequentistic sense. We may imagine a long series of predictions for which $[L,U]$ is constant, where the unknown μ-values behave as independent stochastic variables, following the subjective prediction distribution conditional on $[L,U]$.

The subjective relative prediction vector is

$$(15) \qquad \underline{\mu}_{rs} = [\mu_{rs1},\ldots,\mu_{rsm}] = [\mu_{s1},\ldots,\mu_{sm}]/\sum \mu_{sl}\,,$$

where μ_{si} is the mid value of the subjective prediction interval $[L_i,U_i]$ for μ_i. We assume that the subjective prediction intervals for different strata are uncorrelated. Analogous to eq.(7) we apply the following approximate expression in order to estimate $\mathrm{var}(\mu_{ris})$:

$$(16) \qquad \sigma_{ris}^{2*} = \mathrm{var}^*(\mu_{ris}) = \mu_{ris}^{*2} \cdot \left(\frac{\sigma_{is}^{2*}}{\mu_{is}^{*2}} - 2 \cdot \frac{\sigma_{is}^{2*}}{\mu_{is}^{*} \sum_l \mu_{ls}^{*}} + \frac{\sum_l \sigma_{ls}^{2*}}{\left(\sum_l \mu_{ls}^{*}\right)^2} \right), \; i = 1,\ldots,m$$

where the sums are over strata.

The predicted optimal number of trawl stations in each stratum

We now outline how we predict $\underline{\mu}_r$ by a combination of historical data (index h, henceforth) and subjective prediction (index s, henceforth). The elements in the combined relative prediction vector $[\mu_{r1C}^*,\ldots,\mu_{rmC}^*]/\sum \mu_{riC}^*$ are determined as a linear combination of μ_{ris}^* and μ_{rih}^*, $i = 1,\ldots,m$. The weights are based on the involved variance estimates, i.e.

$$(17) \qquad \mu_{riC}^* = K \cdot \left(\frac{\sigma_{ris}^{2*}}{\sigma_{ris}^{2*} + \sigma_{rih}^{2*}} \cdot \mu_{rih}^* + \frac{\sigma_{rih}^{2*}}{\sigma_{ris}^{2*} + \sigma_{rih}^{2*}} \cdot \mu_{ris}^* \right), \; i = 1,\ldots,m\,,$$

where K is a normalising constant to ensure that $\sum \mu_{riC}^* = 1$ and the other elements are defined previously. Once the elements of the combined relative prediction vector are determined, eq.(5) is applied with μ_{ri} replaced by μ_{riC} in order to estimate the optimal number of stations within each stratum.

Simulation of the variance-reducing effect of subjective prediction

The goal of introducing subjective prediction in the context that we have outlined is to reduce the cv of the abundance estimate, $\mathrm{cv}(\mu_A^*)$. In order to study the effect of $\underline{\mu}_{rC}$ we may perform some simulations based on case studies. We limit the situation to the case where $\underline{\mu}_r$ has changed from $\underline{\mu}_{rh}$ in the past to $\underline{\mu}_{r0}$ in the future situation where the survey is conducted. We assume that the subjective prediction distribution is normal with true mean. Further, we assume that the gamma distribution is a reasonable approach to the population distribution. We also include the unstratified situation (proportional allocation), $N_i \propto A_i$, corresponding to a relative stratum mean vector $\underline{\mu}_{rA} = [1/m, \ldots, 1/m]$. The simulation process then goes as follows:

1. Choose an appropriate stratum relative mean vector $\underline{\mu}_{rh} = [\mu_{r1h}, \ldots, \mu_{rmh}]$ based on historical data.
2. Establish $[L_i, U_i]$-data based on subjective prediction, and calculate the subjective prediction vector $\underline{\mu}_{rs} = [\mu_{r1s}, \ldots, \mu_{rms}]$. Determine the corresponding normal subjective prediction distributions $f_{si}(\mu_i)$ for each stratum i.
3. Determine the combined relative vector $\underline{\mu}_{rC}$ based on 1 and 2 and determine the predicted optimal number of stations in each stratum.
4. Simulate a random true stratum relative mean vector $\underline{\mu}_{r0}$ from the subjective prediction distributions in 2.
5. Repeat step 4 n_{boot} = 10 000 times, say, and compare the distributions of cv(μ_A^*) based on $\underline{\mu}_{rA}$, $\underline{\mu}_{rh}$, $\underline{\mu}_{rC}$ and $\underline{\mu}_{r0}$.

Results

The relationship between the empirical strata means and strata standard deviations are shown in Fig. 2, along with the fitted line $\sigma = k\mu = \mu/\sqrt{2}$. As is clearly seen, the log-log plot results in a more constant variance of S for a given value of $\bar{y}$. We also see that both estimates appear to be rather unbiased, and that k_2^* should be preferred before k_1^* due to the lower standard deviation of the former. The difference between the bootstrap and gamma distribution approaches is negligible, indicating that the gamma distribution is a reasonable population distribution model. The value $k^* = 1/\sqrt{2}$ is chosen for mathematical convenience, because it appeared to be close to the estimated values for k, see Tab. 1. This corresponds to a gamma population distribution $\gamma(\mu/2,2)$ and a $\gamma(1/2,2)$ distribution for the standardised biomass-values Y/μ. The latter is shown in Fig. 3 as a fit to a histogram based on 861 standardised observations $y_{ij}/\bar{y}_{ij}$ from the surveys. As is seen, a reasonably good fit is obtained, though the large histogram peak at the leftmost bin is noteworthy.

The histogram of 10 000 simulations of Q (see eq.(10)) based on the gamma population distribution $\gamma(\mu/2,2)$ is shown in Fig. 2 with a fitted chi-square distribution with 6 degrees of freedom. As is seen, an apparently good fit is obtained. The fixed number of strata observations and strata means used in each simulation are representative for one survey, see Tab. 1 and 2. The empirical estimates of the 95, 99 and 99.5 percentiles in the q-distribution, along with the corresponding percentiles in the chi2(6)-distribution (in parenthesis) were: 13.2 (12.6), 18.3 (16.8) and 20.5 (18.5). The q-distribution thus seems to have a somewhat "heavier" tail than accounted for by the chi2(6)-distribution.

The hypothesis of a constant $\underline{\mu}_{rh}$-value in the period 1992-1999 estimated by applying eq.(11) was rejected for the years 1993 and 1996. Therefore the three last years were used to estimate $\underline{\mu}_{rh}$, providing the vector

$$\text{(18)} \qquad \underline{\mu}_{rh} = [0.073, 0.175, 0.190, 0.109, 0.305, 0.148]$$

The chi-square sums Q then were 3.9, 5.2 and 5.4 for 1997, 1998 and 1999, respectively, i.e. reasonably close to the mean value 6 in the chi-square distribution with 6 degrees of freedom.

All 4 persons who participated in the subjective prediction experiment appeared to provide intervals which tended to have widths $(U - L)$ proportional to the mid value $(L + U)/2$. The cruise leader appeared to have the best result in terms of lowest interval factor ($\varepsilon^* = 0.74$) combined with the lowest subjective confidence level (66 %). When translating her $[L_i, U_i]$ intervals by replacing the M_i-values with the previous haul value, Y_{i-1}, or the last year value, $Y_{i,97}$, the subjective confidence level war reduced to 44 % and 52 %, respectively. This is a promising indication of the potential in applying subjective prediction in this case.

We now restrict the attention to the cruise leader results and the application of these. Based on her 62 (L, M, U)-values the following results were found:

1. $\alpha^* = 0.19$, $\beta^* = 0.15$
2. $U - L \approx 0.74 \cdot (L + U)/2$
3. $\bar{z} = -0.004,\ \ s_z = 0.75$
4. Negligible deviation between M-values and interval mid values.

As is seen in from point 3 above, there is a negligible bias in the z-distribution. The intervals as a function of the modal values, M, are shown in the left part of Fig. 5, along with the true y-values. In the right part of Fig. 5 the estimated standardised subjective distribution $f_{s0}(z)$ with a fitted normal distribution is shown. Because α^* is close to β^*, we let $\alpha = \beta = .17$ in the simulation of the effect of applying a normal subjective prediction distribution in a new survey. The standard deviations σ_{is} to be used are then

$$\text{(19)} \qquad \sigma_{is} = \frac{e_i(U_i + L_i)/2}{2\Phi^{-1}(0.83)} = 0.262 \cdot e_i(U_i + L_i),\ \ i = 1,\ldots,m,$$

where L_i and U_i are the subjective lower and upper borders of the predicted stratum means. Based on subjective predictions by the cruise leader the following relative mean vector was used:

$$\text{(20)} \qquad \underline{\mu}_{rs} = [0.093, \quad 0.155, \quad 0.207, \quad 0.078, \quad 0.363, \quad 0.104]$$

with corresponding ε-values 0.181, 0.208, 0.250, 0.250, 0.179 and 0.488.

Random values for the future μ-value in each stratum are simulated from the appropriate normal distribution with expectation value $(L+U)/2$ and standard deviation given by eq.(19). For each simulation the cv-values are calculated based on proportional allocation ($\underline{\mu}_{rA}$), predicted optimal allocation based on $\underline{\mu}_{rh}$, predicted optimal allocation based on a combination of historical data and subjective prediction ($\underline{\mu}_{rC}$) and optimal allocation based on true values ($\underline{\mu}_{r0}$). The results of 10 000 simulations are shown in Fig. 6. The time available for the survey was chosen to be 250 hr, the vessel speed between stations to be 12 nm/hr and the trawling and handling time at each station was set to $t_0 = 1$ hr. As we see, a considerable improvement is obtained by applying subjective prediction, despite the rather small difference between $\underline{\mu}_{rh}$ and $\underline{\mu}_{rs}$.

Discussion

A major basis for our approach to quantify the statistical properties of subjective prediction is the assumption of a relatively constant subjective confidence level. This means that a person who becomes more skilled with experience will tend to provide more narrow prediction intervals, maintaining the same subjective confidence level. Only experience over time can assess the assumption of a constant subjective confidence level.

Another basic assumption is the transferability of the standardised subjective distribution $f_{s0}(z)$ from the simple single-value prediction of Y to the prediction of the stratum mean, μ. An intuitive feeling is that if a normal subjective prediction distribution is appropriate in the single value case, the assumption of a normal distribution is also reasonable for the μ-predictions. Again, this assumption can be assessed by experiments, but a quite long time period is required. The more persons who seriously perform such experiments, however, the shorter time is needed to make this assessment.

We have for simplicity assumed independent subjective predictions between strata. If the basis for the predictions is related to e.g. migration between strata, this independence assumption is dubious. An overestimate of the emigration effect from one stratum to another will correspond to an overestimate of the immigration effect into the other. As a result, the variances based on independence become too small, overemphasising the subjective prediction ability.

At least at an early stage many will hesitate in trusting subjective prediction too much. A general approach is then to determine some limitations on maximum influence. As an example one may limit the influence so that the combined vector $\underline{\mu}_{rC}$ is not dominated by the subjective vector $\underline{\mu}_{rs}$. One way to obtain this is to scale the relative variances σ_{ris}^2 by a common factor so that none of them become smaller than the corresponding historical variances, σ_{rih}^2. In this way the relative subjective variances between strata are maintained.

It is possible to extend the use of subjective prediction as outlined here to adaptive sampling designs. One approach would be to make new subjective predictions before the first trawl haul in each new stratum, to be used in order to determine the optimal number of trawl stations in the next strata. As more strata are finished, the new biomass values can be implemented in the estimate of the historical predictor $\underline{\mu}_{rh}$. Note that this approach is not in conflict with obtaining unbiased estimates of abundance.

The concept of a subjective prediction distribution can also be extended to inference of the abundance, μ_A. Let $[L_A\ , U_A]$ be an unbiased subjective prediction interval for μ_A consistent with corresponding intervals for each stratum. The subjective prediction distribution $f_s(\mu_A)$ can then be used as a prior in Bayesian sense, and we may use e.g. a 95 % posterior credibility interval for μ_A as an inference measure for μ_A. Because it is reasonable to assume that the distribution of μ_A^* based on data is normal, a normal prior will be a conjugate prior. The construction of a credibility interval is then straight forward. Due to the frequentistic interpretation of $f_s(\mu_A)$ it is now meaningful to say that the posterior expectation value is an unbiased estimator of μ_A.

Acknowledgements

The shrimp scientist Michaela Aschan at Norwegian Institute of Fisheries and Aquaculture Ltd. has in several ways made significant contributions. First of all the results of her subjective predictions as cruise leader have motivated the belief in the usefulness of the concepts outlined. Secondly, she has contributed through fruitful discussions and careful reading of former drafts of the manuscript. Thanks also to captain Hans Hansen and chief mate Inge Berg at R/V “Jan Mayen”, along with scientist Jorge dos Santos at Norwegian College of Fisheries, for eager participation in the pilot subjective prediction experiment in 1998.

References

Aitchison, J. 1986. The Statistical Analysis of Compositional Data. Chapman & Hall.

Aschan, M. and Sunnanå, K. MS 1997. Evaluation of the Norwegian Shrimp Surveys conducted in the Barents Sea and the Svalbard area 1980-1997. ICES CM 1997/Y:7, 24 pp.

Carlin, B.P. and Louis, T.A. 1996. Chapman & Hall.

Cochran, W.G. 1977. Sampling Techniques. 3rd edition. New York: Wiley.

Efron, B. and Tibshirani, R.J. 1993. An Introduction to the Bootstrap. Chapman & Hall, London.

Gilks, W.R. 1996. Markov chain Monte Carlo in practice. Chapman & Hall.

Harbitz, A., Aschan, M. and Sunnanå, K. 1998. Optimal effort allocation in stratified, large area trawl surveys, with application to shrimp surveys in the Barents Sea. Fisheries Research 37, 107-113.

McAllister, M.K. and Kirkwood, G.P. 1998. Bayesian stock assessment: a review and example application using the logistic model. ICES Journal of Marine Science, 55: 1031-1060.

Neyman, J. 1934. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. J. Roy. Statist. Soc. 97, 558-606.

Thompson, S.K. 1992. Sampling. Wiley, New York.

Thompson, S.K. and Seber, A.F., 1996. Adaptive Sampling. John Wiley & Sons, Inc.

Table 1. Number of samples in each stratum. Barents Sea shrimp surveys 1992-1999. Strata area in 1000 nm^2 in last row.

	Stratum					
Year	A	B	C	D	E	F
1992	16	12	10	24	33	37
1993	22	11	12	26	23	33
1994	11	10	12	22	24	22
1995	14	8	8	22	29	22
1996	11	9	8	29	59	22
1997	14	7	9	10	40	15
1998	16	8	10	17	29	29
1999	8	11	14	14	30	15
Area [1000 nm^2]	6.73	4.71	4.00	9.34	11.48	9.58

Table 2. Estimated strata means and standard deviations in kg per trawl haul. Barents Sea shrimp surveys 1992-1999.

		Stratum					
Year		A	B	C	D	E	F
1992	μ^*	21.8	37.1	56.8	28.6	65.7	28.7
	σ^*	15.9	23.0	33.5	24.4	28.9	20.0
1993	μ^*	14.5	26.2	45.2	17.0	47.7	10.8
	σ^*	12.9	13.1	33.3	12.4	26.3	6.8
1994	μ^*	19.9	11.0	13.3	8.8	30.4	13.1
	σ^*	16.6	6.5	8.5	7.8	22.0	13.5
1995	μ^*	9.7	11.5	18.0	17.6	40.8	15.1
	σ^*	6.5	9.5	20.1	19.9	29.7	14.6
1996	μ^*	23.0	10.8	32.5	24.3	68.6	28.1
	σ^*	22.7	6.3	17.8	21.9	37.0	19.2
1997	μ^*	12.5	41.0	25.4	27.4	50.7	21.3
	σ^*	12.5	36.7	10.5	26.1	47.4	14.3
1998	μ^*	16.9	39.4	53.7	23.3	81.4	58.1
	σ^*	14.7	28.9	37.0	24.0	51.5	32.9
1999	μ^*	16.0	27.4	41.1	16.2	59.3	18.0
	σ^*	10.5	21.0	26.0	15.4	41.8	13.9

Table 3. Properties of the k^* estimators for proportional constant k in the relationship $\sigma = k\mu$ based on 1000 simulations from parameter free population model (bootstrap) and from gamma distribution population model.

Method	k_1^*	k_2^*	$\text{std}^*(k_1^*)$	$\text{std}^*(k_2^*)$	$\text{bias}^*(k_1^*)$	$\text{bias}^*(k_2^*)$
Eqs. (1) and (2)	0.692	0.725				
Bootstrap-simulation			0.0247	0.0217		
Gamma-simulation			0.0252	0.0220	0.0074	–0.0021

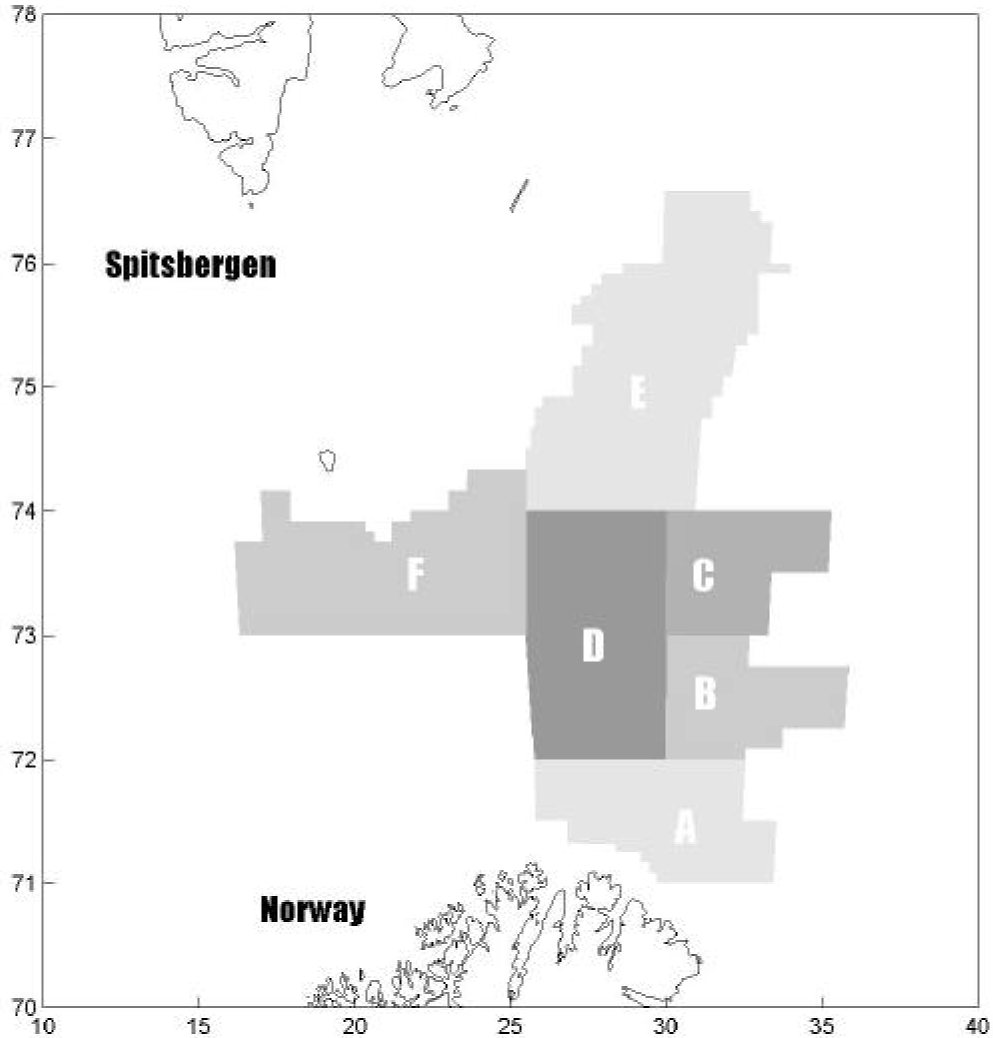

Fig. 1. The strata for the shrimp surveys in the Barents Sea 1992-1999.

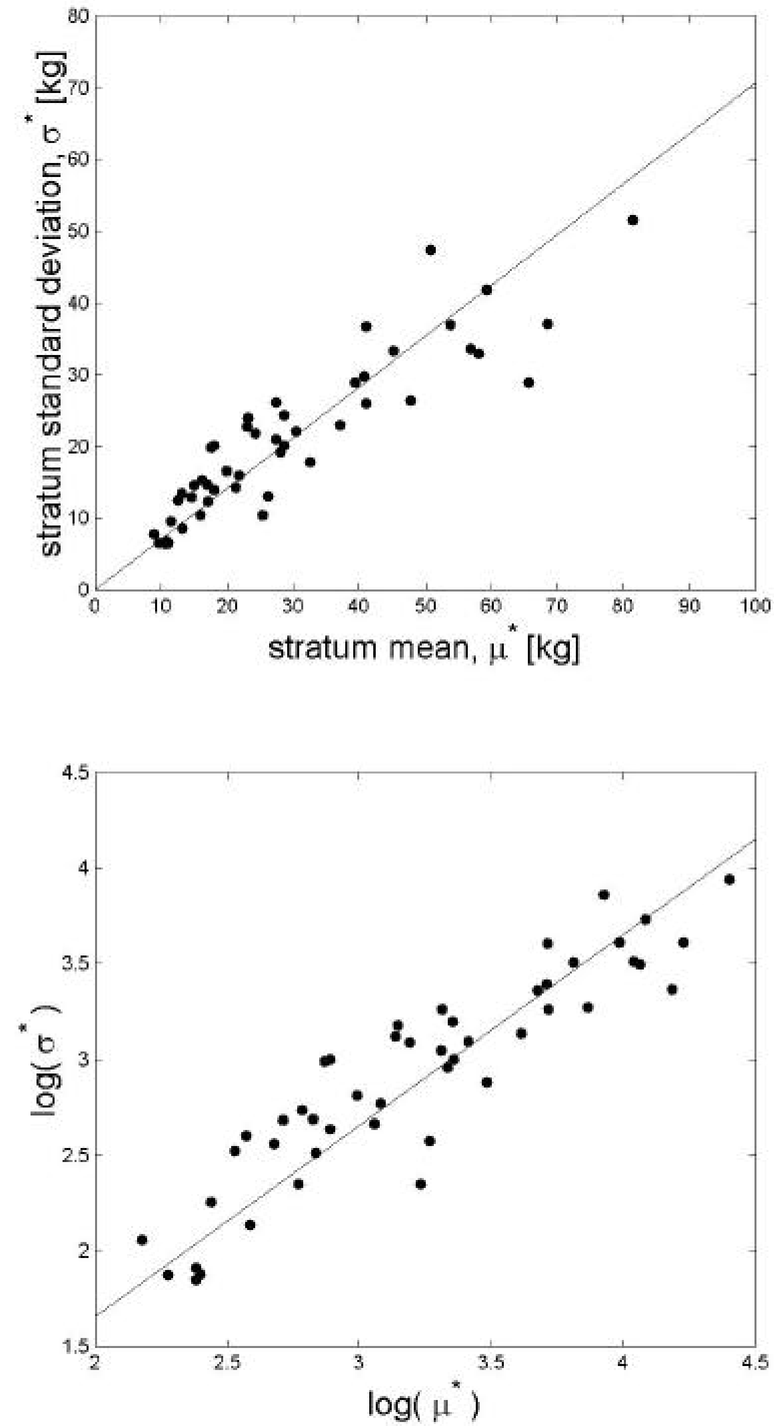

Fig. 2. The relation between empirical stratum means and standard deviations with the fitted curve ♦ ◘❍/ $\sqrt{2}$ from the shrimp surveys in the Barents Sea 1992-1999.

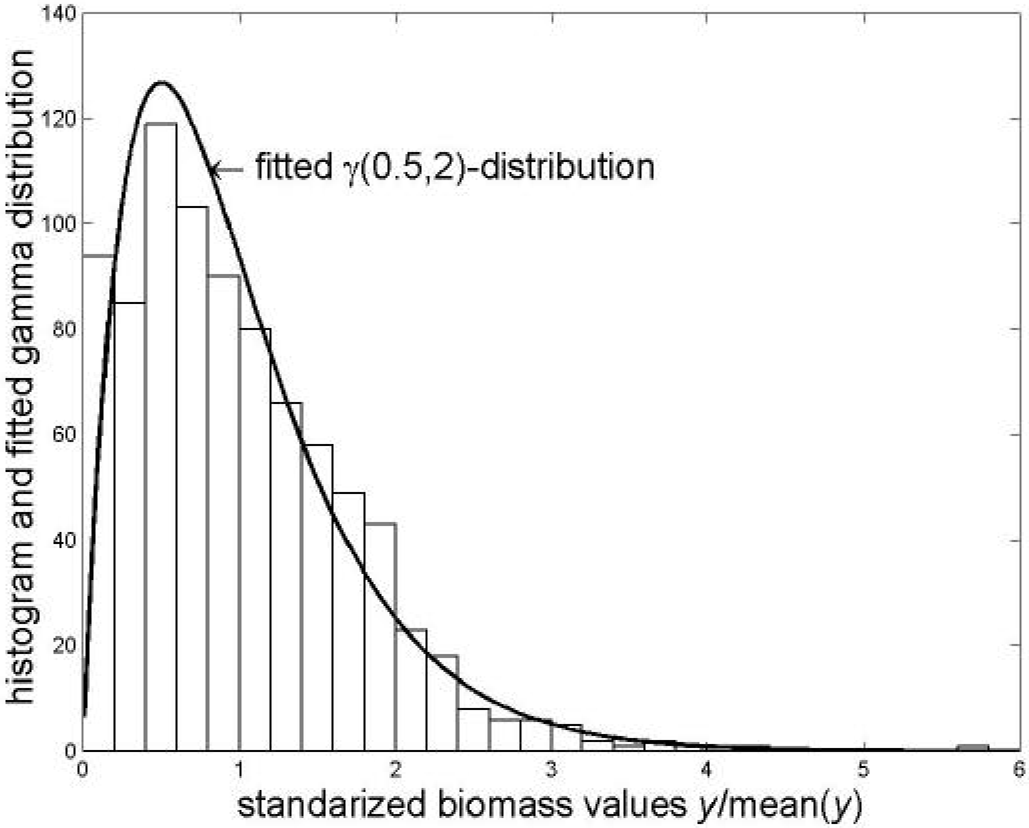

Fig. 3. Histogram of 861 standardised biomass values, $Y_{ij} / \overline{Y}_{ij}$, with fitted γ(0.2, 2)-distribution from the shrimp surveys in the Barents Sea 1992-1999.

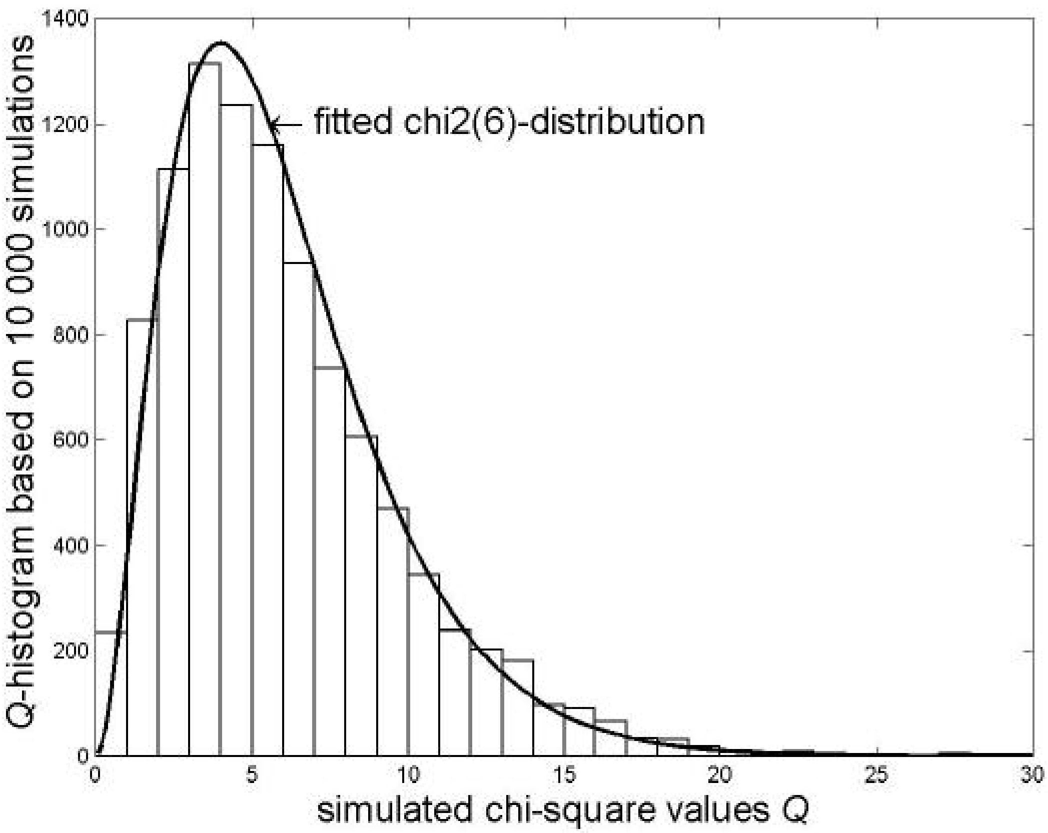

Fig. 4. Simulated distribution of chi-square sum, Q, of squared "normalised" relative stratum means, based on $\gamma(\mu/2, 2)$ as a population distribution with characteristic values for μ and sample sizes from one survey in the Barents Sea, see Tab. 1 and 2.

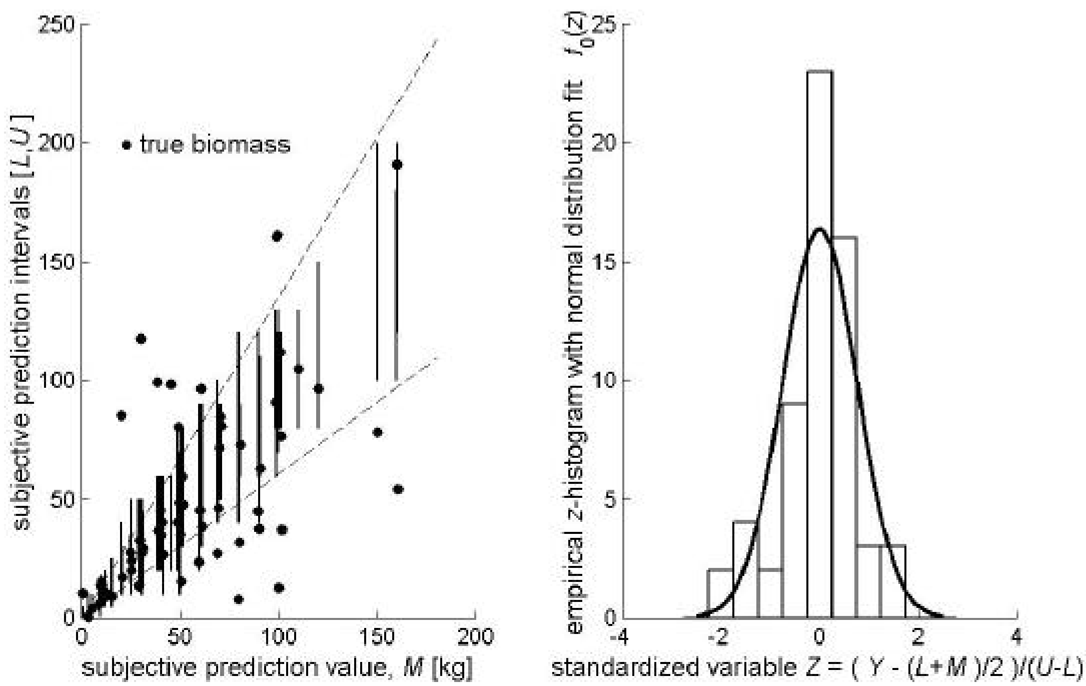

Fig. 5. Left: 62 subjective prediction intervals [L,U] for biomass in next trawl haul as a function of corresponding modal predicted value, M, for cruise leader in the 1998 subjective prediction experiment. The corresponding true biomasses are plotted as dots. Right: Histogram of standardised variable z based on data in left part, with fitted normal standardised subjective prediction distribution, $f_{s0}(z)$. See text.

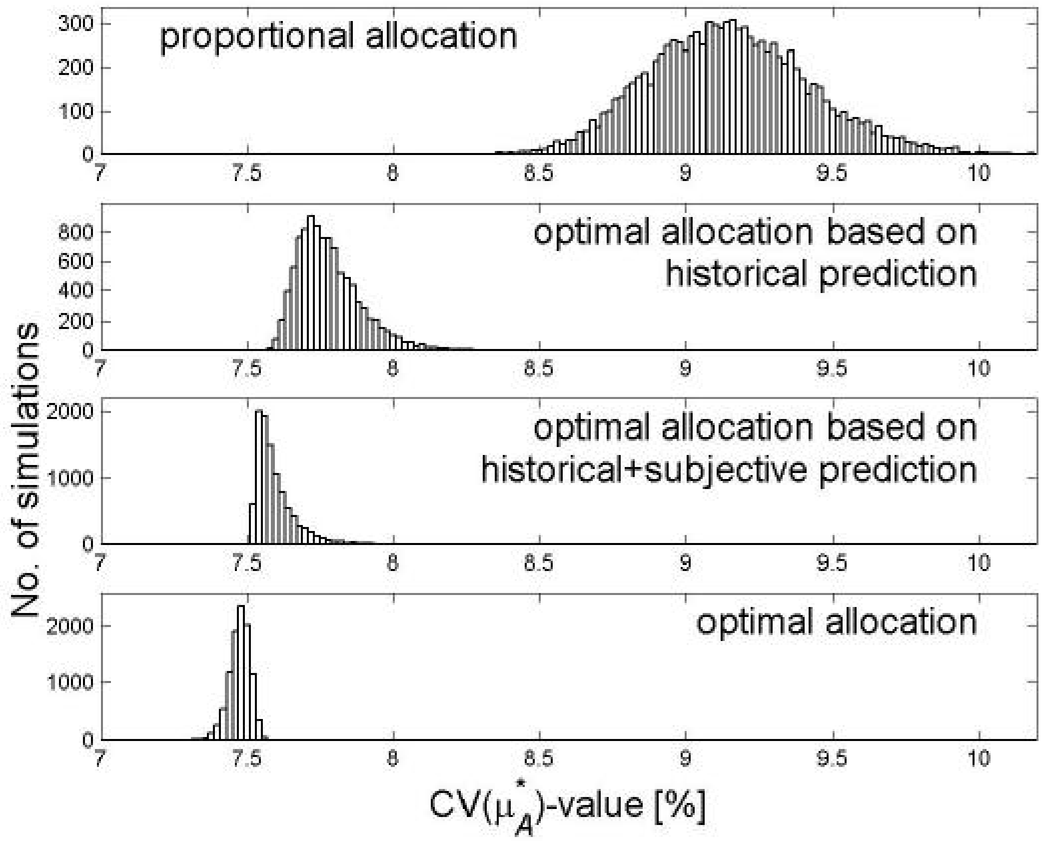

Fig. 6. Simulations of cv(μ_A^*)-distributions based on different predictions of relative strata mean vector $\underline{\mu}_r$, based on 10000 simulations of true strata means from cruise leader's subjective prediction distribution. See text.