NRC Research Press

ARTICLE

Efficient statistical estimators and sampling strategies for estimating the age composition of fish

Sondre Aanes and Jon Helge Vølstad

Abstract: Estimates of age compositions of fish populations or catches that are fundamental inputs to analytical stock assessment models are generally obtained from sample surveys, and multistage cluster sampling of fish is the norm. We use simulations and extensive empirical survey data for Northeast Arctic cod (*Gadus morhua*) to compare the efficiency of estimators that use age–length keys (ALKs) with design-based estimators for estimating age compositions of fish. The design-based weighted ratio estimator produces the most accurate estimates for cluster-correlated data, and an alternative estimator based on a weighted ALK is equivalent under certain constraints. Using simulations to evaluate subsampling strategies, we show that otolith collections from a length-stratified subsample of one fish per 5 cm length bin (~10 fish total) per haul or trip is sufficient and nearly as efficient as a random subsample of 20 fish. Our study also indicates that the common practice of applying fixed ALKs to length composition data can severely underestimate the variance in estimates of age compositions and that "borrowing" of ALKs developed for other gears, areas, or time periods can cause serious bias.

Résumé : Les estimations de la composition par âge de populations ou de prises de poissons, des intrants fondamentaux des modèles analytiques d'évaluation des stocks, sont généralement obtenues à partir d'évaluations sur échantillon, l'échantillonnage en grappes en plusieurs étapes des poissons en constituant la norme. Nous avons utilisé des simulations et de vastes données d'évaluations empiriques pour la morue (*Gadus morhua*) du nord-est de l'Arctique pour comparer l'efficacité, pour estimer les compositions par âge de poissons, des estimateurs qui utilisent des clés âge-longueur (CAL) à celle d'estimateurs fondés sur le plan d'évaluation. L'estimateur de rapport pondéré fondé sur le plan produit les estimations les plus exactes pour les données corrélées par grappes, et un autre estimateur basé sur une CAL pondérée est équivalent moyennant certaines contraintes. En utilisant des simulations pour évaluer des stratégies de sous-échantillonnage, nous démontrons que des collections d'otolites d'un sous-échantillon stratifié selon la longueur d' un poisson par compartiment de 5 cm de longueur (~10 poissons au total) par prise ou sortie sont suffisantes et presque aussi efficaces qu'un sous-échantillon aléatoire de 20 poissons. Notre étude indique également que la pratique répandue consistant à appliquer des CAL fixes à des données de composition selon la longueur peut se traduire par une grave sous-estimation de la variance des estimations des compositions par âge et que « l'emprunt » de CAL définies pour d'autres engins, zones ou périodes peut entraîner d'importants biais. [Traduit par la Rédaction]

Introduction

Estimates of abundance indices and catch in numbers of fish per age-class that are fundamental inputs to analytical stock assessments of many commercially important fish stocks are generally based on data from multistage sampling (Lehtonen and Pahkinen 2004). The primary sampling units (PSUs) in scientific trawl surveys of demersal fish are the trawl stations. Trawl hauls at each station are typically standardized by area swept, towing distance, or tow duration. In sample surveys of commercial landings, the PSUs are often defined by vessel trips, or site-days where trips are secondary sampling units (SSUs) nested within a given landing site and day (ICES 2013, 2014). The catches within PSUs are typically subsampled in one or more stages to obtain biological data, such as individual fish length and weight, and to collect otoliths. The subsamples of otoliths, which are analyzed in the laboratory to determine the age of fish, are referred to as "age samples" in this paper. Age samples are typically collected from a simple random, length-stratified, or systematic subsample of fish measured for length. This complex multistage cluster sampling is an inevitable result of having to use indirect sampling frames (Lavallée 2007) to get access to fish that can be measured for length and for collecting age samples. Unfortunately, cluster sampling of fish to support stock assessments is usually inefficient because it is generally not possible to create clusters that closely represent the length and age compositions of the entire target population, which is ideal (e.g., Thompson 1997, p. 23). In contrast, it is common that lengths and ages of fish sampled in clusters exhibit positive intracluster correlation, which can drastically reduce the effective sample sizes for estimating length and age compositions (Pennington and Vølstad 1994; Aanes and Pennington 2003; Helle and Pennington 2004; Nelson 2014; Stewart and Hamel 2014). It is well known from sampling theory that multistage cluster sampling may greatly inflate the variance in estimates of population characteristics as compared with simple random sampling of the same number of elementary units (specimens of fish in this paper) from a population (e.g., Särndal et al. 1992; Lehtonen and Pahkinen 2004). However, this fact has often been ignored in the field of fisheries research (Nelson 2014).

The classic age–length key (ALK) method introduced by Fridriksson (1934) is widely used by fisheries biologists to estimate age compositions of fish stocks and commercial landings. An ALK defines the proportion of fish in a length bin that fall into a particular age-class. It is common to derive the ALK from a small

Received 9 September 2014. Accepted 23 February 2015.

Paper handled by Associate Editor Verena Trenkel.

S. Aanes. Norwegian Computing Center, P.O. Box 114 Blindern, N-0314 Oslo, Norway.
J.H. Vølstad. Institute of Marine Research, P.O. Box 1870 Nordnes, N-5817 Bergen, Norway.
Corresponding author: Sondre Aanes (e-mail: sondre.aanes@nr.no).

sample of fish cross-classified by age and length and apply it to a length distribution derived from a second independent sample from the same population. An appealing feature about the ALK method is that it can utilize additional length samples as a supplement to the more costly age samples in the estimation of age compositions. In many fisheries-independent and fisheries-dependent sample surveys, concurrent age and length samples may be collected from only a subset of PSUs. PSUs containing data on lengths, but no age samples, are hereafter denoted "length-only data".

A key assumption for the classic ALK method is that the distribution of the ages of fish per length bin remains constant from sample to sample (Kimura 1977; Westrheim and Ricker 1978). There is an extensive body of literature where authors have used statistical estimators of age compositions that can handle length and age data from more complex sampling schemes. The statistical methods presented by Kimura (1977), Sen (1986), Lai (1987, 1993), and Morton and Bravington (2008), for example, assume two-stage sampling where a simple random sample of fish is measured for length in the first stage, and a small number of age samples is taken by simple random or by length-stratified random sampling in the second stage. Kimura and Chikuni (1987) and Hoenig and Heisey (1987) present methods where they combine length-at-age distributions (also commonly used to estimate growth) derived from one sample with a length-frequency distribution derived from another independent sample of fish measured for length only. Kimura and Chikuni (1987) ignore sampling variability in the estimated distribution of length-at-age, but they account for sampling variability in the independent estimates of length-frequency distributions. The Hoenig and Heisey (1987) method accounts for sampling variability in estimates from both samples, and Hoenig et al. (2002) extends the method to data from multiple surveys. The above methods do not require that the two independent samples be drawn from the same target population, in contrast with the classic ALK method (Fridriksson 1934). This relaxed condition is achieved by utilizing the distribution of length-at-age, which is independent of the age composition of the target population. However, it is assumed that the distribution of length-at-age is the same for both populations. Common to all the above methods is the implied assumption of simple random sampling of fish from the target population. Because fish generally are sampled using a hierarchical sampling design, the assumption is rarely if ever met in practice.

In this paper, the efficiency of estimators is evaluated in a simulation study by comparing estimates of age compositions to a reference distribution of true values from a synthetic population adapted to Northeast Arctic (NEA) cod (*Gadus morhua*) in the Barents Sea. We compare weighted and unweighted statistical estimators of age compositions of fish based on multistage cluster sampling (e.g., Cochran 1977; Särndal et al. 1992; Lehtonen and Pahkinen 2004) and estimators based on ALKs (e.g., Quinn and Deriso 1999) for sampling designs including concurrent subsamples of age for each length sample. We follow Jessen (1978) and use the term accuracy as a measure of the proximity of an estimate to the true value (i.e., high accuracy signifies high precision and low bias). Precision in estimates of proportions-at-age is measured by the standard error (SE) or relative standard error (RSE = SE/$\hat{p}_k$) (Jessen 1978) for each estimator, while accuracy is measured by the mean square error: MSE = SE^2 + (bias)^2 (Cochran 1977).

We also use the simulation study to evaluate the efficiency of an estimator of proportions-at-age based on an ALK assumed to be known exactly. It is often argued that this assumption can be justified when lengths and age samples are available from a large number of fish. In practice, however, sampling errors in ALKs can be large because of the multistage cluster sampling, even though age–length data are obtained from a large number of fish (e.g., Nelson 2014).

A special case of the use of a fixed ALK is the common practice of "borrowing" ALKs in stratified surveys to fill data gaps in the estimation of the age composition, for example when concurrent age and length samples are lacking for some geographic areas, quarters, or gears (ICES 2014). To illustrate the risk of bias due to the borrowing of ALKs, we provide an example based on empirical data for NEA cod.

We apply resampling methods (Efron 1982) to extensive empirical age and length data from scientific trawl surveys and catch sample data for the NEA cod fishery to determine cost-effective subsampling strategies for collecting age samples. We compare two commonly used subsampling strategies for collecting age samples from a subsample of fish measured for length within PSUs: (*i*) otolith collections from a simple random sample of fish and (*ii*) systematic otoliths collections from a fixed number of fish within length bins (length-stratified sampling). Finally, we evaluate alternative survey designs that on average yield similar number of age samples in total and assess the value of collecting additional length-only data with respect to accuracy of estimated age compositions.

Methods

We address three main questions in this paper. The first avenue of inquiry is which estimator provides the most accurate estimates of age composition given that data are collected using a hierarchical sample design. Six estimators are tested: a weighted and an unweighted ALK, a weighted and an unweighted ALK that is assumed to be known and constant, and a weighted and an unweighted design-based estimator. Symbols and parameters are described in Table 1, and a summary of the estimators is given in Table 2. The efficiency of these estimators are based on simulations and assessed by overall accuracy (MSE, goodness of fit, and confidence interval coverage rates using a synthetic population as the reference. Variances are estimated nonparametrically using bootstrapping methods. The second evaluation undertaken in this work is to assess the impact of borrowing ALKs for imputation. This practice is common in analysis of both fishery-independent and fishery-dependent monitoring data, although the resulting impacts in terms of bias and overall accuracy is not well established. Lastly, given the results of the first two steps, we compare the effectiveness of different sampling strategies with respect to expected precision in proportions-at-age, again using simulation methods for both trawl survey and commercial fishery sample data (Table 3).

Estimators of age composition of fish

Estimators based on ALKs

An estimator for the proportion of fish in the target population M in age-class k, p_k, expressed as the product of the length distribution of all fish M in the population of N PSUs and the ALK, is

$$\hat{p}_k^{\mathrm{w,ALK}} = \sum_{j=1}^{L} \hat{p}_j \times \hat{q}_{jk} \quad (1)$$

where $\hat{p}_j$ and $\hat{q}_{jk}$ are estimators for p_j and q_{jk}, respectively (Table 1).

Tanaka (1953), Kimura (1977), and Lai (1993) provide variance estimators for eq. 1 under the assumption that a simple random sample of fish is cross-classified by length and age. The estimator from eq. 1 can yield reliable estimates of age compositions for the realistic situation where a cluster of fish is sampled in the first stage, and the associated variance can be estimated by bootstrapping (Efron 1982). The estimator $\hat{q}_{jk}$ can be expressed as the ratio

$$\hat{q}_{jk} = \frac{\sum_i M_i \times \hat{p}_{ij} \times \hat{q}_{ijk}}{\sum_i M_i \times \hat{p}_{ij}} \quad (2)$$

Table 1. List of symbols and parameters.

Symbol	Definition
M	Total number of fish in the target population
N	Total number of primary sampling units (PSUs) that together comprise the study population*
B	The number of length bins
A	The number of age-classes
i	An index for the PSU
j	An index for the length bin
k	An index for the age-class
M_i	The total number of fish in the ith PSU
M_{ij}	The total number of fish in length bin j for the ith PSU
n	The number of PSUs sampled from a total of N
m_i	The number of fish subsampled and measured for length from the ith PSU†
m'_i	The number of fish subsampled for otoliths collections (age samples) from sample m_i‡
m_{ij}	The number of fish in length bin j in subsample m_i from the ith PSU
m'_{ij}	The number of fish subsampled for otoliths collections (age samples) from the m_{ij} fish in length bin j for the ith PSU
m'_{ijk}	The number of fish in the subsample m'_{ij} assigned to age-class k for the ith PSU
p_j	Proportion of fish in the target population (M) in length bin j
p_{ij}	Proportion of fish in length bin j for the ith PSU
p_k	Proportion of fish in the target population (M) in age-class k
p_{ik}	Proportion of fish in age-class k for the ith PSU
q_{ijk}	Proportion of fish for the ith PSU and length bin j that fall into age-class k
q_{jk}	Proportion of fish in length bin j that fall into age-class k in the target population (M)
$\boldsymbol{q}_k$	Vector of proportions of fish in length bins that fall into age-class k in the target population (M), defined by $\boldsymbol{q}_k = \{q_{jk}\}_{j=1,\ldots,B}$
ALK	The age–length key for the target population (M) is the matrix defined by ALK = $\{\boldsymbol{q}_k\}_{k=1,\ldots,A}$

*Based on indirect sampling frames, for example all possible area-swept trawl hauls for a scientific survey, or total number of fishing trips for a fisheries-dependent survey. For a scientific trawl survey, $N = SA/\alpha$, where SA is the survey area covering the target population, and α is the standardized area swept by the trawl.

†These subsamples are secondary sampling units.

‡These subsamples are tertiary sampling units; note that $m'_i = m_i$ for the empirical fisheries-dependent survey data analyzed in our example.

Table 2. Summary of estimators.

Estimator	Description	Comment
$\hat{p}_k^{w,ALK}$	Use weighted estimates of population length distribution and ALK.	Is equivalent to $\hat{p}_k^{w}$ provided concurrent age and length samples within each PSU.
$\hat{p}_k^{uw,ALK}$	Use unweighted estimates of population length distribution and ALK.	Is not equivalent to $\hat{p}_k^{uw}$, although it may be shown that its performance is similar.
$\hat{p}_k^{w,ALK*}$	Use weighted estimates of population length distribution but assumes population ALK is known and constant.	The point estimate equals $\hat{p}_k^{w,ALK}$ provided q_{jk}^* equals $\hat{q}_{jk}$ for all age-classes. This will thus only differ in performance with respect to variability and coverage, since the variance in the ALK is ignored and thus underestimated.
$\hat{p}_k^{uw,ALK*}$	Use unweighted estimates of population length distribution but assumes population ALK is known and constant.	The point estimate equals $\hat{p}_k^{uw,ALK}$ provided q_{jk}^* equals $\hat{q}_{jk}$ for all age-classes. This will thus only differ in performance with respect to variability and coverage, since the variance in the ALK is ignored and thus underestimated.
$\hat{p}_k^{w}$	Direct estimation of the age distribution utilizing the sampling weights.	Applies an ALK within each PSU to estimate the PSU-specific age distributions, and then the population age distribution is estimated by a weighted mean of the age distributions. This estimator requires concurrent samples of age and lengths within each PSU, in which case it is equivalent to $\hat{p}_k^{w,ALK}$.
$\hat{p}_k^{uw}$	Direct estimation of the age distribution using equal weights.	As for $\hat{p}_k^{w}$, but with an unweighted mean of the PSU-specific age distributions. Its performance may be shown to be similar to $\hat{p}_k^{uw,ALK}$.

Table 3. Summary of data sets used for assessing estimators for age composition.

Data set	Description	Purpose
SynthPop	A synthetic population adapted to mimic key features observed in the joint Norwegian–Russian bottom trawl survey in the Barents Sea during winter.	Used to define reference distribution to evaluate estimators in terms of goodness of fit, mean squared error, relative standard error, and coverage.
TrawlSurv	Trawl survey data from the 1991 joint Norwegian–Russian winter survey. Systematic subsampling of up to 10 ages within each 5 cm length category within each PSU.	Evaluate precision as a function of number of ages sampled within each length category, number of PSUs, and the effect of including additional length-only samples.
ComDat	Commercial catch samples from the Norwegian fishery in 2000. Random and concurrent subsampling of both ages and lengths.	Evaluate simple random subsampling of ages, systematic age sampling, and simple random subsampling, where the sample size is proportional to the catch size for the PSU.
ComGillLine	A subset of ComDat including catch samples from the Norwegian gillnet and longline fisheries in the Lofoten area during March 2000 where catch operations overlap closely in area, depth, and time.	Provide empirical examples of the effect of using fixed and borrowed ALK.

Note: All data sets include Northeast Arctic cod.

where $\hat{p}_{ij} = m_{ij}/m_i$ is an estimator of the proportion of fish in length bin j for the ith PSU
and

$$\hat{q}_{ijk} = \frac{m'_{ijk}}{\sum_k m'_{ijk}} \quad (3)$$

is an estimator for q_{ijk} for the ith PSU (Table 1). The estimator from eq. 2 is a weighted estimator that can be expressed as

$$\hat{q}_{jk} = \sum_i w_{ij} \times \hat{q}_{ijk} \quad (4)$$

with weights defined by

$$w_{ij} = \frac{\hat{M}_{ij}}{\sum_j \hat{M}_{ij}} \quad (5)$$

where $\hat{M}_{ij} = M_i \times \hat{p}_{ij}$ is an estimator of the number of fish in length bin j for the ith PSU.

By the same general approach as above, the weighted estimator for proportion of fish in length bin j is

$$\hat{p}_j = \sum_i w_i \times \hat{p}_{ij} \quad (6)$$

with weights defined by

$$w_i = \frac{M_i}{\sum_i M_i} \quad (7)$$

We also evaluate the performance of an estimator for p_k where an unweighted ALK is applied to an unweighted estimate of the length distribution (i.e., M_i and $\hat{M}_{ij}$ in the weights from eqs. 5 and 7 are replaced by constants, and thus the weights are constant and sum to one). This unweighted (uw) estimator is denoted $\hat{p}_k^{\mathrm{uw,ALK}}$ in the rest of this paper.

Estimates of age compositions of fish are often obtained by applying a fixed ALK to population estimates of length distributions from survey sample data. Using weighted or unweighted estimators, let $\hat{p}_k^{\mathrm{ALK}^*} = \hat{p}_k^{\mathrm{ALK}}|\hat{\boldsymbol{q}}_k = \boldsymbol{q}_k^*$, where the kth column-vector $\hat{\boldsymbol{q}}_k = \{\hat{q}_{jk}\}_{j=1,\ldots,B}$ in the ALK is replaced by single and fixed estimates $\boldsymbol{q}_k^* = \{q_{jk}^*\}_{j=1,\ldots,B}$, and ALK* denotes the fixed ALK. If $\boldsymbol{q}_k^* = E(\hat{\boldsymbol{q}}_k)$, then $E(\hat{p}_k^{\mathrm{ALK}^*}) = E(\hat{p}_k^{\mathrm{ALK}})$, but if $E(\hat{\boldsymbol{q}}_k) \neq \boldsymbol{q}_k^*$, the equality will in general not hold. The effect of ignoring sampling errors in the ALK on estimates of variance can be seen by decomposing the variance of the proportions-at-age, using the conditional variance formula

$$\mathrm{Var}\left(\hat{p}_k^{\mathrm{ALK}}\right) = E_{\hat{\boldsymbol{q}}_k}\left[\mathrm{Var}\left(\hat{p}_k^{\mathrm{ALK}}|\hat{\boldsymbol{q}}_k\right)\right] + \mathrm{Var}_{\hat{\boldsymbol{q}}_k}\left[E\left(\hat{p}_k^{\mathrm{ALK}}|\hat{\boldsymbol{q}}_k\right)\right]$$

From this decomposition it is apparent that the first term includes sampling errors in the estimated length distributions, whereas the second term includes sampling errors in the estimated ALK, in addition to interaction between $\hat{p}_k^{\mathrm{ALK}}$ and $\hat{\boldsymbol{q}}_k$ in both terms. When the ALK is estimated once and fixed, only $\mathrm{Var}(\hat{p}_k^{\mathrm{ALK}}|\hat{\boldsymbol{q}}_k) = \mathrm{Var}(\hat{p}_k^{\mathrm{ALK}^*})$ contributes to the variance of $\hat{p}_k^{\mathrm{ALK}}$, and hence the true sampling variability in age compositions is underestimated. In the special case where fixed ALKs are borrowed to fill data gaps, it is likely that $E(\hat{\boldsymbol{q}}_k) \neq \boldsymbol{q}_k^*$, such that bias is introduced in estimates of p_k for all ages (see, e.g., Quinn and Deriso 1999 pp. 311–317; Hoenig et al. 2002).

Design-based estimators

This approach adopts the classical sampling theory for multistage cluster sampling (e.g., Cochran 1977), where

1. The first stage is a random sample of PSUs (e.g., trawl stations or vessel trips).
2. The second stage is a random subsample of fish (SSU) from the ith PSU that is sorted by species and measured for length.
3. The third stage involves taking either a simple random or a length-stratified age sample from selected species in the SSU.

It should be noted that a truly length-stratified age sample cannot easily be achieved, since the strata sizes defined by the total number of fish in each length bin for the ith PSU often are unknown and have to be estimated. In practice, otoliths are collected systematically from a fixed number of fish from each length bin, and the total number of fish for each length bin (stratum size) is typically estimated using the length composition of fish from the SSU sample. The estimator for the proportion of fish at age k in the population is

$$\hat{p}_k^{\mathrm{w}} = \sum_i w_i \times \hat{p}_{ik} \quad (8)$$

which is a weighted ratio estimator where the weights w_i are given by eq. 7 and

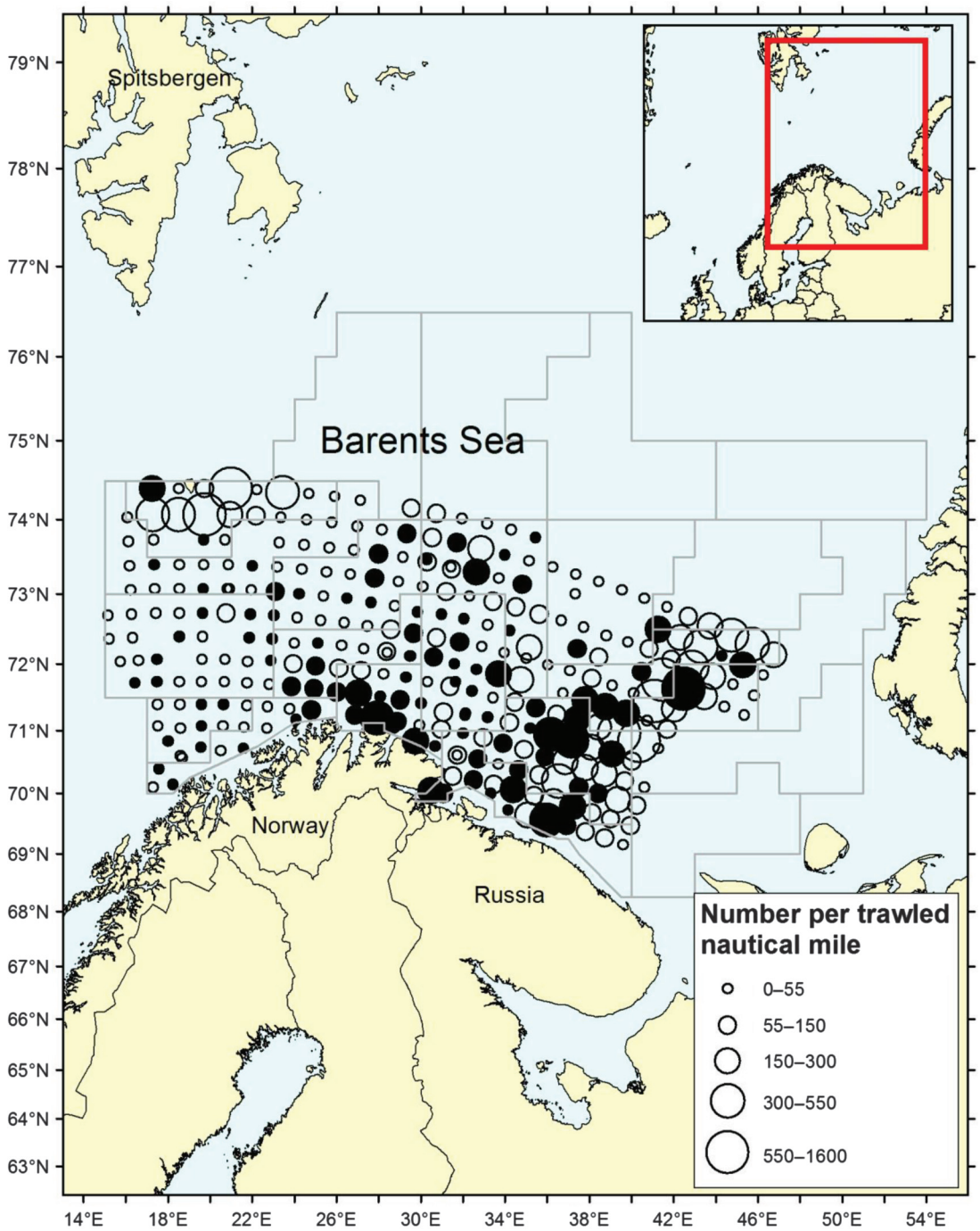
$$\hat{p}_{ik} = \sum_{j=1}^{L} \frac{m_{ij}}{m_i} \times \frac{m'_{ijk}}{m'_{ij}} = \sum_{j=1}^{L} \hat{p}_{ij} \times \hat{q}_{ijk} \quad (9)$$

When concurrent age and length samples are obtained from all PSUs, it can be shown that $\hat{p}_k^{\mathrm{w}} \equiv \hat{p}_k^{\mathrm{w,ALK}}$. This is seen if eq. 9 is substituted into the estimator from eq. 8, and compared with the estimator from eq. 1, with eqs. 4 and 6 substituted for $\hat{p}_j$ and $\hat{q}_{jk}$, respectively. We also consider an alternative unweighted version of the estimator from eq. 8, denoted $\hat{p}_k^{\mathrm{uw}}$, where M_i in the weights w_i (eq. 7) are set to a constant value; note that $\hat{p}_k^{\mathrm{uw}} \neq \hat{p}_k^{\mathrm{uw,ALK}}$ in this case. Aanes and Pennington (2003) provide an estimator for p_k based on data from fisheries-dependent surveys where a random sample of catches (PSUs) is taken in the first stage, and a simple random subsample of fish for ageing is taken from each PSU. Their estimator is a special case of eq. 8 under simple random sampling and is only used here in the comparison of sampling strategies for the commercial fishery survey.

Bootstrap variance estimation

We used nonparametric bootstrapping (Efron 1982) to estimate the variance of age compositions and proportions-at-age for all estimators. The bootstrap resampling procedure replicated the multistage sampling designs. Since the sampling fraction of PSUs generally is negligible, we assumed sampling with replacement in the first stage (see, e.g., Williams 2000). Each bootstrap replicate was generated by first sampling the PSUs at random with replacement. The age samples within each PSU were then selected by simple random sampling (catch data) or by systematic sampling from length bins (stratified sampling; trawl survey data). Since the age samples generally were collected from a relatively small finite population of fish measured for length, the bootstrap subsamples within PSUS were selected without replacement, using methods described in Booth et al. (1994) and Davison and Hinkley (1997). The effect of subsampling sizes on the precision of age compositions was studied through simulations, where we varied the num-

Fig. 1. Catch (in numbers) of NEA cod per nautical mile (1 n.mi. = 1.852 km) trawled during the joint Russian–Norwegian winter survey in the Barents Sea in 1991. NEA cod were caught at 297 out of the 300 trawl stations. A sample of approximately 100 fish was measured for length from each catch, while otoliths were collected from a subsample of fish measured for length at 102 stations (solid circles). A total of 26 597 individuals were measured for length, and 3736 fish were aged.

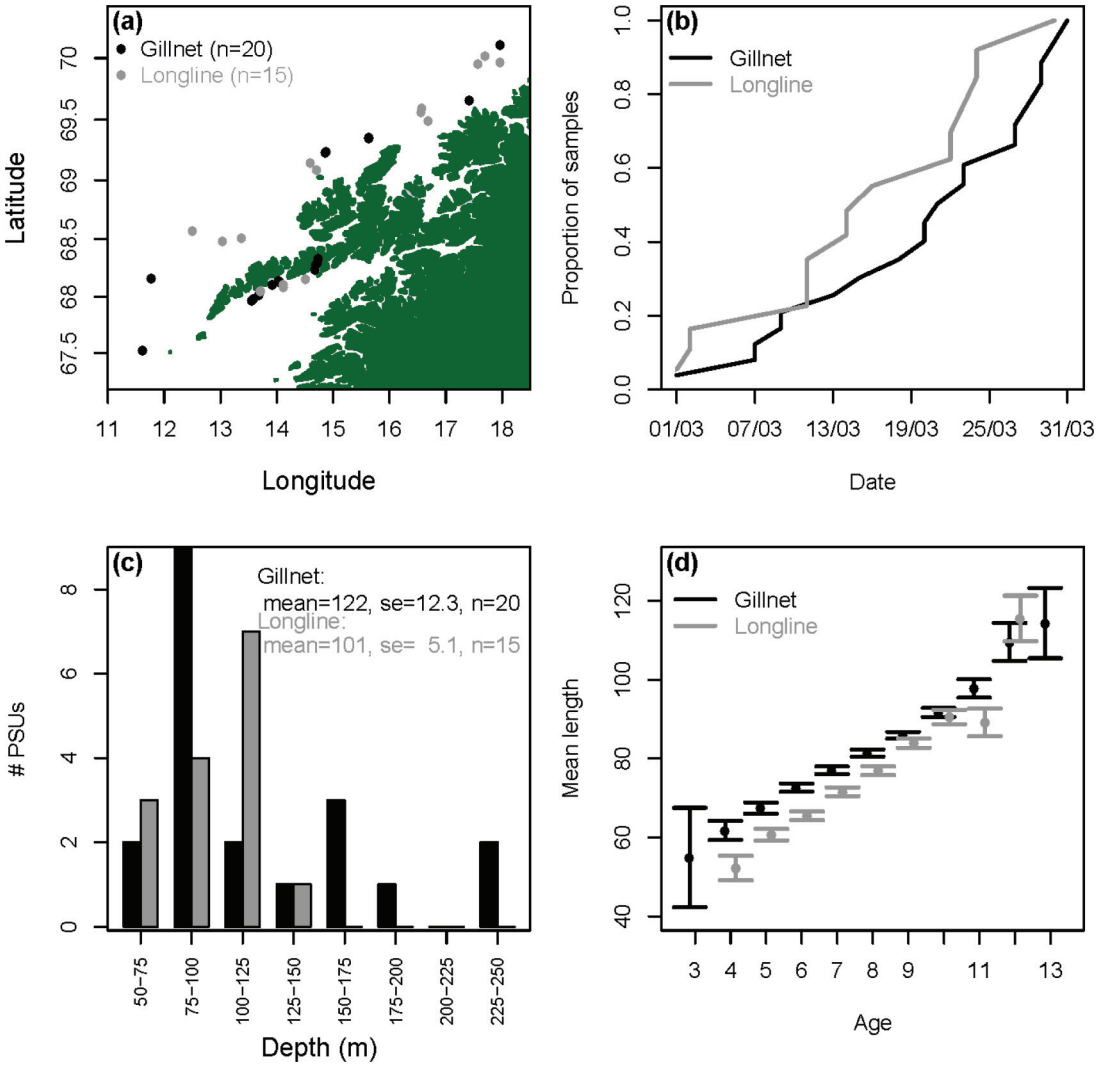


ber of age samples from each PSU in the bootstrap resampling. The bootstrap estimates of variance for each survey were based on 2000 replicates.

Evaluation of estimators

Since it is not possible to get reliable estimates of bias in estimates of proportions-at-age from empirical survey data, we assessed efficiency of estimators (Table 2) in terms of accuracy by comparing estimates with true values for a synthetic population in a simulation study. We used a stochastic simulation model to create a realistic synthetic population of PSUs for NEA cod (SynthPop; Table 3, $N = 100\,000$). This population contains concurrent age and length data, varying numbers of fish per cluster, and the complex variance–covariance structure in age and length compositions typically observed in trawl surveys for NEA cod (see Appendix A). The true proportions-at-age p_k were derived for all fish M in SynthPop and used as a reference in the evaluation of estimators. Using the PSUs in SynthPop as a sampling frame, we simulated 400 trawl surveys with a probabilistic multistage cluster sampling

Fig. 2. Comparing samples from gillnet and longline from the commercial fishery of cod in the Lofoten area during March in 2000 (ComGillLine) by (*a*) area, (*b*) the time of sampling (day/month), (*c*) depth, and (*d*) the estimated mean length-at-age. Gillnet samples are shown in black, while longline samples are shown in gray. The time of sampling is represented by the cumulative proportions of the total respective number of samples by date. In panel (*d*), the estimated mean length-at-age is shown by dots, while the bars represent 95% confidence intervals.



design, where a simple random sample of PSUs ($n = 200$) was taken from the synthetic population N, and age samples were collected concurrently with lengths from all PSUs from each trawl survey.

We used a goodness of fit statistic (Fleiss et al. 2003) F to identify which of the estimators (Table 2) produces the most accurate estimates of age composition for each simulated trawl survey (i.e., closest to the true proportions-at-age in the reference distribution). The F statistic for an estimator $\hat{p}_k$ is defined by

$$F_{\hat{p}} = \sum_{k=1}^{A} \frac{(\hat{p}_k - p_k)^2}{p_k} \quad (10)$$

where $F_{\hat{p}}$ is the difference between the point estimate of the age composition $\hat{\boldsymbol{p}} = \{\hat{p}_k\}_{k=1,\ldots,A}$ from a simulated survey and the true age composition $\boldsymbol{p} = \{p_k\}_{k=1,\ldots,A}$ for SynthPop. $F_{\hat{p}}$ is zero for $\hat{\boldsymbol{p}} = \boldsymbol{p}$ and increases with increasing difference between estimated and true proportions-at-age. The F statistic (eq. 10) was used to compare point estimates of the age composition for the weighted and unweighted estimators for each of the 400 simulated trawl surveys. We used MSE and coverage probability (Fleiss et al. 2003) of

Table 4. Summary statistics for evaluation of the estimators $\hat{\boldsymbol{p}}^{w}$, $\hat{\boldsymbol{p}}^{uw}$, and $\hat{\boldsymbol{p}}^{uw,ALK}$ for proportions-at-age, where $\hat{\boldsymbol{p}}^{\bullet} = \{\hat{p}_k^{\bullet}\}_{k=1,\ldots,A}$.

Estimator	$\tilde{F}_{\hat{p}^{\bullet}}$	$P(F_{\hat{p}^w} < F_{\hat{p}^{\bullet}})$	%|rel bias|
$\hat{\boldsymbol{p}}^{w}$	0.020	—	0.9
$\hat{\boldsymbol{p}}^{uw}$	0.082	0.99	29.8
$\hat{\boldsymbol{p}}^{uw,ALK}$	0.075	0.98	22.5

Note: We considered age-classes 1–8, which accounts for 98% of the total number of fish in SynthPop. The statistics are based on estimates from 400 replicated surveys of SynthPop, each with sample size of 200 PSUs. For each simulated survey, the value of the F statistic for an estimator is obtained by comparing the estimated age composition with the true age composition in SynthPop. $\tilde{F}_{\hat{p}^{\bullet}}$ denotes the median of the F statistic (eq. 10), $P(F_{\hat{p}^w} < F_{\hat{p}^{\bullet}})$ denotes the probability of the F statistic $F_{\hat{p}^w}$ for a weighted design-based estimator being smaller than the F statistic for an alternative unweighted design-based estimator ($P(F_{\hat{p}^w} < F_{\hat{p}^{uw}})$) or an unweighted ALK estimator ($P(F_{\hat{p}^w} < F_{\hat{p}^{uw,ALK}})$), and %|rel bias| is the mean relative bias in proportions-at-age across age-classes.

Table 5. The average mean squared error (MSE) × 10^{-3}, mean standard error (SE) × 10^{-3}, and percent mean relative bias for the estimators $\hat{p}_k^{w}$, $\hat{p}_k^{uw}$, and $\hat{p}_k^{uw,ALK}$ by age-classes 1 through 8 derived by comparing the respective estimates of proportions-at-age from simulated surveys to the reference proportions-at-age p_k defined by SynthPop.

		Average MSE			Mean SE			% mean relative bias		
Age-class	p_k	$\hat{p}_k^{w}$	$\hat{p}_k^{uw}$	$\hat{p}_k^{uw,ALK}$	$\hat{p}_k^{w}$	$\hat{p}_k^{uw}$	$\hat{p}_k^{uw,ALK}$	$\hat{p}_k^{w}$	$\hat{p}_k^{uw}$	$\hat{p}_k^{uw,ALK}$
1	0.28	2.78	8.14	7.96	35.67	16.82	16.72	−1.06	−31.27	−30.91
2	0.19	1.33	1.17	1.56	24.58	15.05	14.24	−0.91	−13.96	−17.62
3	0.18	0.96	0.51	0.66	21.37	15.24	14.69	0.14	2.79	7.96
4	0.14	0.70	0.98	1.52	18.57	15.15	14.22	0.44	16.62	24.43
5	0.14	0.77	3.16	3.44	19.12	17.77	16.56	1.46	35.67	38.20
6	0.06	0.33	1.29	0.51	12.15	13.56	11.05	2.60	52.47	28.14
7	0.01	0.03	0.06	0.02	3.35	4.25	3.01	2.34	45.97	15.89
8	0.00	0.00	0.01	0.00	1.14	1.56	1.20	−0.33	40.12	18.00

Note: The means are based on 400 replicated sample surveys of SynthPop, each with sample size of 200 PSUs, and for each survey replicate the variance and bias is estimated using 2000 bootstrap replicates.

95% confidence intervals to assess accuracy of proportions-at-age estimates. The coverage probability of the 95% confidence interval for $\hat{p}_k^{w}$ is the proportion of the time that the estimated interval contains the true proportion value p_k in the reference distribution over repeated surveys. Efficient estimators will produce accurate interval estimates of proportions-at-age with coverage probability close to the nominal 95% level. Since proportions are restricted in the interval [0, 1], the distributions of the estimates cannot be expected to be symmetrical. We therefore considered several methods for generating non-parametric bootstrap confidence intervals: (*a*) the first order normal approximation, (*b*) the basic bootstrap interval, (*c*) the studentized bootstrap interval, (*d*) the bootstrap percentile interval, and (*e*) the adjusted bootstrap percentile (BCa) interval (Davison and Hinkley 1997) estimated using the boot library in R (Canty and Ripley 2013; Davison and Hinkley (1997)). Bootstrapping was used to provide point estimates of age compositions and confidence intervals around proportions-at-age for each survey.

Borrowing ALKs

In a case study we use empirical commercial catch sample data for NEA cod where ALKs were derived for different fishing gears to evaluate the effects of borrowing ALKs on the accuracy of estimates of proportions-at-age. We use data from gillnet and longline fleets that target the same population of NEA cod. We first estimate ALKs and proportions-at-age in the landings for both fleets and then compare the standard estimate for one fleet to an estimate based on a borrowed ALK from the other.

Evaluating age-sampling strategies based on empirical data

Trawl survey data

Survey sample data for NEA cod from the joint Norwegian–Russian winter survey (Pennington et al. 2011) were used to assess sampling strategies for estimating age composition from trawl surveys. This trawl survey has been conducted yearly since 1981 to provide abundance indices for cod and other demersal species, with yearly sampling effort varying between 176 and 394 trawl stations and approximately uniform distribution of stations in space. The PSU in the winter survey is a standardized area-swept trawl haul. Since the purpose of our analysis is to compare the performance of estimators and evaluate subsampling strategies for age, we collapsed the spatial strata and assumed simple random sampling of PSUs. This approach is reasonable because of the fairly uniform distribution of stations and was done since the variance estimates are based on bootstrapping, which performs poorly for small sample sizes (Korn and Graubard 1995).

TrawlSurv (Fig. 1; Table 3) were analyzed using resampling methods to assess the efficiency of the Norwegian age-sampling strategy for NEA cod. For the winter surveys conducted since 1996, otoliths have been collected from all trawl hauls. Otoliths from a subsample of one fish per 5 cm length bin are collected from a larger length sample of approximately 100 cod per haul. The typical strategy for age sampling from 1981 to 1995 was to collect five age samples per 5 cm length bin, but only from a subset of trawl stations. For the data from the 1991 survey (Fig. 1), length samples were obtained from 297 trawl stations in 1991, and otoliths were collected from a subsample of up to 10 fish per length bin from a subset of 102 stations. These data were used in a simulation study to assess alternative survey designs and sampling strategies for estimating age compositions:

(a) Random sample of 102 PSUs, with (up to) 10 age samples per length bin from all PSUs.

(b) Random sample of 297 PSUs, with (up to) 10 age samples per length bin from a random subset of 102 PSUs. This is the strategy employed in the winter survey in 1991 for NEA cod. We assumed that PSUs with concurrent age and length samples and PSUs with length-only samples were two independent simple random samples from the target population.

(c) Random sample of 297 PSUs, with one age sample per length bin from all PSUs. This is the age-sampling strategy for NEA cod applied since 1996 for the winter surveys.

We used bootstrapping to estimate the mean RSE in estimates of proportions-at-age for each survey design. The estimators $\hat{p}_k^{w,ALK}$ (eq. 1) and $\hat{p}_k^{w}$ (eq. 8) were applied for strategies (a) and (c) where age samples were collected concurrently with lengths from all PSUs, while $\hat{p}_k^{w,ALK}$ was applied for strategy (b), since only this estimator could accommodate the extra PSUs with length-only samples. Note that the estimator $\hat{p}_k^{w,ALK}$ differs from the estimator $\hat{p}_k^{w}$ in this case.

Survey sample data from a commercial fishery

We used extensive sample data for the NEA cod fishery in 2000 (ComDat, Table 3; see Aanes and Pennington 2003 for more details) to compare the performance of systematic (length-stratified) and simple random subsampling of otoliths. We assumed a simple random sample of fishing trips (PSUs). For each of 126 fishing trips sampled in 2000, concurrent length measurements and otoliths for age determination were obtained from a subsample of ~85 cod from the catches, on average. In the simulations, we resampled a fixed number of otoliths at random (without replacement) within 5 cm length bins, using the bootstrapping method described in Booth et al. (1994). The design-based weighted and unweighted estimators were used to estimate the age composition p_k for simulated systematic age sampling, while the estimator in Aanes and Pennington (2003) was used for simple random age samples. We also evaluated simple random subsampling where the number of age samples per PSU is proportional to the catch size for the trip (proportional allocation).

We used a subset of ComDat, restricted to 20 gillnet trips and 15 longline trips in Lofoten during March in 2000 (ComGillLine;

Table 6. Estimated percent coverage of the estimated 95% confidence interval for proportions-at-age.

	Age-class														
Estimator	1	2	3	4	5	6	7	8	9	10	11	12	13	Mean	Mean 1–8
$\hat{p}_k^{w}$	92	91	94	94	92	92	85	89	88	73	65	65	44	82	91
$\hat{p}_k^{uw}$	0	60	93	62	12	24	78	90	88	76	59	68	40	58	52
$\hat{p}_k^{uw,ALK}$	0	39	82	30	6	63	91	92	91	82	63	74	42	58	50
$\hat{p}_k^{w,ALK*}$	89	77	67	68	76	66	39	41	47	43	36	49	40	57	65
$\hat{p}_k^{uw,ALK*}$	0	19	47	5	1	21	41	41	48	50	41	55	36	31	22

Note: The coverage is based on 400 replicated sample surveys of SynthPop, each with sample size of 200 PSUs. For each survey replicate the confidence interval is estimated using the BCa method (see Methods) and 2000 bootstrap replicates.

Fig. 3. Estimated population characteristics of the commercial landings of NEA cod by gillnet and longline in the Lofoten area during the first quarter in 2000: (*a*) age compositions, (*b*) length compositions, (*c*) ALKs by dots with size proportional to the probability of age within each length category, and (*d*) age compositions of longline based on ALK for gillnet (borrowed). For the age composition in panel (*a*), the proportion at age 3 for gillnet is too small (0.0003) to show, whereas the proportion for longline is 0.

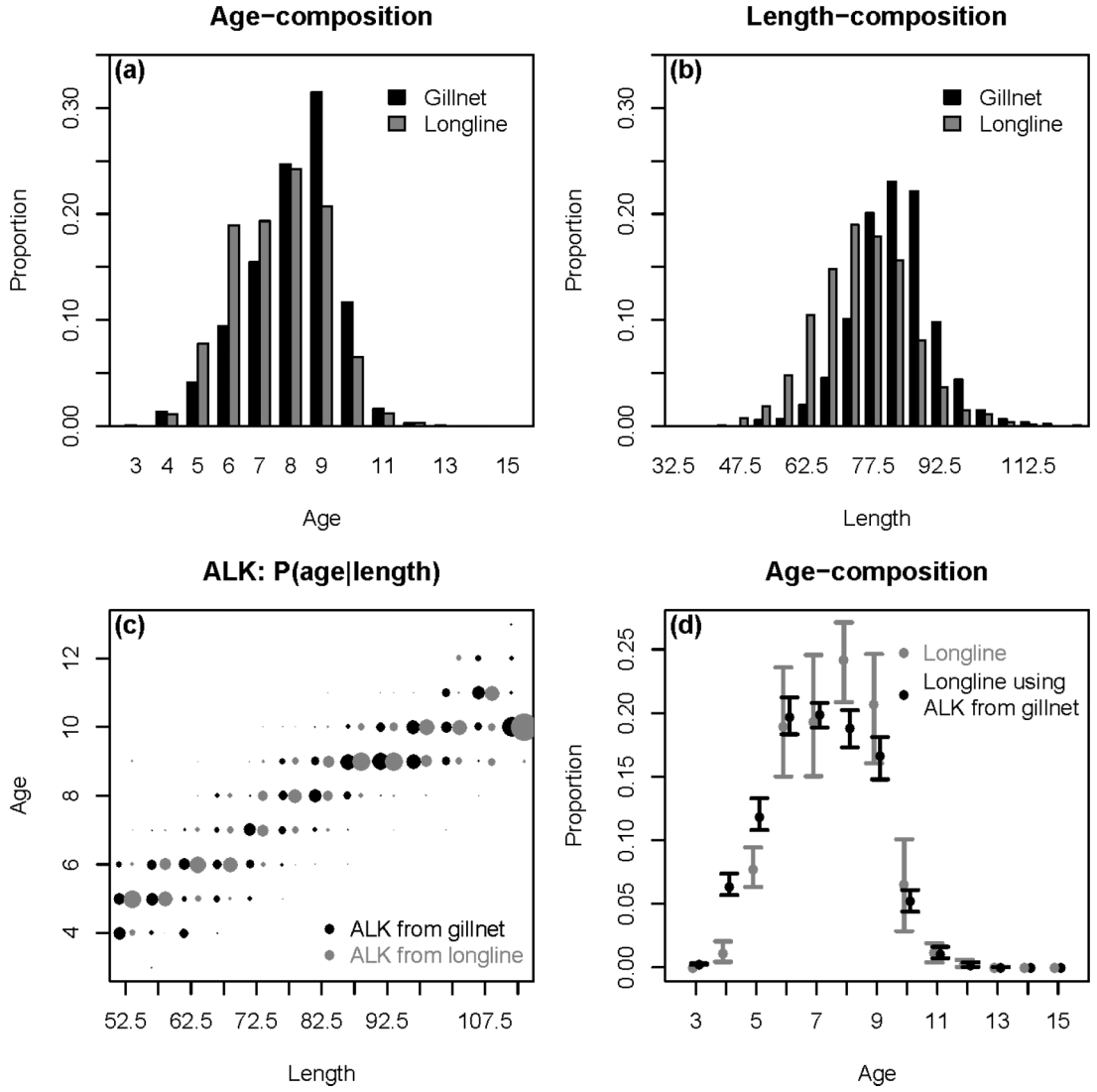


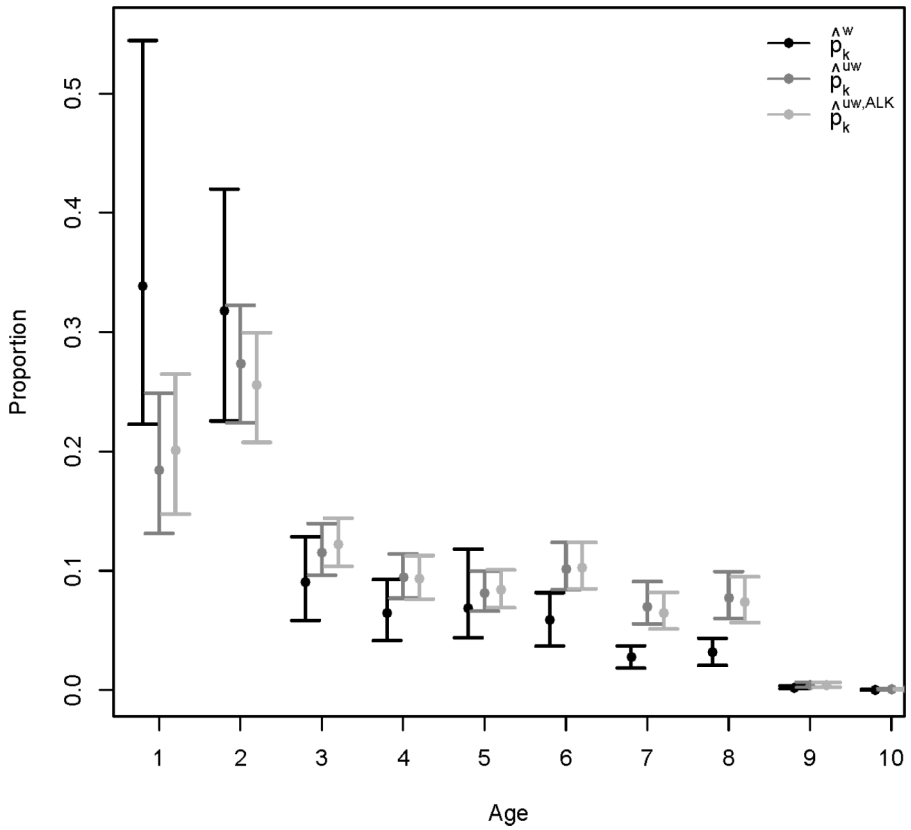
Table 3; Fig. 2), for the case study to illustrate the effect of borrowing ALKs. The fishing operations from the two fleets overlapped closely in space (Fig. 2*a*) and time (Fig. 2*b*), and there was no significant difference in mean depth of the fishing operations (Fig. 2*c*). Nevertheless, the total catch taken by each gear had different characteristics. Fish of ages 3 and 13, for example, were only caught by gillnets, and the mean lengths for age-classes 4–8 and for age-class 11 were significantly larger for gillnet than for longline (Fig. 2*d*).

Results

Evaluation of estimators

The simulation study using data from SynthPop with concurrent length and age sampling shows that the equivalent weighted estimators $\hat{p}_k^{w,ALK}$ (eq. 1) and $\hat{p}_k^{w}$ (eq. 8) outperformed the unweighted estimators. These weighted estimators produced estimates of age compositions with the smallest median F values, lowest bias, and the lowest F values for 98%–99% of the surveys

Fig. 4. Estimated proportions-at-age for NEA cod for three different estimators based on data (102 trawl hauls including age samples) from the joint Russian–Norwegian winter survey in 1991 (TrawlSurv). The error bars are estimated 95% confidence intervals using the BCa method (see Methods) and 2000 bootstrap replicates.



(Table 4). The unweighted estimators $\hat{p}_k^{uw}$ and $\hat{p}_k^{uw,ALK}$ perform poorly for all age-classes. With few exceptions, the weighted estimators also produced estimates of proportions-at-age with the lowest average MSE, and the lowest mean MSE across age-classes, even though the estimates from the weighted estimator generally are less precise than estimates from unweighted estimators (Table 5).

The different methods considered for estimating bootstrap confidence intervals had similar performance, but the BCa intervals were marginally closer to the nominal confidence level and method (*e*) was therefore used for estimating coverage probability. The coverage probability of the estimated 95% confidence interval for proportions-at-age based on the weighted estimator $\hat{p}_k^{w}$ was estimated at 90% on average across age-classes 1 through 8, which comprise 98% of the fish in the target population, but was reduced to an average of 81% across all age-classes because of the poor confidence interval coverage for those age-classes that comprise small portions of the target population (Table 6). The coverage probability was drastically reduced to approximately 56% on average when using the estimator $\hat{p}_k^{w,ALK*}$ with a fixed ALK. Interval estimates of proportions-at-age based on $\hat{p}_k^{uw,ALK*}$ with a fixed ALK have coverage probability lower than 15% for age-classes 1–5.

Borrowing ALKs

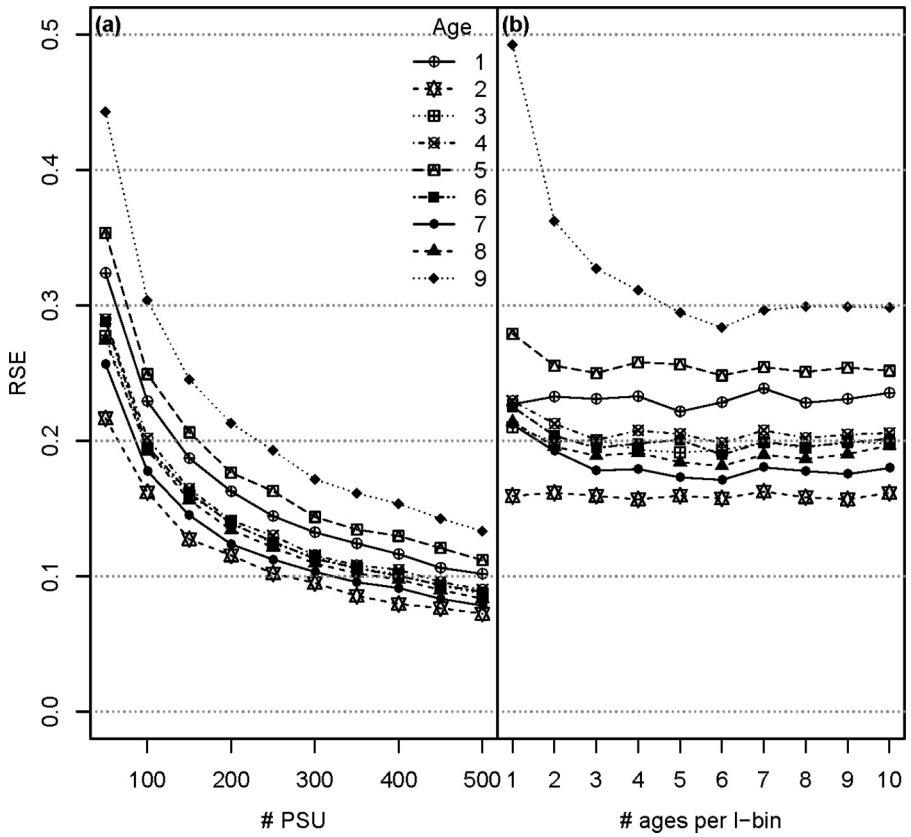
Results from analyses of ComGillLine show that the borrowing of a fixed ALK to fill data gaps in addition to the underestimation of variance also can introduce substantial bias in estimates of age compositions. The estimated age and length compositions of the gillnet and longline catches are rather different (Figs. 3*a*, 3*b*) even though the two fisheries target the same population of fish. The estimates for gillnet landings are dominated by 9-year-old fish in the length bins 75–90 cm, while the estimates for longline landings are dominated by 8-year-old fish in the length bins 65–80 cm. Nevertheless, their respective estimated ALKs appear rather similar by a visual inspection (Fig. 3*c*). The estimated proportions-at-age of fish landed in the longline fishery based on an ALK borrowed from the gillnet fishery are significantly different from the standard estimates for age-classes 4, 5, and 8 (Fig. 3*d*).

Evaluation of sampling strategies for age sampling based on empirical data

For TrawlSurv there was stronger negative correlation between mean age of fish per cluster and cluster size M_i than for the SynthPop data used in the simulation study to evaluate accuracy of estimators (Table A2). The estimated proportions-at-age from unweighted estimators (Fig. 4) based on TrawlSurv are therefore more precise than estimates from the weighted estimator $\hat{p}_k^{w}$, but they are likely to have larger bias and be less accurate. The differences in point estimates of proportions-at-age for all age-classes in TrawlSurv (Fig. 4) can be explained by the expected bias, Bias($\hat{p}_k^{uw}$), of the unweighted estimator. It can be shown that Bias($\hat{p}_k^{uw}$) = $-\mathrm{Cov}(M_i, \hat{p}_{ik})/\overline{M}$, where $\overline{M}$ is mean cluster size, which is independent of sample size n (Helle and Pennington 2004; Pennington and Helle 2011; Deming 1960, p. 379). On average, mean age per cluster in TrawlSurv is negatively correlated with cluster size (Table A2). However, for age-class 1, for example, the proportion-at-age is positively correlated with cluster size, resulting in a negative bias: Bias($\hat{p}_{k=1}^{uw}$) = −0.16 ($\mathrm{Cov}(M_i, \hat{p}_{ik=1})$ = 22.12 and $\overline{M}$ = 137.2). This bias matches the difference in the point estimates of proportion-at-age for age-class 1 for the weighted and the unweighted estimators (Fig. 4). For age-classes 6 through 8 the bias is positive, since the proportions-at-age is negatively correlated with cluster size.

Analyses of TrawlSurv show that the precision in proportions-at-age for NEA cod is largely driven by the number of PSUs (Fig. 5*a*).

Fig. 5. Expected relative standard error (RSE) for the estimated proportions-at-age for NEA cod for varying numbers of (*a*) PSUs and (*b*) number of age samples per 5 cm length bin in each PSU, respectively, based on simulations using data from the joint Russian–Norwegian winter survey in 1991 (TrawlSurv) using the weighted estimator $\hat{p}_k^w$. Each RSE is estimated using 2000 bootstrap replicates. The RSEs for ages 3, 4, 6, and 8 have similar RSEs and are therefore difficult to separate in the figure.



The RSEs of proportions-at-age have not reached their asymptotic limit even for sample sizes as large as 500 trawl stations (Fig. 5*a*), which is prohibitively expensive for the winter survey. The subsample sizes of otoliths per length bin from each PSU have minimal effect on precision (Fig. 5*b*). For NEA cod, one fish per length group (age samples of ~11 fish per PSU) is sufficient to estimate the age composition of the NEA cod from trawl surveys. For the age-classes 1–8, little if any improvement in precision of proportions-at-age is achieved by increasing the number of age samples per length bin from 1 up to 10 (Fig. 5*b*). The mean number of length bins per PSU was ~11 in TrawlSurv, and hence, one age sample per length bin resulted in ~11 age samples per PSU, on average. For age groups 9 and 10 (which account for a marginal proportion of fish in the population), there is noticeable improvement in precision when the sample size is increased from one to four age samples per length bin, but there is little improvement in precision from collecting more than five age samples per length bin. The higher variances in estimates of proportions of fish in age-classes 9 and 10 are largely driven by the few PSU samples (26 and 9 stations, respectively) containing fish in these age-classes. We conducted analysis of several more years of winter survey data, with similar results.

The analyses of ComDat show that there is little gain in precision of proportions-at-age by collecting more than 20 random age samples per PSU (Fig. 6). These analyses furthermore imply that the use of "length-stratified subsampling" also is more effective than simple random subsampling in the collection of age samples from commercial catches. However, random subsampling with number of age samples per PSU being proportional to catch size could improve precision for some age-classes (Fig. 6). Length-stratified subsampling, where a fixed number of age samples are collected per length bin, is easy to implement in the field and provides estimates with similar precision as simple random age samples for age-classes 4–10, but significantly more precise estimates for age-class 3 (Fig. 6). The analyses of ComDat (~10 length bins) show that there is only a small gain in precision from collecting more than one age sample per length bin except for age-class 3, which accounts for a small proportion of the catch.

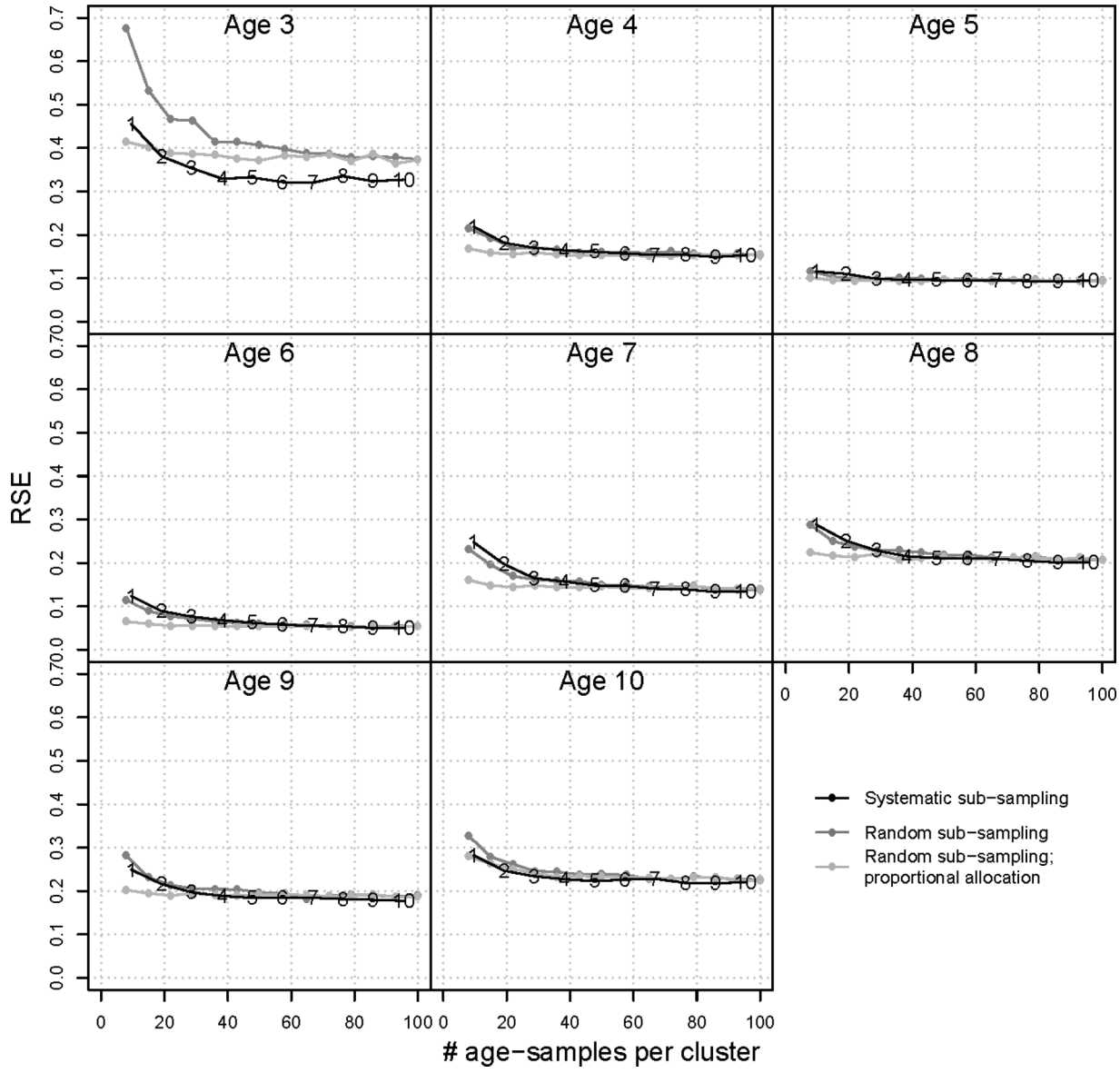
The analyses of TrawlSurv (297 stations with nonzero catches, 3736 age samples overall) provides further support for length-stratified age sampling from as many PSUs as possible. The strategy of subsampling one age sample per 5 cm length bin from all trawl stations with nonzero catches of NEA cod (3237 age samples overall) is the most efficient for all age groups (Fig. 7). The collection of 10 age samples per length bin from 102 stations, in comparison, generally produced less precise estimates of proportion-at-age, although more age samples were collected overall (3736 age samples). The precision in proportions-at-age increased substantially when the 195 extra stations with length-only data were included, especially for younger fish. The reason is that there is less overlap in lengths-at-age for young fish than for older fish that grow slowly.

Discussion

Evaluation of estimators

In general, estimates of age compositions are based on data from multistage sample surveys, and our study suggests that the weighted ratio estimator $\hat{p}_k^w$ is most efficient for such cluster-correlated data. We show that an estimator of proportions-at-age based on properly weighted ALKs is equivalent to a design-based weighted estimator provided that age samples are collected con-

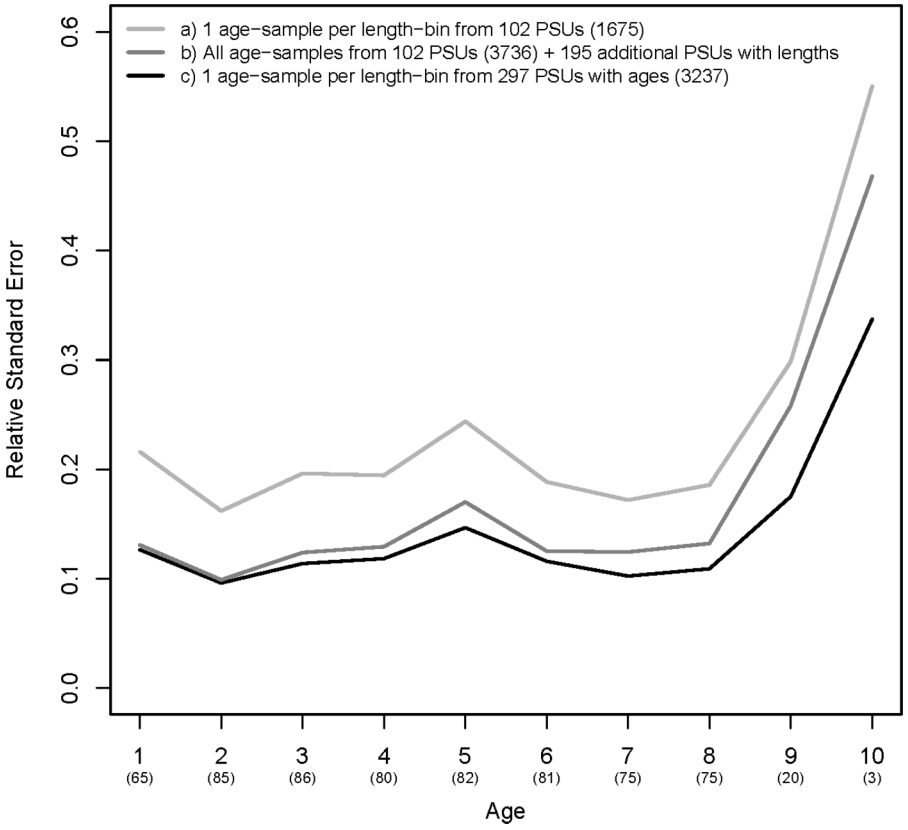
Fig. 6. Expected relative standard error (RSE) for the estimated proportions-at-age for commercial landings of NEA cod for varying number of age samples within each PSU, based on ComDat. Simple random age sampling is compared with systematical subsampling of a fixed number of age samples per 5 cm length bin (length-stratified) and simple random age sampling with sample size proportional to catch size such that the mean number of age samples is according to number of age samples per cluster. The data points numbered from 1 to 10 represent number of age samples per length bin varying from 1 to 10. The mean number of length bins per PSU is ~10, and hence a sample of one fish per 5 cm length bin will result in ~10 age samples. The RSEs are based on the weighted estimator $\hat{p}_k^{w}$ using 2000 bootstrap replicates.



current with the length samples from all PSUs. The variance of estimated proportions-at-age from these weighted estimators can be obtained using bootstrapping methods that replicate the complex multistage survey design, provided that the survey database provides age and length data by PSU. The weighted estimators can also facilitate the estimation of sampling errors in stock assessment parameters. Using bootstrapping to create input data in a statistical assessment model (e.g., Nielsen and Berg 2014), the propagation of sampling errors to estimates of spawning stock biomass and fishing mortality may be quantified.

The weighted estimators $\hat{p}_k^{w}$ and $\hat{p}_k^{w,ALK}$ are preferred because they generally produce estimates with substantially higher accuracy, even though the unweighted estimators can produce substantially more precise estimates. However, our study show that the use of fixed ALKs can severely underestimate the variance in estimates of proportions-at-age since the sampling errors in the ALKs is ignored. The weighted ratio estimator $\hat{p}_k^{w}$ is the best linear unbiased estimator if the relationship between $\hat{p}_{ik}^{w}$ and the cluster size M_i is a straight line through the origin and the variance of M_i about this line is proportional to $\hat{p}_{ik}^{w}$ (Cochran 1977, p. 158). In other cases it is approximately unbiased, although it may yield lower precision than the unweighted estimator $\hat{p}_k^{uw}$. The performance of the weighted estimators agrees with theory and empirical results from complex sample surveys in other research fields. Using four examples from health surveys, for example, Korn and Graubard (1995) demonstrate that unweighted estimators of population parameters using data from multistage cluster sampling can be highly biased, whereas weighted estimators are approximately unbiased but may be more variable than unweighted estimators. In general, hence, there is a trade-off between bias and precision in the choice of unweighted and weighted estimators. For simplic-

Fig. 7. The expected relative standard error (RSE) in estimates of proportions-at-age for NEA cod for different survey designs, based on simulations using data from the joint Russian–Norwegian winter survey in 1991(TrawlSurv). The mean total number of age samples is shown in parentheses in the legends, while number of PSUs with fish of a given age group is provided in parentheses on the *x* axis. The RSEs for panels (*a*) and (*c*) are based on the weighted estimator $\hat{p}_k^{w}$ using the bootstrap method, while the RSEs for panel (*b*) is based on $\hat{p}_k^{w,ALK}$. Each RSE is estimated using 2000 bootstrap replicates.



ity, we presented methods that apply under simple random sampling of PSUs. The estimators and methods for evaluating sampling strategies can easily be extended to more complex survey designs such as stratified sampling of PSUs.

In scientific abundance surveys using bottom trawls with small mesh size, there is often a relationship between cluster size (number of fish in the PSU, M_i) and the size and age structure of fish for many species. This can, for example, be due to catches of abundant young fish in nursery areas being larger, on average, than catches of mature fish in the general survey area. Individual fish of similar size often aggregate together (Pitcher and Parrish 1993; Pennington and Vølstad 1994), and their spatial distribution is often related to size (Milinsky 1993; Osenberg et al. 1994). From theory, it is known that the unweighted estimator $\hat{p}_k^{uw}$ will produce biased estimates when the proportions-at-age for a cluster (PSU) is correlated with cluster size M_i, and this bias will not vanish even for large sample sizes (Cochran 1977, p. 305).

In this paper we focused on the estimation of proportions-at-age. Estimates of numbers-at-age for the target population can be obtained by multiplying the proportions-at-age with an estimate of total number of fish in the target population, and the accuracy will thus be affected by sampling errors and bias in the estimated total number of fish in the target population in addition to sampling errors and bias in estimates of proportions-at-age. A simple approach to estimate numbers-at-age and associated precision is to multiply the proportions-at-age in each PSU with the total number of fish in the PSU and then use the standard design-based estimators and bootstrapping to estimate numbers-at-age with associated variance. The scaling to obtain numbers-at-age estimates should not alter the results presented here on efficiency of estimators and sampling strategies.

Model-based methods based on poststratification can improve efficiency of estimators (Holt and Smith 1979; Valliant 1993). Increased precision in estimates from fishery-dependent surveys, for example, may be achieved by poststratifying fishing trips by area and fishing gear and then adjusting the base weights so that the "sizes" of poststrata or estimation cells are equal to known population totals. Catches and catch composition within poststrata (estimation cells) may be more homogeneous, thus reducing effects of cluster correlation and variable cluster sizes, which can be accommodated using a model-based estimator (see, e.g., Hirst et al. 2012). For fisheries-independent surveys, the total number of fish (poststrata or strata weights) are unknown and must be estimated and becomes a factor in the variance estimates (see Pennington and Vølstad 1994; Folmer and Pennington 2000). In this paper we focused on comparison of estimators, and model-based approaches for improved efficiency is beyond the scope of this paper. Särndal et al. (1978) and Gregoire (1998) provide a good discussion about the differences between design- and model-based inferences.

Borrowing ALKs

Even for commercially important stocks, age samples may be scarce or even unavailable for segments of the target population (i.e., for a particular time period, gear, or spatial area), while extensive length-frequency data may be available. In the estimation of age compositions to support stock assessments, ALKs are therefore often borrowed from another time period, stratum, or gear to

fill data gaps (ICES 2014) and is effectively assumed to be known exactly. This imputation method will produce biased estimates of the age composition unless the ALK is representative for the age composition of the target population (e.g., Quinn and Deriso 1999, pp. 311–317; Hoenig et al. 2002). Gerritsen et al. (2006) and Berg and Kristensen (2012) showed large regional differences in ALKs for a range of species, suggesting that the borrowing of ALKs between regions cause bias (see also Hoenig and Heisey 1987). In the case study for NEA cod, where we estimated the age composition in longline catches using an ALK derived for gillnet catches, it is likely that the ALKs for the total landings from gillnet and longline differ because of differences in gear selectivity. Faster-growing fish are more likely to be caught in gillnet fisheries at a younger age because gillnets in general are highly selective with respect to fish size (e.g., Handford et al. 1977), which is consistent with our findings.

Evaluation of sampling strategies for age sampling based on empirical data

Our NEA cod case study show that the precision in estimates of proportion-at-age largely is driven by the number of PSUs and that little gain in precision of proportions-at-age is achieved by collecting more than 20 random age samples per PSU. This agrees with findings in many other studies (e.g., Aanes and Pennington 2003; Helle and Pennington 2004; Stewart and Hamel 2014). Compared with the actual 102 trawl stations with age samples in the 1991 trawl survey, a fairly big gain in precision has been achieved by increasing the number of trawl stations with age samples to 200 or higher, which has been realized in the winter surveys since 1996. Our analyses furthermore imply that the use of length-stratified subsampling, which is standard in the winter survey, is more effective than simple random subsampling in the collection of age samples. When age samples are collected from a small subsample of fish measured for length, then length-stratified (systematic) age sampling will ensure a better coverage of the age composition than simple random sampling, since age groups that accounts for a small proportion of the catch are more likely to be included. Our study also suggest that the efficiency of the commercial catch sampling could be improved by increasing the number of PSUs and collecting one age sample per length bin from each PSU (see Aanes and Pennington 2003). Subsampling where numbers of age samples per PSU are proportional to the catch sizes (proportional to size) could potentially improve precision for some age-classes. However, we believe that this protocol is likely to be difficult to implement routinely in the field and that the improvement likely will be marginal, since precision is driven by the number of PSUs.

Conclusions

We conclude from this work that the most accurate estimator of proportion-at-age is the weighted design-based ratio estimator and that taking many small age samples from as many PSUs as possible is a better strategy in general than taking more age samples from fewer PSUs. We also show that the common practice of the borrowing of ALKs derived for different strata or fishing gears to fill gaps when data are scarce will underestimate variance and may introduce bias to the age composition estimates. We explicitly show that failing to account for hierarchical sample designs in the data analysis, which is common in the fisheries literature, can underestimate the variance and cause severe bias in age composition estimates.

Acknowledgement

This work was supported by Institute of Marine Research, the Norwegian Research Council through the Statistics for Innovation Centre, and the project ADMAR (Adaptive management of living marine resources by integrating different data sources and key ecological processes; project No. 200497/I30). We are grateful to Michael Pennington, Pat Sullivan, Magne Aldrin, two anonymous reviewers, and the associate editor for valuable comments.

References

Aanes, S., and Pennington, M. 2003. On estimating the age composition of the commercial catch of Northeast Arctic cod from a sample of clusters. ICES J. Mar. Sci. **60**: 297–303. doi:10.1016/S1054-3139(03)00008-0.

Berg, C.W., and Kristensen, K. 2012. Spatial age–length key modelling using continuation ratio logits. Fish. Res. **129**: 119–126. doi:10.1016/j.fishres.2012.06.016.

Booth, J.G., Butler, R.W., and Hall, P. 1994. Bootstrap methods for finite populations. J. Am. Stat. Assoc. **89**: 1282–1289. doi:10.1080/01621459.1994.10476868.

Canty, A., and Ripley, B. 2013. boot: Bootstrap R (S-Plus) functions. R package version 1.3-9.

Cochran, W.G. 1977. Sampling techniques. 3rd ed. John Wiley and Sons, New York.

Davison, A.C., and Hinkley, D.V. 1997. Bootstrap methods and their applications. Cambridge University Press, Cambridge.

Deming, W.E. 1960. Sample design in business research. John Wiley & Sons, New York. ISBN 0-471-52370-4.

Efron, B. 1982. The jackknife, the bootstrap and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, Pa.

Fleiss, J.L., Levin, B., and Paik, M.C. 2003. Statistical methods for rates and proportions. 3rd ed. John Wiley & Sons, Hoboken, N.J.

Folmer, O., and Pennington, M. 2000. A statistical evaluation of the design and precision of the shrimp trawl survey off West Greenland. Fish. Res. **49**: 165–178. doi:10.1016/S0165-7836(00)00196-X.

Fridriksson, A. 1934. On the calculation of age-distribution within a stock of cod by means of relatively few age-determinations as a key to measurements on a large scale. Rapports et Procès-verbaux des Réunions du Conseil Permanent International pour l'Exploration des Mers, **86**: 1–14.

Gerritsen, H.D., McGrath, D., and Lordan, C. 2006. A simple method for comparing age–length keys reveals significant regional differences within a single stock of haddock (*Melanogrammus aeglefinus*). ICES J. Mar. Sci. **63**: 1096–1100. doi:10.1016/j.icesjms.2006.04.008.

Gregoire, T.G. 1998. Design-based and model-based inference in survey sampling: appreciating the difference. Can. J. For. Res. **28**(10): 1429–1447. doi:10.1139/x98-166.

Handford, P., Bell, G., and Reimchen, T. 1977. A gillnet fishery considered as an experiment in artificial selection. J. Fish. Res. Board Can. **34**(7): 954–961. doi:10.1139/f77-148.

Helle, K., and Pennington, M. 2004. Survey design considerations for estimating the length composition of the commercial catch of some deep-water species in the Northeast Atlantic. Fish. Res. **70**: 55–60. doi:10.1016/j.fishres.2004.06.011.

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. Can. J. Fish. Aquat. Sci. **69**(12): 2064–2076. doi:10.1139/cjfas-2012-0075.

Hoenig, J.M., and Heisey, D.M. 1987. Use of a log-linear model with the EM algorithm to correct estimates of stock composition and to convert length to age. Trans. Am. Fish. Soc. **116**: 232–243. doi:10.1577/1548-8659(1987)116<232:UOALMW>2.0.CO;2.

Hoenig, J.M., Choudary Hanumara, R., and Heisey, D.M. 2002. Generalizing double and triple sampling for repeated surveys and partial verification. Biomet. J. **44**: 603–618. doi:10.1002/1521-4036(200207)44:5<603::AID-BIMJ603>3.0.CO;2-4.

Holt, D., and Smith, T.M.F. 1979. Post stratification. J. Royal Stat. Soc. Series A, **142**(1): 33–46.

ICES. 2013. Report of the second Workshop on Practical Implementation of Statistical Sound Catch Sampling Programmes, 6–9 November 2012, ICES Copenhagen. ICES CM2012/ACOM:54.

ICES. 2014. Report of the Third Workshop on Practical implementation of Statistical Sound Catch Sampling Programs (WKPICS3), 19–22 November 2013, ICES Copenhagen. ICES CM2013/ACOM:54.

Jessen, R.J. 1978. Statistical survey techniques. John Wiley & Sons, New York.

Kimura, D.K. 1977. Statistical assessment of the age–length key. J. Fish. Res. Board Can. **34**(3): 317–324. doi:10.1139/f77-052.

Kimura, D.K., and Chikuni, S. 1987: Mixtures of empirical distributions: an iterative application of the age–length key. Biometrics, **43**: 23–35. doi:10.2307/2531945.

Korn, E.L., and Graubard, B.I. 1995. Examples of differing weighted and unweighted estimates from a sample survey. Am. Stat. **49**: 291–295. doi:10.1080/00031305.1995.10476167.

Lai, H.L. 1987. Optimum allocation for estimating age composition using age-length key. Fish. Bull. **85**(2): 179–186.

Lai, H.L. 1993. Optimal sampling design for using the age-length key to estimate age composition of a fish population. Fish. Bull. **91**: 382–388.

Lavallée, P. 2007. Indirect sampling. Springer Series in Statistics. Springer.

Lehtonen, R., and Pahkinen, E. 2004. Practical methods for design and analysis of complex surveys. John Wiley & Sons.

Milinsky, M. 1993. Predation risk and feeding behavior. *In* Behavior of teleost fishes. 2nd ed. *Edited by* T.J. Pitcher. Chapman and Hall, Fish and Fisheries Series 7, New York. pp. 285–306.
Morton, R., and Bravington, M. 2008. Comparison of methods for estimating age composition with application to Southern Bluefin Tuna (*Thunnus maccoyii*). Fish. Res. **93**: 22–28. doi:10.1016/j.fishres.2008.02.009.
Nelson, G.A. 2014. Cluster Sampling: a pervasive, yet little recognized survey design in fisheries research. Trans. Am. Fish. Soc. **143**: 926–938. doi:10.1080/00028487.2014.901252.
Nielsen, A., and Berg, C.W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. **158**: 96–101. doi:10.1016/j.fishres.2014.01.014.
Osenberg, C.W., Olson, M.H., and Mittelbach, G.G. 1994. Stage structure in fishes: resource productivity and competition gradients. *In* Theory and application in fish feeding ecology. *Edited by* D.J. Stouder, K.L. Fresh, and R.J. Feller. University of South Carolina Press, Belle W. Baruch Library in Marine Science 18, Columbia, S.C. pp. 151–170.
Pennington, M., and Helle, K. 2011. Evaluation of the design and efficiency of the Norwegian self-sampling purse-seine reference fleet. ICES J. Mar. Sci. **68**: 1764–1768. doi:10.1093/icesjms/fsr018.
Pennington, M., and Vølstad, J.H. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from trawl surveys. Biometrics, **50**: 725–732. doi:10.2307/2532786.
Pennington, M., Shevelev, M.S., Vølstad, J.H., and Nakken, O. 2011. Bottom trawl surveys. *In* The Barents Sea. Ecosystem, resources, management. Half a century of Russian–Norwegian cooperation. *Edited by* T. Jakobsen and V.K. Ozhigin. Tapir Academic Press, Trondheim. pp. 570–584.
Pitcher, T.J., and Parrish, J.K. 1993. Functions of shoaling behavior in teleosts. *In* Behavior of teleost fishes. 2nd ed. *Edited by* T.J. Pitcher. Chapman and Hall, Fish and Fisheries Series 7, New York. pp. 363–439.
Quinn, T.J., and Deriso, R.B. 1999. Quantitative fish dynamics. Oxford University Press, New York.
Särndal, C.E., Thomsen, I., Hoem, J.M., Lindley, D.V., Barndorff-Nielsen, O., and Dalenius, T. 1978. Design-based and model-based inference in survey sampling [with discussion and reply]. Scan. J. Stat. **5**: 27–52.
Särndal, C.E., Swensson, B., and Wretman, J. 1992. Model assisted survey sampling. Springler-Verlag, New York.
Sen, A.R. 1986. Methodological problems in sampling commercial rockfish landings. Fish. Bull. **84**(2): 409–421.
Stewart, I.J., and Hamel, O.S. 2014. Bootstrapping of sample sizes for length- or age-composition data used in stock assessments. Can. J. Fish. Aquat. Sci. **71**(4): 581–588. doi:10.1139/cjfas-2013-0289.
Tanaka, S. 1953. Precision of age-composition of fish estimated by double sampling method using the length for stratification. Bull. Jpn. Soc. Sci. Fish. **19**(5): 657–670.
Thompson, M.E. 1997. Theory of sample surveys. Chapman and Hall, London.
Valliant, R. 1993. Poststratification and conditional variance estimation. J. Am. Stat. Assoc. **88**: 89–96. doi:10.1080/01621459.1993.10594298.
Westrheim, S.J., and Ricker, W.E. 1978. Bias in using an age–length key to estimate age-frequency distributions. J. Fish. Res. Board Can. **35**(2): 184–189. doi:10.1139/f78-030.
Williams, R.L. 2000. A note on robust variance estimation for cluster-correlated data. Biometrics, **56**: 645–646. doi:10.1111/j.0006-341X.2000.00645.x.

Appendix A. Simulating trawl survey data

To build a simulation model that mimics empirical data for trawl surveys, we take the following approach. On each trawl station (PSU), $i = 1, \ldots, n$, the age composition $\{p_{ik}\}_{k=1,\ldots,A}$ is estimated applying the standard estimator for length-stratified data yielding $\{\hat{p}_{ik}\}_{k=1,\ldots,A}$. The estimated cluster-specific proportions-at-age are transformed by the multinomial logit transformation to convert proportions to real numbers (i.e., $x_{ik} = \ln(\hat{p}_{ik}/\hat{p}_{iB})$ for clusters $i = 1, \ldots, n$ and $k = 1, \ldots, A-1$ for a reference age category B). We chose B as the highest age in the data. The cluster-specific correlation structure is estimated by correlating age proportions within cluster with distance in age (see also Hrafnkellson and Stefánsson 2004 for a similar approach to estimate the empirical correlation structure). We assume that the correlation structure is independent of age category and cluster and only depends on distance in age between proportions within cluster (analogue to isotropic correlation function in spatial statistics; see, e.g., Cressie 1993). To estimate the mean correlation structure for the population, a linear model is fitted to the Fisher-transformed (e.g., Larsen and Marx 1986, p. 488) values of the empirical correlations, r, as a function of absolute difference in age $|k - j| = h$; i.e., $z = \phi + \gamma h$, where $z = 0.5\ln\left(\frac{1+r}{1-r}\right)$. The parameters ϕ and γ are replaced by the least-squares estimates using the data and thus specifies the entire correlation matrix with elements $\text{cor}(x_k, x_j)$, $k, j = 1, \ldots, A-1$. The elements of the covariance matrix $\boldsymbol{\Sigma}_x$, $\text{cov}(x_k, x_j)$, is then specified using the empirical age-specific variances $\text{SD}(x_k) = \sqrt{\text{var}(x_k)}$ and correlations $\text{cor}(x_k, x_j)$ through the relationship, $\text{cov}(x_k, x_j) = \text{cor}(x_k, x_j)\text{SD}(x_k)\text{SD}(x_j)k, j = 1, \ldots, A-1$. To complete the model for the process-generating cluster-specific age vectors, we assume the vector of cluster-specific transformed age proportions $\boldsymbol{x}_i = \{x_{ik}\}_{k=1,\ldots,A-1}$ is multivariate normally (MVN) distributed with mean $\boldsymbol{\mu}_x = \{\mu_{x,k}\}_{k=1,\ldots,A-1}$ and covariance $\boldsymbol{\Sigma}_x$ (i.e., $\boldsymbol{x}_i \overset{iid}{\sim} \text{MVN}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$). Values for $\boldsymbol{\mu}_x$ are replaced by the empirical means obtained by Fisher-transformed estimates of $\hat{p}_{ik}$ across PSUs for each age. Sampling from $\boldsymbol{x}_i \overset{iid}{\sim} \text{MVN}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ for $i = 1, \ldots, N$ thus defines the PSU-specific age composition for the population formed by N PSUs.

To specify the simultaneous age and length distribution, a model for size-at-age is needed. The assumption of equal growth for all fish across all clusters is often considered to be realistic (see, e.g., Hoenig and Heisey 1987; i.e., equal probability of fish being in a specific length bin given the age category). We use the von Bertalanffy growth function (see e.g., Quinn and Deriso 1999) fitted to a large data set for cod to distribute the individuals to length given age within a cluster. This is achieved by sampling lengths independently from the model $\log(l|a) \sim N\{\log[L_\infty(1 - e^{-\kappa a})], \sigma_l^2\}$ (see Table A1).

To generate realistic cluster sizes accounting for potential dependencies between age structure and cluster size, we use the

Table A1. Parameter estimates for the parameters in the von Bertalanffy growth model fitted to all individual cod measurements of age–length data in Institute of Marine Research's database for the years 1990–2000.

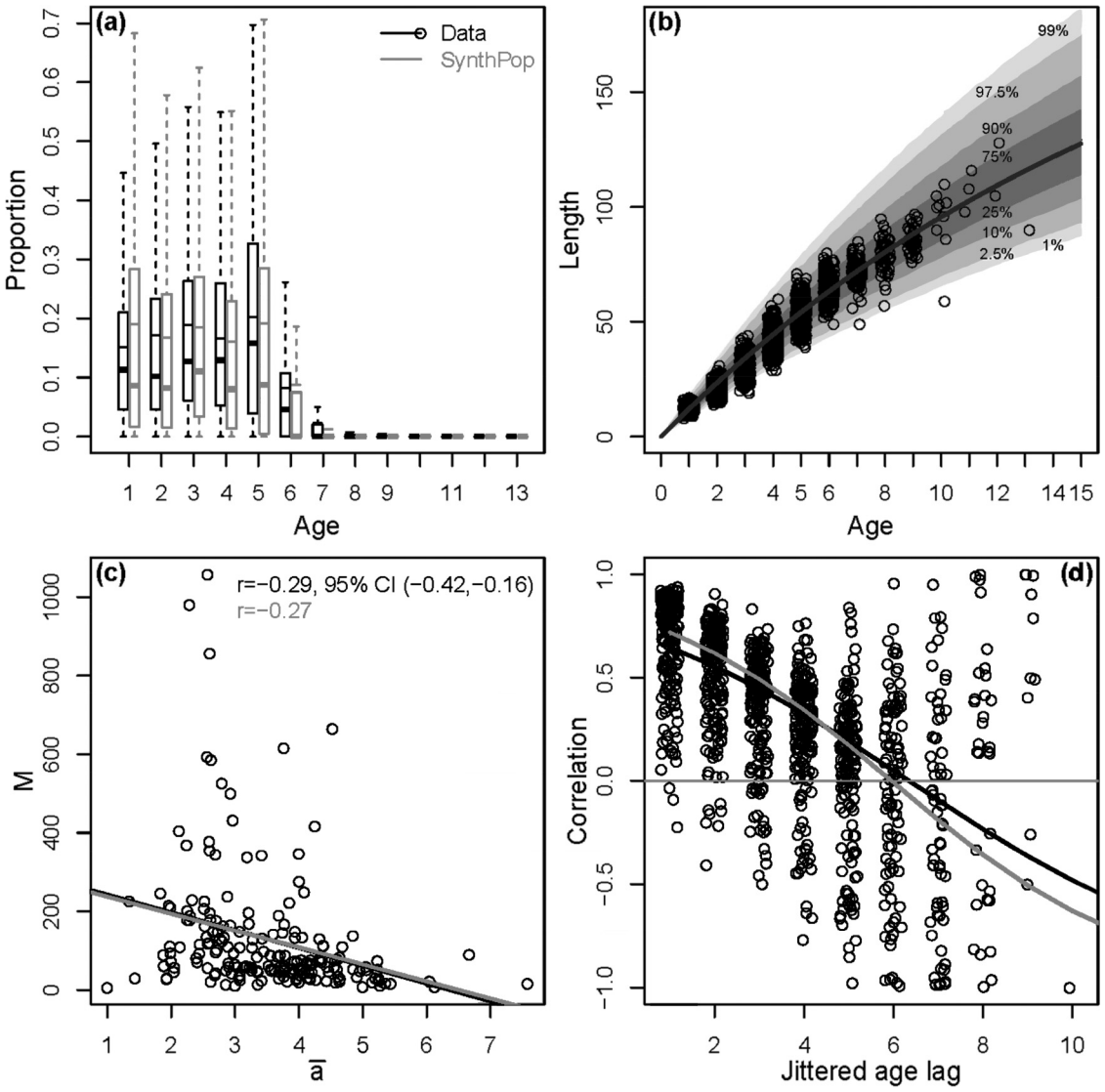
L_∞	κ	σ_l
232.9803	0.0523	0.1618

Note: The growth model is $l_i|a_i, L_\infty, \kappa = L_\infty(1 - e^{-\kappa a_i})e^{\varepsilon_i}$, where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_l^2)$ and is fitted to individual data by maximizing the likelihood with respect to L_∞, κ, and σ_l.

Table A2. Correlation between mean age ($\bar{a}$) per cluster and cluster size M_i for the empirical data for Northeast Arctic (NEA) cod (*Gadus morhua*) from the joint Norwegian–Russian winter survey for the years 1990 through 2011.

Year	Corr(M, $\bar{a}$)	95% confidence interval
1990	0.06	(−0.16, 0.27)
1991	−0.46	(−0.59, −0.29)
1992	−0.24	(−0.41, −0.05)
1993	−0.20	(−0.37, −0.02)
1994	−0.34	(−0.47, −0.18)
1995	−0.33	(−0.46, −0.19)
1996	−0.33	(−0.43, −0.22)
1997	−0.26	(−0.39, −0.11)
1998	−0.36	(−0.48, −0.23)
1999	−0.30	(−0.48, −0.10)
2000	−0.29	(−0.42, −0.16)
2001	−0.27	(−0.37, −0.16)
2002	−0.21	(−0.31, −0.11)
2003	−0.24	(−0.36, −0.12)
2004	−0.13	(−0.25, −0.01)
2005	−0.23	(−0.34, −0.11)
2006	−0.21	(−0.32, −0.09)
2007	−0.21	(−0.32, −0.09)
2008	−0.20	(−0.32, −0.07)
2009	−0.07	(−0.19, 0.06)
2010	−0.12	(−0.23, −0.01)
2011	0.04	(−0.07, 0.15)
Mean	−0.24	

Fig. A1. Summary of data on age composition of NEA cod from the joint Russian–Norwegian winter survey in 2000 used to parameterize the synthetic population of PSUs SynthPop: (*a*) the empirical distribution of proportions-at-age across clusters by boxplots, (*b*) the functional relationship between length and age used to determine the length and age distributions of fish in the simulated synthetic populations of clusters, where percentiles are given by the different shades, (*c*) the correlation between and mean age ($\bar{a}$) and cluster size (M) by PSU, and (*d*) the within-cluster correlation structure in proportions-at-age, shown as the correlation by distance in age (age lag). The solid line is the fitted correlation coefficient based on Fisher *z*-transformation using age lag as covariate (see text for details). The resulting synthetic population is shown in gray. The actual data points representing single hauls are added to panels (*b*) through (*d*). Note that for panel (*c*), the black and gray lines are overlapping and are difficult to distinguish.



empirical relationship between cluster size and mean age through $\log(M_i) = \alpha + \beta \log(\bar{a}_i) + \varepsilon_i$, where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_M^2)$ and α, β, and σ_M^2 are replaced by least-squares estimates. A log-linear relationship is chosen simply because it fits the data better than a linear relationship on the original scale. That is, in one realization the cluster size from cluster i is simulated from $N(\alpha + \beta \log(\bar{a}_i), \sigma_M^2)$, where $\bar{a}_i = \sum_{j=1}^{A} j p_{ij}$ is available from the cluster-specific age proportions.

Then proportion-at-age $p(a)$, proportion-at-length given age $p(l|a)$, and hence the simultaneous distribution $p(a, l)$, the ALK $p(a|l)$, and the length distribution $p(l)$ are defined by standard application of conditional probabilities. Given all cluster sizes, all clusters are determined with numbers-at-age and length. To avoid decimal counts (which will appear since the model for cluster sizes not is defined as a discrete distribution and since numbers-at-age and length are derived by multiplying numbers and proportions), all numbers are rounded to the nearest integer to fully specify cluster-specific population counts. Rounding of a high dimensional vector of proportions-at-age and length may cause large rounding errors compared with the total given by the cluster size. Therefore, the remainder after rounding is distributed randomly according to the simultaneous multinomial distribution of numbers-at-age and length.

The population characteristics are established by generating a large number of PSUs by the above approach (we use $N = 100\,000$), which serves as the sampling frame for a trawl survey. We use data from the joint Russian–Norwegian winter survey from 2000 to parameterize the simulation model, and the resulting synthetic population produces realistic population characteristics: The fairly large variability in age compositions among clusters is

fairly typical for the winter survey (Fig. A1*a*), and the length-at-age relationship for the synthetic population is similar to the empirical age–length observations of NEA cod from the 2000 survey (Fig. A1*b*). The mean correlation ($\bar{r}$ = –0.29 with 95% confidence interval (–0.42, –0.16)) between mean age ($\bar{a}$) per cluster and cluster size (M_i) is close to the mean correlation for the time series from the winter trawl surveys ($\bar{r}$ = –0.24; Table A2; Fig. A1*c*). Proportions-at-age for closely spaced age groups are highly positively correlated, on average, and the mean correlation decreases and remains positive up to an age difference of 4 years and turns negative for age groups that are 5 years or more apart (Fig. A1*d*).

References

Cressie, N. 1993. Statistics for spatial data. Wiley, New York.

Hoenig, J.M., and Heisey, D.M. 1987. Use of a log-linear model with the EM algorithm to correct estimates of stock composition and to convert length to age. Trans. Am. Fish. Soc. **116**: 232–243. doi:10.1577/1548-8659(1987)116<232:UOALMW>2.0.CO;2.

Hrafnkelsson, B., and Stefánsson, G. 2004. A model for categorical length data from groundfish surveys. Can. J. Fish. Aquat. Sci. **61**(7): 1135–1142. doi:10.1139/f04-049.

Larsen, R.J., and Marx, M.L. 1986. An Introduction to mathematical statistics and its applications. Englewood Cliffs, New Jersey.

Quinn, T.J., and Deriso, R.B. 1999. Quantitative fish dynamics. Oxford University Press, New York.