Quantifying age-reading error for use in fisheries stock assessments, with application to species in Australia's southern and eastern scalefish and shark fishery

André E. Punt, David C. Smith, Kyne KrusicGolub, and Simon Robertson

Abstract: Age-reading error occurs when estimates of age based on reading hard structures differ from the true age of the animal concerned. This error needs to be accounted for when conducting stock assessments. Common methods for quantifying age-reading error include the average percent error, the coefficient of variation, age bias plots, and age difference tables, but these techniques cannot be used to construct age-reading error matrices. A method for constructing age-reading error matrices that accounts for both ageing bias and ageing imprecision is outlined. Simulation evaluation of this method suggests that it is able to estimate both ageing bias (assuming that one reader is unbiased) and ageing imprecision for relatively large sample sizes and for the ages that constitute the bulk of the ages in the sample. However, the performance of the method is poor when sample sizes are small, age-reading error is correlated among readers, when both readers are biased, and for ages that are poorly represented in the sample. The method is applied for illustrative purposes to data on multiple-aged fish in Australia's southern and eastern scalefish and shark fishery.

Résumé : Les erreurs de lecture de l'âge se produisent lorsque les estimations de l'âge basées sur la lecture de structures dures diffèrent de l'âge réel de l'animal en question. Il est nécessaire de tenir compte de cette erreur dans les évaluations de stocks. Les méthodes courantes pour quantifier l'erreur de lecture de l'âge comprennent le pourcentage moyen d'erreur, le coefficient de variation, les graphiques d'erreur en fonction de l'âge et les tables de différences d'âge, mais ces méthodes ne peuvent pas servir à contruire des matrices d'erreurs de lecture de l'âge. Nous décrivons une méthode pour contruire des matrices d'erreurs de lecture de l'âge qui tiennent compte autant de l'erreur que de l'imprécision de la détermination de l'âge. Une évaluation de la méthode par simulation indique qu'elle est capable d'estimer tant l'erreur de la lecture d'âge (si on suppose qu'un des lecteurs n'est pas biaisé) que l'imprécision de la détermination d'âge pour des échantillons de taille relativement élevée et pour les âges qui constituent la majorité des âges dans l'échantillon. Cependant, la performance de la méthode est médiocre lorsque l'échantillon est petit, lorsqu'il y a corrélation de l'erreur de lecture entre les lecteurs, lorsque les deux lecteurs sont biaisés et pour les âges qui sont peu représentés dans l'échantillon. Comme exemple, nous appliquons la méthode à des données de poissons d'âges multiples obtenues des pêches commerciales de poissons à écailles et de requins du sud et de l'est de l'Australie.

[Traduit par la Rédaction]

Introduction

Quantitative stock assessments that involve fitting population dynamics models to data collected from fishery-dependent and -independent sampling programs are an important component of the basis for the scientific management advice for many of the world's major fisheries. Many of the stock assessments conducted for marine species off New Zealand, Australia, South Africa, and the west coast of North America are currently based on the integrated analysis paradigm (e.g., Fournier and Archibald 1982; Methot 2000, 2007). Integrated analysis assessments can make use of a wide range of data types, including indices of relative and absolute abundance, catch and survey age and size composition information, and tagging data. Central to the use of age composition data in stock assessments based on integrated analysis is the relationship between the ages obtained by reading hard structures such as otoliths and spines and the

Received 26 September 2007. Accepted 12 March 2008. Published on the NRC Research Press Web site at cjfas.nrc.ca on 4 September 2008.
J20197

A.E. Punt.[1] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195-5020, USA; CSIRO Marine and Atmospheric Research, GPO, Hobart, TAS 7001, Australia.
D.C. Smith. School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98195-5020, USA.
K. KrusicGolub and S. Robertson. Marine and Freshwater Systems, Department of Primary Industries, Queenscliff, VIC 3225, Australia.

[1]Corresponding author (e-mail: andre.punt@csiro.au).

true age of the animal. There are two sources of uncertainty in this relationship, bias and imprecision, both of which occur regularly. Ageing bias occurs when there is a systematic difference between the true age of an animal and the age assigned to it, whereas ageing imprecision occurs when age-reading errors occur at random. Reeves (2003) examined the impact of age-reading error on the performance of stock assessments based on extended survivors analysis (Shepherd 1999), an assessment framework that ignores age-reading error. Reeves (2003) found that although estimates of spawning stock biomass and fishing mortality were broadly correct, those of recruitment were smoothed out and those of total allowable catch under a control rule that approximates how International Council for the Exploration of the Sea advice is provided were too optimistic in the presence of age-reading error. The smoothing of recruitment estimates leads to difficulties when estimating a stock–recruitment relationship, quantifying the risk associated with future management strategies, and determining the impact of environmental factors on recruitment (Fournier and Archibald 1982; Richards et al. 1992).

In principle, age-reading error can be represented using a matrix that specifies the probability of an animal of true age a being aged to be that age or some other age a', $P(a'|a)$. The model predictions upon which the likelihood component in the stock assessment for the age composition data is based are then a function of the model estimate for the observed catch of animals of age a after accounting for age-reading error. Given $P(a'|a)$, this prediction would be

$$(1) \qquad C_{a'} = \sum_a P(a'|a)C_a$$

where C_a is the model estimate of the catch of animals of age a, and $C_{a'}$ is the model estimate of the catch of animals of (perceived) age a' after accounting for age-reading error.

The ability to take account of age-reading error is already included in several of the most commonly applied stock assessment packages, e.g., stock synthesis (Methot 2000, 2007), Coleraine (Hilborn et al. 2003), and CASAL (Bull et al. 2003), but these packages do not include the facility to estimate age-reading error matrices.

The extent of age-reading imprecision and bias is regularly assessed using measures such as the average percent error (APE) (Beamish and Fournier 1981), the coefficient of variation, age bias plots (Campana et al. 1995), and age difference tables (Morison et al. 1998), but these measures cannot be used to construct age-reading error matrices for use in stock assessments because they focus on either precision or bias but not both. Richards et al. (1992) developed a method for estimating the parameters of functions that could be used to model the relationship between true and estimated age and hence construct the function $P(a'|a)$. This method, which has unfortunately not been used much in practice, perhaps because of its relative complexity, is based on maximum likelihood and allows for considerable flexibility in the relationship between (true) age and the expectation and imprecision of the estimate of this age.

This paper first conducts a simulation study to determine the performance of an extension of the method developed by Richards et al. (1992) given different sample sizes (numbers of animals read by two or more readers) and the structure of the relationship between true and estimated age for the particular case of silver warehou (*Serioellela punctata*) (previously also known as spotted warehou). The ability of model selection methods based on the Bayesian information criterion (BIC) (Schwarz 1978) to correctly detect the nature of age-reading error is also assessed for this species. It then applies the method to estimate age-reading error for another seven species in Australia's southern and eastern scalefish and shark fishery.

Materials and methods

Data and ageing protocol

The data used in the analyses of this paper were obtained from the Central Ageing Facility (CAF) (Queenscliff, Victoria, Australia), a specialist fish-ageing laboratory established in Australia during the early 1990s (Morison et al. 1998). The CAF is a fully integrated processing and ageing unit that has provided age estimates for more than 200 species ranging from temperate to tropical species and from freshwater species to species found on the continental slope and on seamounts.

Otoliths are collected from the southern and eastern scalefish and shark fishery throughout the year, and the entire year's sample is read after all of the otoliths have been collected. This provides for efficiency and requires fewer calibration checks of readers. It helps maintain consistency by reducing the likelihood of a change or drift over time in interpretation of the incremental structure. APE, age bias plots, and age difference tables are used qualitatively to determine bias and precision. Specifically, if a new reader is assigned to a species or a substantial amount of time has elapsed since a particular species has been last aged (maximum >1 month), the reader reads a subset of 100 otoliths from a calibration set (previously aged otoliths from various years) and APE and age bias plots are examined. If the APE is <5% and there is no evidence for bias, then the reader is allowed to continue. As such, an APE >5% is considered to indicate poor precision, an extremely difficult species to age, or insufficient training of new readers. A more detailed description of the ageing procedures used by the CAF can be found in Morison et al. (1998).

Data from eight species are used in this paper: silver warehou, blue grenadier (*Macruronus novaezelandiae*), school whiting (*Sillago flindersi*), tiger flathead (*Platycephalus richardsoni*), deepwater flathead (*Platycephalus conatus*), pink ling (*Genypterus blacodes*), gemfish (*Rexea solandri*), and black bream (*Acanthopagrus butcheri*). The analyses for school whiting and gemfish are based on whole otolith readings, while those for the remaining species (except silver warehou) are based on reading of sectioned otoliths. The analyses for silver warehou also include readings based on whole otoliths to illustrate how ageing bias can be quantified.

The total number of samples aged at the CAF for each species examined in this study and the numbers of repeat reads (secondary and tertiary) by reader available for inter- and intra-rereader comparisons are provided in Table 1. The reader index (A, B, C, etc.) in Table 1 refers to specific readers, not necessarily the primary, secondary, or tertiary age estimate. The primary reads of otoliths for a new species

Table 1. Numbers of samples aged at the Central Ageing Facility for each species examined in this study and the numbers of repeat reads (secondary and tertiary) by reader.

			Maximum	Reader A			Reader B			Reader C			Other readers			
Scientific name	Family	Habitat	age (years)	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd	Whole
Seriolella punctata	Centrolophidae	Marine, temperate	23		36		2703	1407	31							1480
Macruronus novaezelandiae	Merluciidae	Marine, temperate	25	940						4520	4512					940
Sillago flindersi[a]	Sillaginidae	Marine, temperate	8	160				184					758	719	122	
Platycephalus richardsoni	Platycephalidae	Marine, temperate	25		101		541	276		420	332	72	314	351	269	169
Platycephalus conatus	Platycephalidae	Marine, temperate	33		88	44	398	334			86	21	529	460	257	
Genypterus blacodes	Ophidiidae	Marine, temperate	28	1515	678	27	998	1615	18		311		448	438	13	
Rexea solandri[a]	Gempylidae	Marine, temperate	17	543	659	133	176	164		40	97		388	488	50	
Acanthopagrus butcheri	Sparidae	Estuarine to marine	39	738	159		1178	853		110			1498	951		

Note: 1st, primary read; 2nd, secondary read; 3rd, tertiary read.
[a]All age estimates are based on readings of whole otoliths.

or those read by an inexperienced reader are reread by an experienced secondary reader to determine intrareader precision and variability. This is indicated in Table 1, where only secondary and tertiary readings are available for some of the readers for some of the species.

Since its inception in 1991, CAF has employed more than 10 different readers. It is rarely the case that a species will be aged by only one primary reader (indicated in Table 1 by multiple readers providing the primary reads). Silver warehou is an exception; reader 2 has been the primary reader for this species since age estimates started being obtained from sectioned otoliths.

Analysis technique

The functional form used to model the probability of reader i (of I readers) assigning an animal of true age a an age of a', $P^i(a'|a)$, can be very general but needs to satisfy the constraint that $\sum_{a'} P^i(a'|a) = 1$. For example, assuming that (*i*) ageing bias depends on reader and the true age of an animal, (*ii*) the age-reading error standard deviation depends on true age and reader, and (*iii*) age-reading error is normally distributed about the expected age (i.e., the expected age given any bias in age reading) leads to the following model for $P^i(a'|a)$:

$$(2)\qquad P^i(a'|a,\boldsymbol{\varphi}) \propto \exp\left[\frac{-(a' - b_a^i(\varphi))^2}{2(\sigma_a^i(\varphi))^2}\right]$$

where b_a^i is expected age when reader i determines the age of an animal of true age a, σ_a^i is the standard deviation for reader i of the age-reading error for animals of true age a, and $\boldsymbol{\varphi}$ is the vector of parameters that determines the age-reading error matrix.

For the analyses of this paper, the probabilities obtained from eq. 2 are set to zero for values of $a' < 0$ and larger than a prespecified maximum age. Following Richards et al. (1992) and Heifetz et al. (1998), and for illustrative purposes, the age-reading error standard deviation and ageing bias are modeled in this paper using the following three-parameter models:

$$(3)\qquad \sigma_a = \begin{cases} \sigma_L + (\sigma_H - \sigma_L)\dfrac{1 - e^{-\alpha(a-L)}}{1 - e^{-\alpha(H-L)}} & \text{if } \alpha \neq 0 \\ \sigma_L + (\sigma_H - \sigma_L)\dfrac{a-L}{H-L} & \text{if } \alpha = 0 \end{cases}$$

where σ_L is the age-reading error standard deviation for a prespecified minimum age L, σ_H is the age-reading error standard deviation for a prespecified maximum age H, and α determines the extent of nonlinearity between age and the age-reading error standard deviation and

$$(4)\qquad b_a = \begin{cases} b_L + (b_H - b_L)\dfrac{1 - e^{-\lambda(a-L)}}{1 - e^{-\lambda(H-L)}} & \text{if } \lambda \neq 0 \\ b_L + (b_H - b_L)\dfrac{a-L}{H-L} & \text{if } \lambda = 0 \end{cases}$$

where b_L is the expected age of an animal of true age L, b_H is the expected age of an animal of true age H, and λ deter-

mines the extent of nonlinearity between age and the expected age.

The ageing techniques for which bias and imprecision are to be quantified can be based on a variety of ageing structures (including spines, otoliths, or scales). The values for the parameters that determine the age-reading error matrix for each age reader are estimated by maximizing the following likelihood function:

$$(5) \qquad L(\mathbf{A}|\boldsymbol{\beta}, \boldsymbol{\varphi}) = \prod_{j=1}^{J} \sum_{a=L}^{H} \beta_a \prod_{i=1}^{I} P^i(a_{i,j}|a, \boldsymbol{\varphi})$$

where $a_{i,j}$ is the age assigned by reader i to the jth ageing structure (of J ageing structures), and $\mathbf{A}$ is the entire data set of age readings.

The βs are nuisance parameters that can be interpreted as the relative frequency of animals of (true) age a in the sample (rather than in the population from which the sample was taken). An additional term can be added to eq. 5 to make use of data for known-age animals, if such data are available (Heifetz et al. 1998). In general, not all ageing structures are read by all readers. Therefore, the likelihood function is more generally the product of eq. 5 over sets of ageing structures that were all read by the same group of readers and a separate set of βs is estimated for each such set of ageing structures.

Simulation evaluation

A simulation study is used to evaluate the performance of the approach outlined above. The aim of the simulation study was to determine how often the correct ("true") model of age-reading error is selected (which age reader is biased and the functional form for the relationship between true age and bias and that between true age and age-reading error standard deviation), as well as the extent of bias and imprecision associated with the outputs from the model.

The simulation study involved developing a true model based on one specific model of age-reading error and applying several variants of the approach outlined above to data sets that could have been observed had that "true" model been correct. Specifically, the simulation study was based on 250 replicates, each of which involves a single data set with two readers, one of whom (reader 2) produces biased age readings and both of whom make age readings subject to random age-reading error (Figs. 1*a* and 1*b*). This is equivalent to one age reader reading different structures (i.e., reader 1 could be age estimates based on sectioned otoliths, while reader 2 could be age estimates by the same reader based on whole otoliths). The true age structure of the population sampled (Fig. 1*c*) is based on the results of applying eqs. 2, 3, 4, and 5 to the data for silver warehou. The variants of the estimator considered in the simulation study are listed in Table 2. Variants 1–4 form the focus for the simulation study (variant 3 is the "true" model given the structure of the simulations), while the remaining variants form the basis for tests of sensitivity. In general, model selection criteria (such as Akaike's information criterion (AIC) and BIC) are used to select among alternative variants when the method is applied in practice, so the simulations also examine how well the BIC can select among the variants. The calculation of BIC was based on the number of parameters in the model being those of eqs. 3 and 4, as well as the number of nuisance parameters that define the true age structure of the sample. BIC was preferred to AIC (Burnham and Anderson 2002) for these simulation analyses (and when it was applied to the actual data for fish species in Australia's southern and eastern scalefish and shark fishery) because AIC tends to select the most complicated model, even when it is not the true model, because of the large number of data points.

Fig. 1. Specifications of the simulations used to evaluate the performance of the estimator of age-reading error: (*a*) ageing bias for the two readers (reader 1, solid line; reader 2, broken line), (*b*) age-reading error standard deviation, and (*c*) relative frequency of different ages in the sampled population. The solid circles in the bottom panel highlight the ages that form the focus for the results of the simulation study. The dotted lines in the upper panel indicate the 95% intervals for an age estimate for each reader.

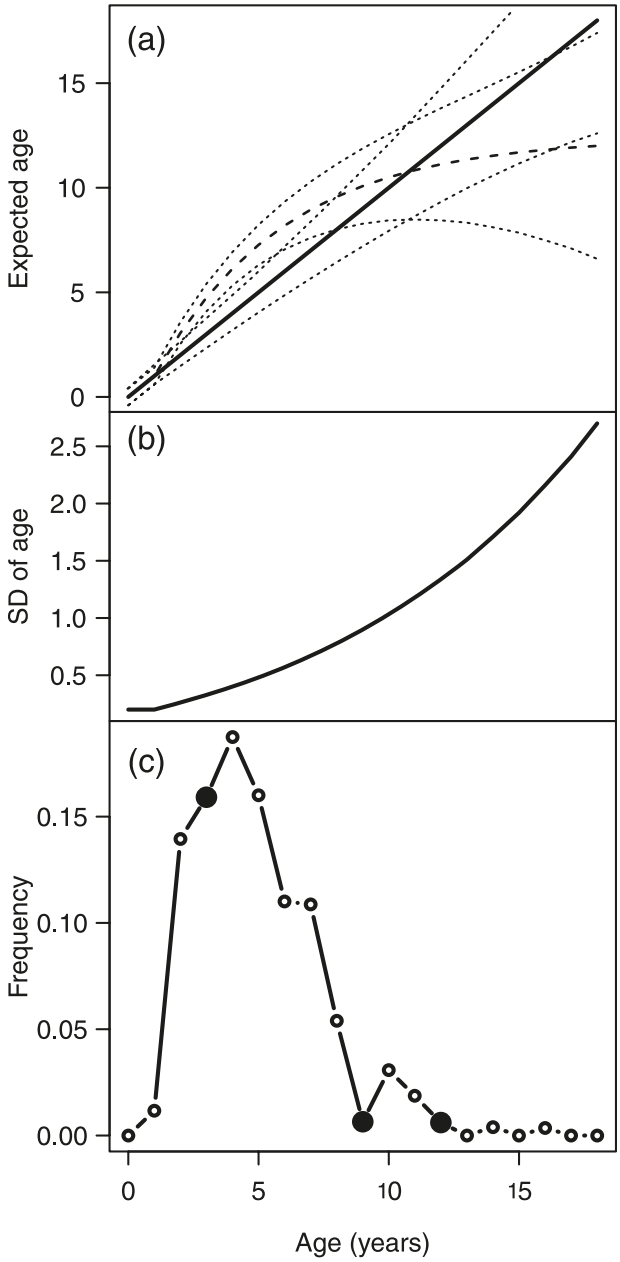


Table 2. The eight variants considered in the simulation study.

Variant	Description
1	Both readers provide unbiased age estimates and both have the same age-reading error standard deviations
2	Both readers provide unbiased age estimates, but they differ in terms of the standard deviations of their age estimates
3	Reader 1 provides unbiased age estimates, but reader 2 is biased; both readers have the same age-reading error standard deviations
4	Reader 1 provides unbiased age estimates, but reader 2 is biased; the two age readers differ in terms of the standard deviations of their age estimates
5	As for variant 3, except that eq. 3 pertains to the coefficient of variation rather than the standard deviation
6	As for variant 4, except that eq. 3 pertains to the coefficient of variation rather than the standard deviation
7	Reader 2 provides unbiased age estimates, but reader 1 is biased; both readers have the same age-reading error standard deviations
8	Reader 2 provides unbiased age estimates, but reader 1 is biased; the two age readers differ in terms of the standard deviations of their age estimates

Fig. 2. Distributions of percent relative error for the age-reading error standard deviations for the two readers and the expected age by the biased reader. Results are shown for four estimator variants and for the variant selected using BIC. The results in this figure are based on a sample size of 500 double-read ageing structures. The quantities on the *x* axis are the age-reading standard deviations for ages 3, 9, and 12 for reader 1, the age-reading standard deviations for ages 3, 9, and 12 for reader 2, and the expected ages for ages 3, 9, and 12 for reader 2. The "true" variant for these simulations is "with bias / same SDs" (variant 3).

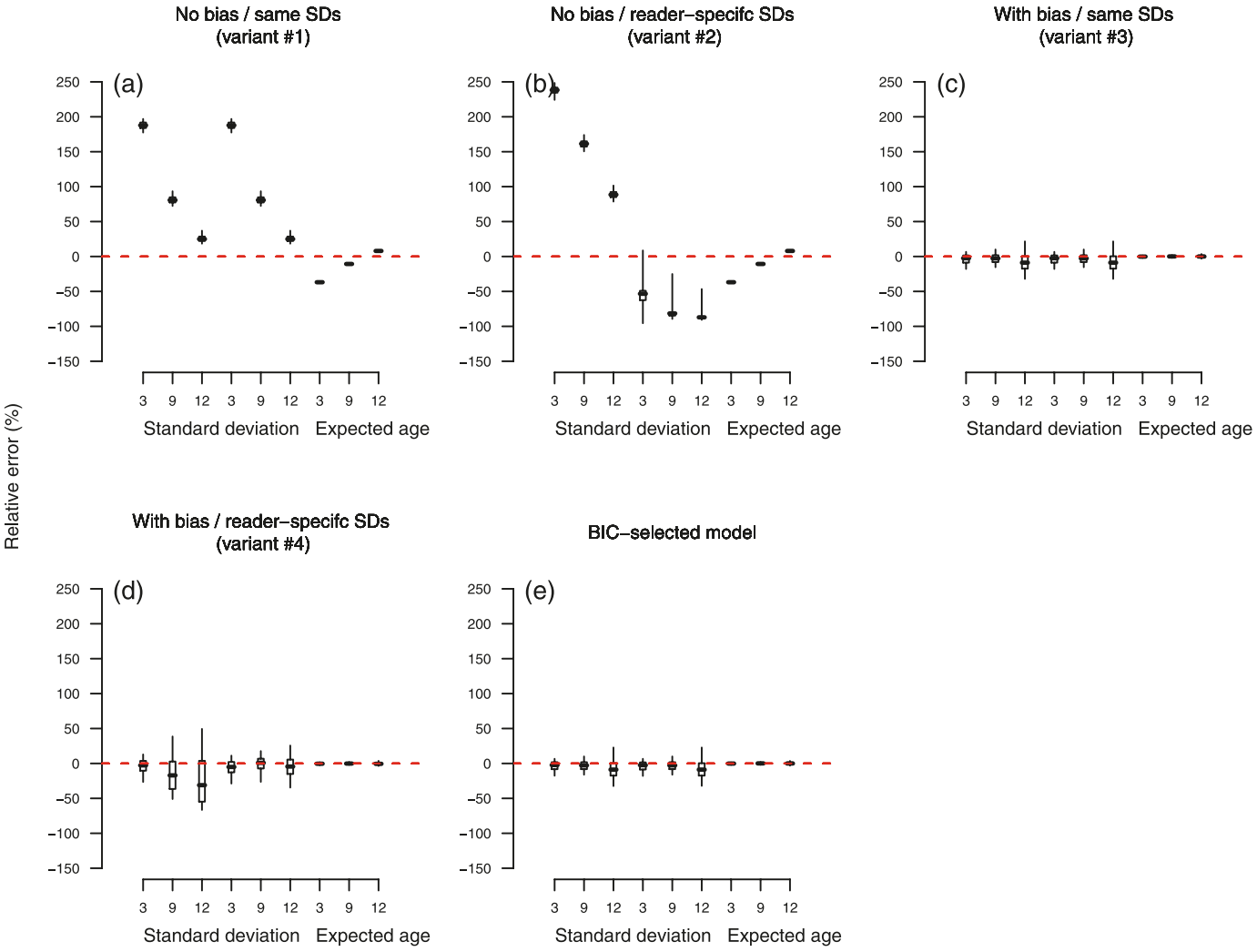
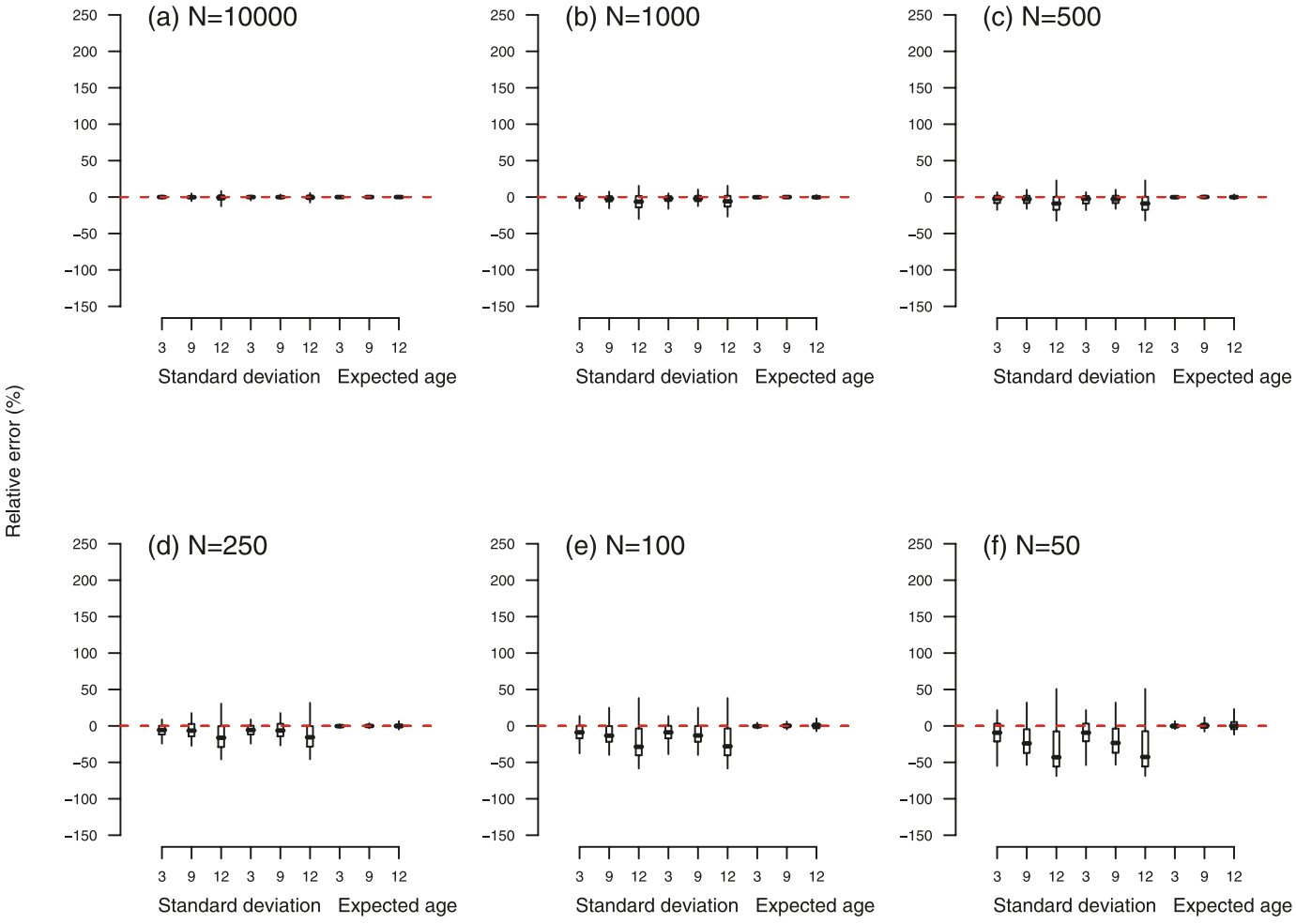


Fig. 3. Distributions of percent relative error for the age-reading error standard deviations for the two readers and the expected age by the biased reader. Results are shown for various sample sizes for an estimator that fits estimator variants 1–4 and selects a "best" variant using BIC.



The sensitivity of the results to the number of double-aged animals is examined by varying the simulated sample sizes from 50 to 10 000. Analyses were conducted in which both age readers are unbiased and in which the number of age classes is greater or smaller than those for silver warehou. The results of these simulations are not shown, however, because they were qualitatively the same as the reported results.

Results and discussion

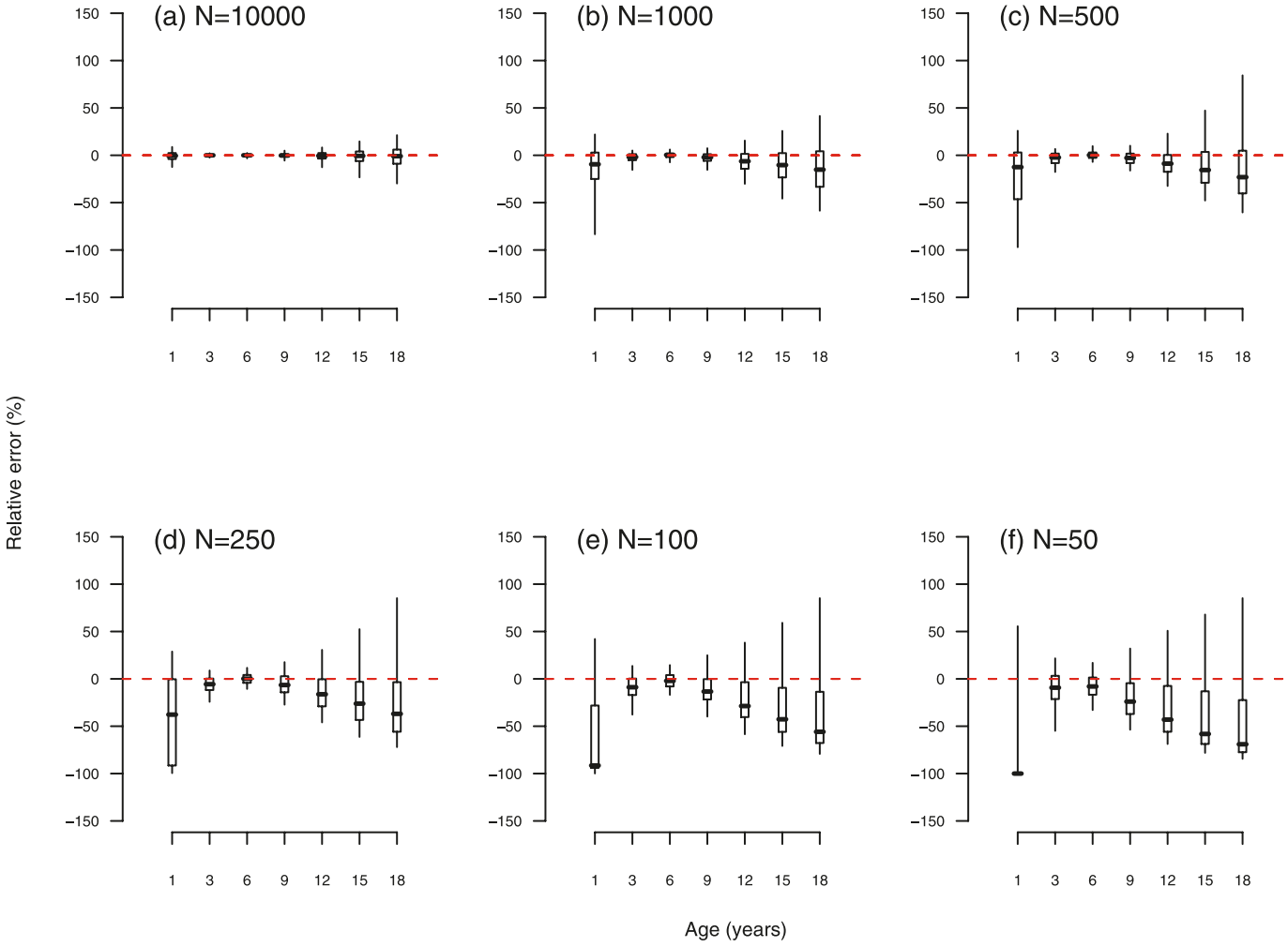
Simulation study

The results of the simulation study are summarized by distributions of percent relative error for the standard deviations of age-reading error and for the expected ages. Ideal performance in the simulation study would involve all of the errors being zero; any systematic deviation from zero would indicate the possibility of some bias, while variability in the errors highlights a lack of precision.

The distributions of percent relative error for the standard deviations of age-reading error for ages 3, 9, and 12 (ages that are spread over the sampled population; see Fig. 1*c*), as well as for the expected ages for these ages for reader 2, suggest that assuming that the two readers are unbiased when they are not leads to estimates of both ageing bias and age-reading error standard deviations with substantial bias (Figs. 2*a* and 2*b*). This bias largely disappears when allowance is made for the possibility that reader 2 is biased. However, the estimates of age-reading error standard deviation tend to be negatively biased, particularly when allowance is made for the possibility that the age readers differ in terms of ageing imprecision (Fig. 2*d*). The results for the BIC-selected model (Fig. 2*e*) are similar to those when the true model (i.e., "with bias / same SDs") is assumed, suggesting that the BIC is capable of correctly identifying the "true" model when that model is in the set of models considered, at least for sample sizes of 500 (the "true" model was selected in over 99% of simulations).

Increasing sample sizes leads to estimates that are both more accurate and more precise (Fig. 3). However, the pattern that the estimates of the age-reading error standard deviations are negatively biased (evident in Fig. 2*e*) is exacerbated for small sample sizes. This occurs because the age-reading error standard deviations are relatively small, and for small sample sizes, almost all of the data will reflect agreement between the two age readers so there are few data

Fig. 4. Distributions of percent relative error for the age-reading error standard deviations for the biased reader (the results for the unbiased age reader are qualitatively identical). Results are shown for various sample sizes for an estimator that fits estimator variants 1–4 and selects a final variant using BIC.



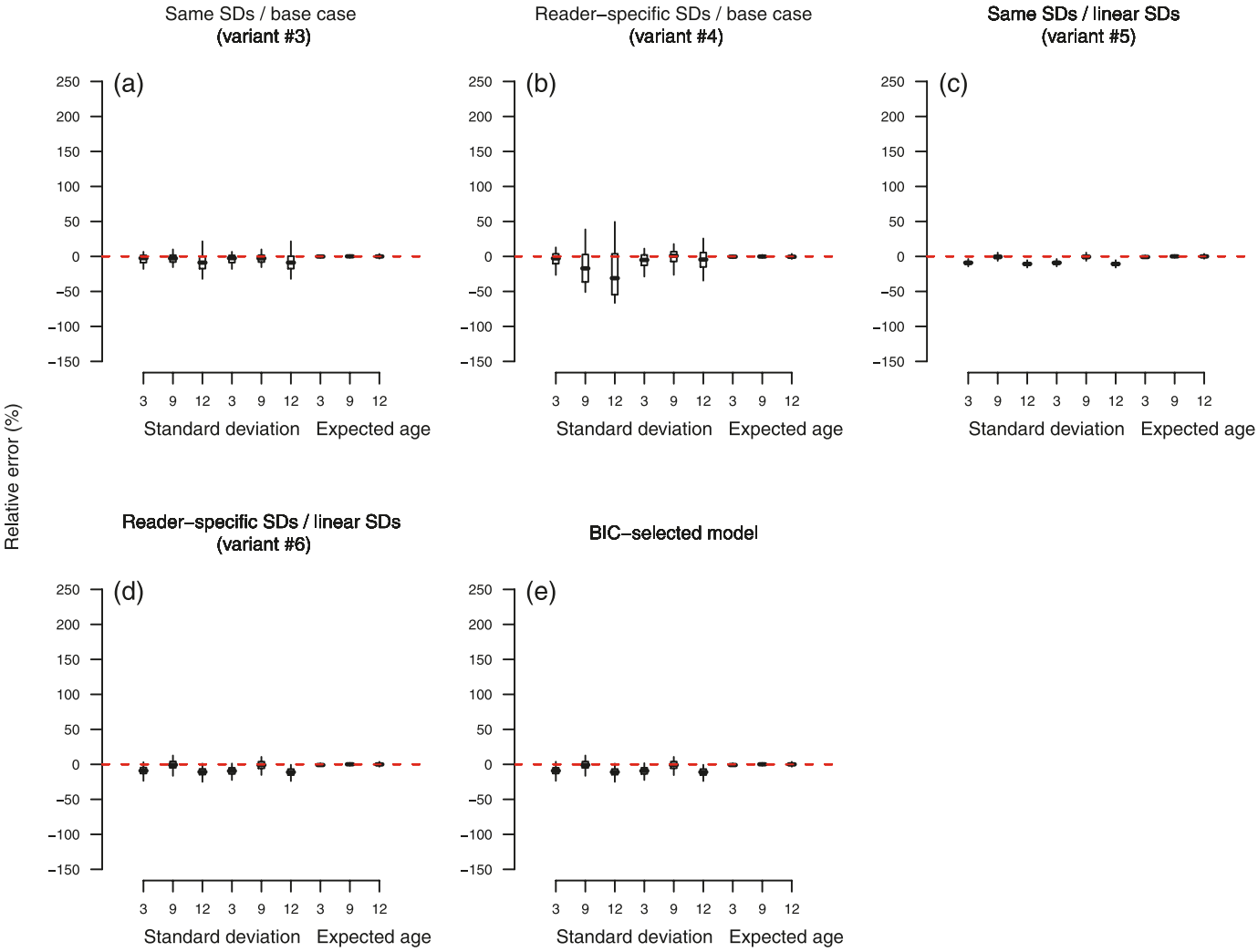
to detect the magnitude of age-reading errors. The biases and poor precision for small sample sizes (evident in Fig. 3) are more substantial for ages 1, 15, and 18, ages that are poorly represented in the sampled population (Fig. 4). This occurs because unless the sample size is very large, only a few animals of these underrepresented age-classes will be included in the sample.

Although the age-reading error standard deviation is a nonlinear function of true age (Fig. 1*b*), it is possible that the data are insufficient to adequately determine the relationship between true age and the age-reading error standard deviation (particularly for small sample sizes). The estimates of age-reading error standard deviation are more precise when the coefficient of variation of age-reading error is assumed to be independent of age (variants 5 and 6) than when the age-reading error standard deviation is assumed to change with age according to eq. 3 (variants 3 and 4) (Figs. 5*c* and 5*d*). However, BIC does not select the model that visually performs best in this case (variant 5), instead generally selecting the model with different, but age-independent, coefficients of variation for the two readers (variant 6). However, the BIC-selected model outperforms the "true" model (i.e., that in which the age-reading error standard deviation increases with age according to eq. 3; Fig. 5*a*) in terms of median relative errors.

The analyses on which Figs. 4 and 5 are based assume that the biased reader can be correctly identified. Performance is (predictably) poor when reader 1 rather than reader 2 is assumed to be biased (Fig. 6). BIC is able to select the "true" model with high probability in this case, although this may be an artifact of there being some model misspecification when reader 1 is assumed to be biased.

There are two situations in which it would be anticipated that the performance of the estimator would be poor. The estimates of age-reading standard deviations are negatively biased if the age estimates are correlated (e.g., Fig. 7*b*). Furthermore, although it is possible in theory to attempt to estimate ageing bias for all readers, the estimates become very imprecise when only one of the readers is biased (contrast Figs. 7*a* and 7*c*) and imprecise and biased when both readers are biased (Figs. 7*d* and 7*e*). This poor performance is not particular to this estimator but would occur for any estimator, as the age-reading errors would be confounded. The only way to overcome the biases (Figs. 7*b*, 7*d*, and 7*e*)

Fig. 5. Distributions of percent relative error for the age-reading error standard deviations for the two readers and the expected age by the biased reader. Results are shown for estimator variants that assume that age-reading error is governed by eq. 3 ("base case"), that assume that the coefficient of variation of age-reading error is independent of age ("linear SDs"), and for an estimator that selects among these four variants using BIC. The results in this figure are based on a sample size of 500 double-read ageing structures and the "true" model is variant 3.



would be to obtain age estimates for animals for which the true age was known.

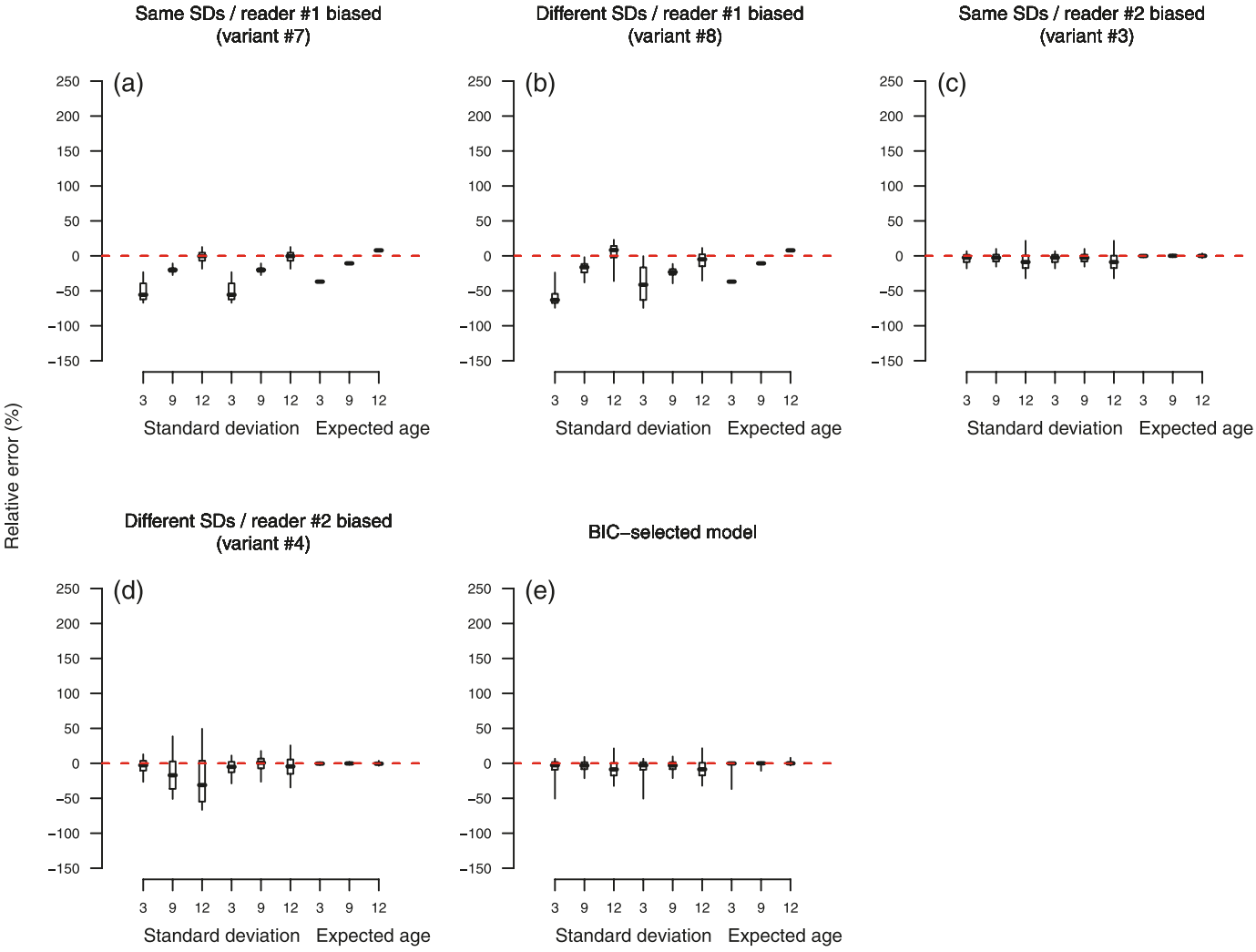
Application to actual data

Example application: silver warehou

Silver warehou is a species (maximum age 23 years) that is distributed throughout southeastern Australia, including New South Wales, Tasmania, and Victoria. Adults of this species are found on the continental shelf and upper slope where they are one of the target species of a bottom trawl fishery (Smith 1994). Assessments of this species have been conducted using statistical catch-at-age models fitted to age and size composition information, as well as to standardized catch rates (e.g., Thomson and He 2001). The data set on which age-reading error matrices for silver warehou can be based consists of age estimates made using sectioned and whole otoliths: 1042 otoliths were read sectioned twice, 1104 were read sectioned once and whole once, and 358 were read sectioned twice and whole once. Silver warehou is therefore a fairly complicated case in which some otoliths were read three times and there are multiple methods for age determination, one of which (reading of otoliths whole) was expected to be biased.

The fits of 10 models to this data set are summarized in Table 3. Different models are considered for the first and second sectioned reads of an otolith and for the whole read of the otolith. These models examine whether the expected ages are the same for the two sectioned reads (as might be expected a priori), whether eq. 3 should relate to the standard deviation or the coefficient of variation of age-reading error, and whether it is reasonable to assume that the coefficient of variation of age-reading error is independent of age (Table 3). All of the models shown assume that age readings based on whole otoliths are biased relative to those based on sectioned otoliths and that the level of age-reading error differs between whole and sectioned otoliths (this is clearly apparent from a visual examination of the raw data).

Fig. 6. Distributions of percent relative error for the age-reading error standard deviations for the two readers and the expected age by the biased reader. Results are shown for two estimator variants in which the biased reader (reader 2) is assumed to be biased (variants 3 and 4), for two variants in which the unbiased reader (reader 1) is assumed to be biased (variants 7 and 8), and for the variant selected using BIC. The results in this figure are based on a sample size of 500 double-read ageing structures and the "true" model is variant 3.



BIC selects model 9 (the two readings based on sectioned otoliths lead to the same expected ages but they differ in terms of precision). Models that assume that the two sectioned reads lead to (slightly) different expected ages and that the standard deviations of age-reading error are the same for the two sectioned reads (models 1 and 2) have BIC values most similar to that for the selected model. The models that assume that the coefficient of variation follows eq. 3 and (particularly) that the coefficient of variation is independent of age (models 5–8) have much higher BIC values than model 9. Comparing the BIC values for models 5 and 6 with those for models 7 and 8 suggests that there is very strong evidence against the assumption that the coefficient of variation for the whole readings is independent of age.

There are marked differences between the standard deviations of age-reading error for the two sectioned reads (Fig. 8*a*), with the model suggesting that the age-reading error standard deviation is much less for the second read for animals of age 6 and older. It should be noted, however, that there are very few animals of age 6 and older in the samples on which the analyses for silver warehou are based (Fig. 8*c*) and that the estimates of the standard deviations of age-reading error are imprecise for animals 10 years and older.

The APE based on sectioned readings is 3.6%, while the average coefficient of variation is 5.1%. These values are markedly lower than the coefficients of variation in Fig. 8*b*. Moreover, the use of APE or a coefficient of variation fails to indicate that the coefficient of variation of age-reading error likely changes as a function of age.

The quality of the fit of the selected model is explored, plotting the observed number of animals for each combination of age from the first sectioned read and age from the second sectioned read with the associated model prediction (Fig. 9). The fits are generally adequate (the observed and model-predicted values lie close to the 1:1 line), although the fits are poorer for combinations for which the number of observations is low. There is no evidence in Fig. 9 for overdispersion associated with the fit of model 9 (in con-

Fig. 7. Distributions of percent relative error for the age-reading error standard deviations for the two readers and the expected age by reader 2. (*a*) Variant 3 applied when variant 3 is the true model, (*b*) as for Fig. 7*a* except that the age-reading errors are correlated (ρ = 0.71), (*c*) as for Fig. 7*a* except that ageing bias is estimated for both readers, (*d*) variant 3 applied when both age readers are biased (reader 1 positively biased, reader 2 negatively biased), and (*e*) as for Fig. 7*d* except that ageing bias is estimated for both readers. All of the results in the figure are based on a sample size of 500 double-read ageing structures.

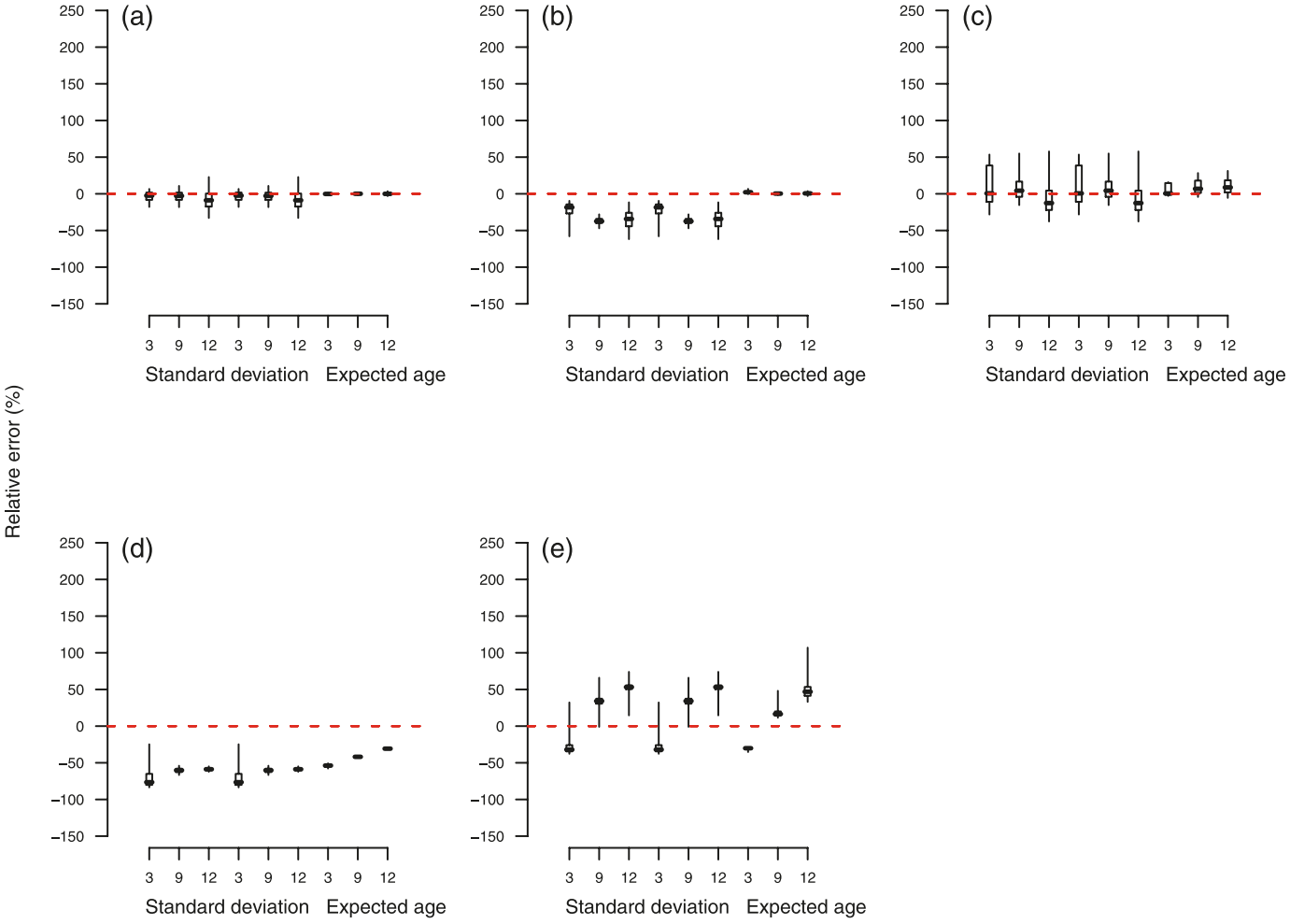


Table 3. Alternative models for age-reading error for silver warehou and the results of applying BIC to select among these models.

	First sectioned read		Second sectioned read		Whole read			
Model	Bias	Precision	Bias	Precision	Bias	Precision	No. of parameters	ΔBIC
1	Unbiased	Eq. 3	Unbiased	As for 1	Eq. 4	Eq. 3	60	7.82
2	Unbiased	Eq. 3	Eq. 4	Eq. 3	Eq. 4	Eq. 3	66	4.32
3	Unbiased	Eq. 3[a]	Unbiased	As for 1	Eq. 4	Eq. 3[a]	60	39.56
4	Unbiased	Eq. 3[a]	Eq. 4	Eq. 3[a]	Eq. 4	Eq. 3[a]	66	70.46
5	Unbiased	Con CV	Unbiased	As for 1	Eq. 4	Con CV	56	114.16
6	Unbiased	Con CV	Eq. 4	Con CV	Eq. 4	Con CV	60	106.86
7	Unbiased	Con CV	Unbiased	As for 1	Eq. 4	Eq. 3	58	14.49
8	Unbiased	Con CV	Eq. 4	Con CV	Eq. 4	Eq. 3	62	10.07
9	Unbiased	Eq. 3	Unbiased	Eq. 3	Eq. 4	Eq. 3	63	0
10	Unbiased	Eq. 3	Eq. 4	As for 1	Eq. 4	Eq. 3	63	14.36

Note: Con CV denotes that the coefficient of variation of age-reading error is assumed to be independent of age, and "as for 1" denotes that the precision of the age-reading error matrix is assumed to be the same for the two sectioned reads.

[a]Eq. 3 pertains to the coefficient of variation rather than the standard deviation.

Fig. 8. (*a*) Standard error and (*b*) coefficient of variation of age-reading error for the three types of age readings for silver warehou (broken line, whole reads; solid and dotted lines, sectioned reads; dashed–dotted lines, asymptotic 95% confidence intervals about the point estimates). (*c*) Estimated age structures underlying the three samples for silver warehou based on the selected model.

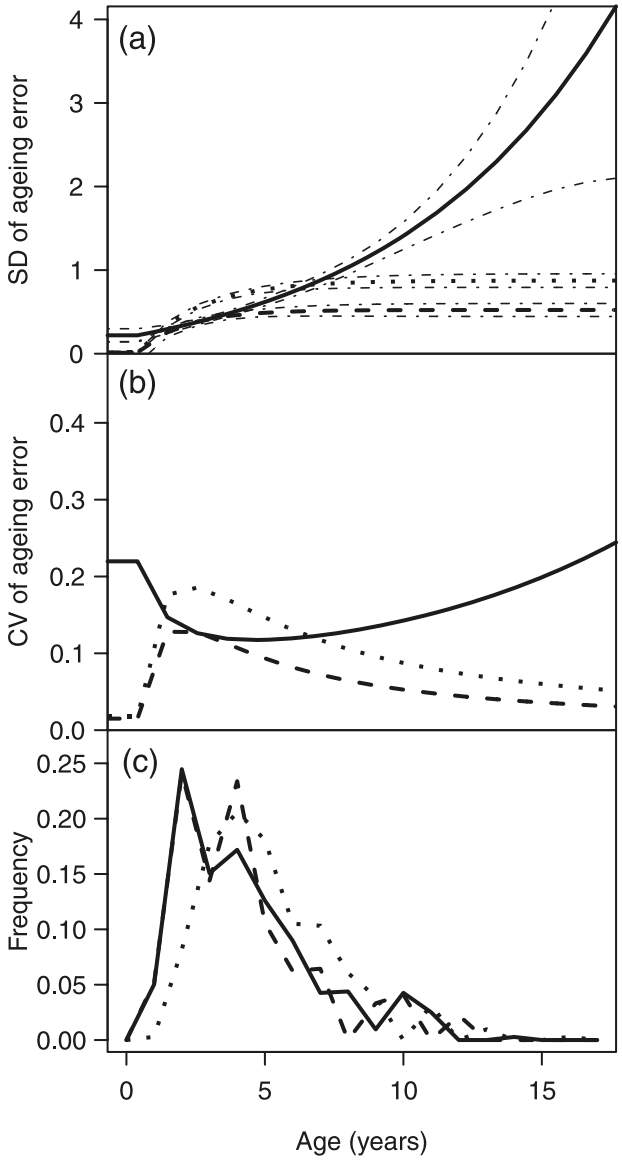
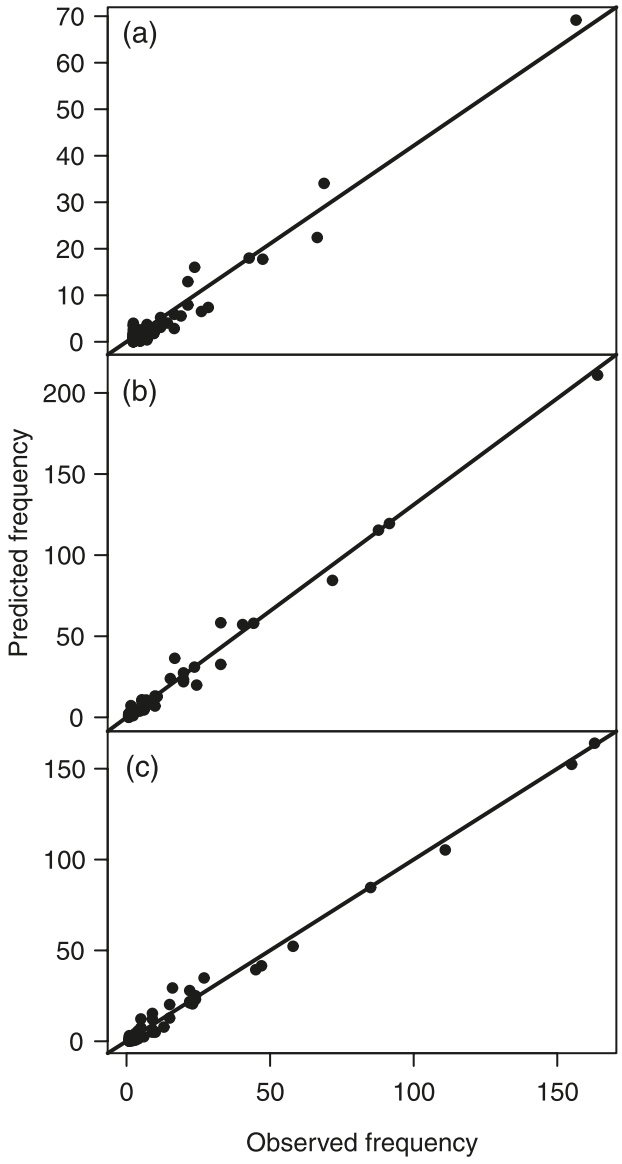


Fig. 9. Diagnostic plots for the fit of the age-reading error model to the data for silver warehou. Samples sizes for the data sets: (*a*) 358, (*b*) 1104, and (*c*) 1042.



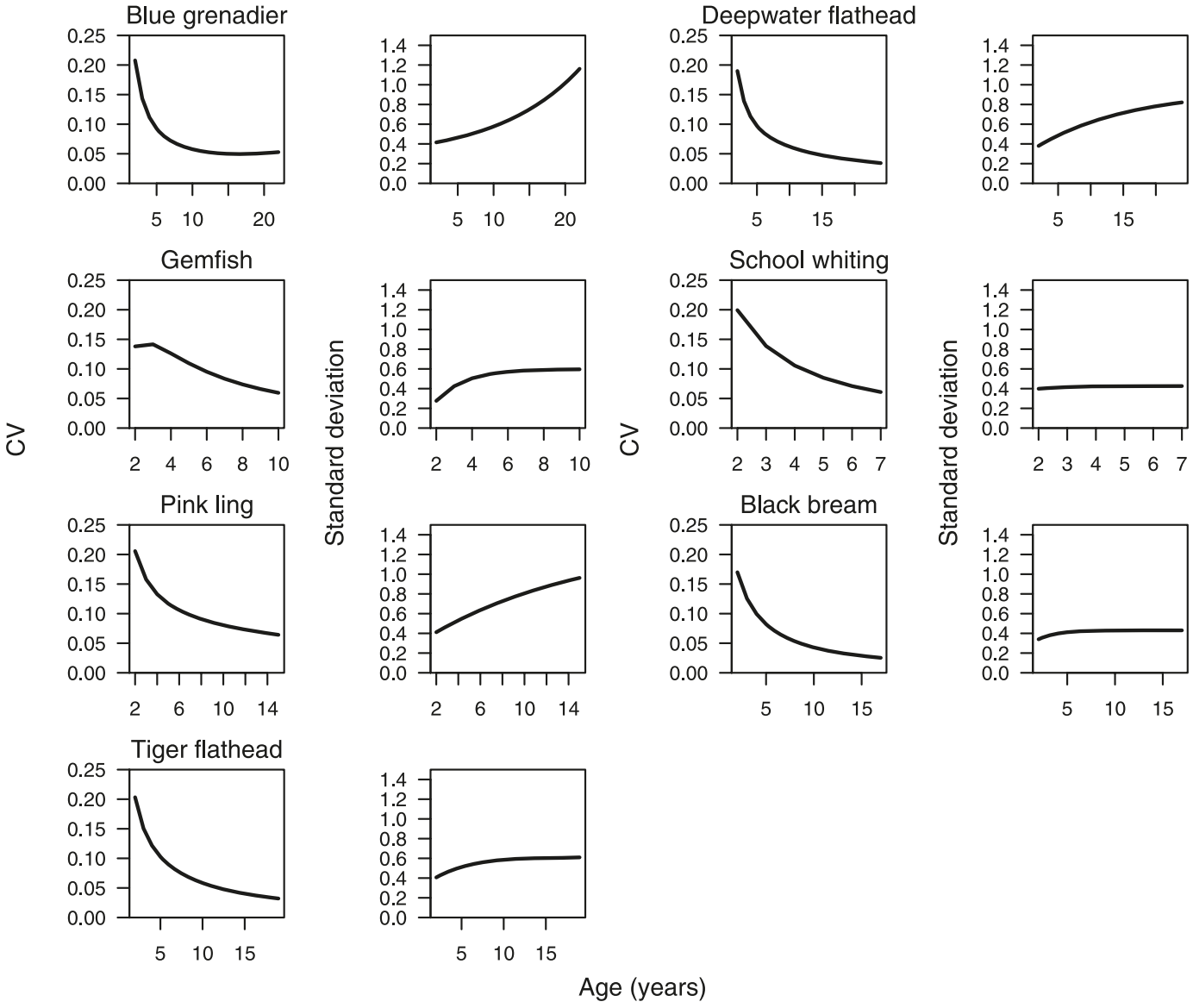
trast, there is strong evidence for overdispersion for some of the models that led to higher BIC values than the selected model; results not shown).

Application to other species

The approach outlined above has been applied to a variety of the fish species harvested off southern Australia, and the results have been included in stock assessments based on the integrated analysis paradigm. Some of these applications are summarized (Fig. 10) in terms of the coefficients of variation and standard deviations of age-reading error as a function of true age (for the primary age reader) for seven species (the results for silver warehou are shown in Fig. 8) for a model in which all readers are assumed to be unbiased and in which the curvature parameter α (eq. 3) is estimated. This model was supported by BIC for most of the species. No plots of age versus bias are provided, as the primary reader is assumed to be unbiased.

The coefficients of variation of age-reading error decline with (true) age. The coefficients of variation of age-reading error for animals of age 2 are generally 0.12–0.20, whereas those for the oldest ages depend on species. Tiger and deepwater flathead and black bream have the lowest coefficients of variation for the oldest ages (<0.05). When expressed in terms of standard deviations, the species for which age estimates are the least precise are blue grenadier, silver warehou, and pink ling. The relationship between the age-reading error standard deviation and true age can be convex (silver warehou and blue grenadier), concave (deepwater flathead and pink ling), or asymptotic (gemfish, school whiting, black bream, and tiger flathead), highlighting the need

Fig. 10. Coefficients of variation (CV) and standard deviations of age-reading error versus true age for seven species in Australia's southern and eastern scalefish and shark fishery. The results in this figure pertain to the primary age reader.



to have a method for quantifying age-reading error that is sufficiently flexible to capture a wide range of alternative models of age-reading error.

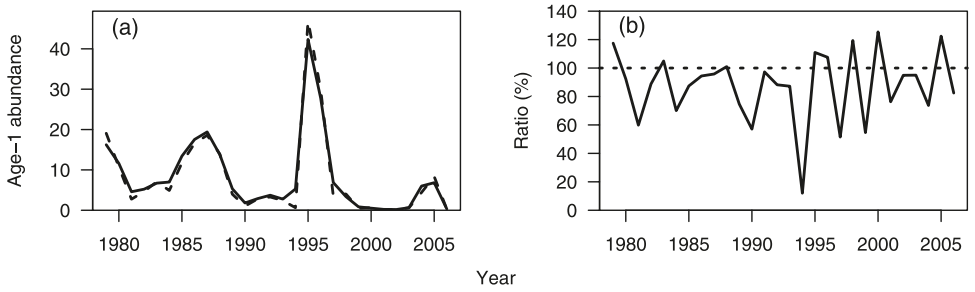
General discussion

The simulation study suggests that the method is able to correctly determine both ageing bias and the extent of age-reading imprecision if sample sizes are large, particularly for those ages that are well represented in the sampled population. An implicit assumption of both the way the pseudo data sets were generated for the simulation experiment and the structure of the estimator is that age-reading errors occur at random. However, there will be correlation in age-reading errors among readers if age-reading errors are caused by factors that are common to all age-readers, i.e., misinterpretation of visual cues on, for example, an otolith or spine or due to institutional ageing protocols can lead to negatively biased estimates of the extent of age-reading error (e.g., Fig. 7*b*). Similarly, the method is unable to estimate ageing bias and age-reading imprecision when both readers are biased without auxiliary information on the true age of the sampled animals (e.g., Figs. 7*d* and 7*e*).

It is not possible to estimate the extent of correlation in age estimates among age readers and whether all of the readers are biased without auxiliary information (such as age estimates for known-age animals). Therefore, having a method that is able to estimate the extent of ageing bias and stochastic age-reading error does not replace the need for age validation using direct techniques (e.g., Francis et al. 1992; Smith et al. 1995; Kalish et al. 1997) or by tracking strong cohorts in catch-at-age matrices (e.g., Punt et al. 2001).

The estimates of recruitment for blue grenadier for analyses in which account is taken of age-reading error and in which age-reading error is assumed to be exact indicate that ignoring age-reading error leads to the sizes of weak cohorts

Fig. 11. (*a*) Estimates of recruitment (age-1 abundance) for blue grenadier off southeast Australia based on an assessment that accounts for age-reading error (broken lines) and an assessment that ignores age-reading error and hence assumes that the age estimates are exact (solid line) and (*b*) ratio of the estimates of recruitment accounting for age-reading error to those in which age-reading error is ignored.



being overestimated and those of strong cohorts being underestimated, with the effect that the extent of variation in recruitment is underestimated (Fig. 11). For blue grenadier at least, the size of this effect is relatively small and the general patterns of year-class strength are easily determined even if age-reading error is ignored.

The inability to account for correlations in the errors among age readers is not the only limitation of the method. The results suggest considerable variation in year-class strength for blue grenadier, and it is known that animals from the strong cohorts have grown slower than those from the weak cohorts (Punt and Smith 2001). An original aim of this work was to assess whether age-reading error was different for strong and weak cohorts of blue grenadier (an expectation being that age readers will have assigned "uncertain" otoliths to the strong cohorts; Kimura et al. 1992). However, the nature of the otolith sampling program, which is designed to provide data for stock assessment purposes, is such that the vast bulk of the samples are for the strong cohorts so that the power to detect differences in age-reading standard deviations between strong and weak cohorts is very low. The otolith sampling strategy would need to be modified considerably to address a question such as this.

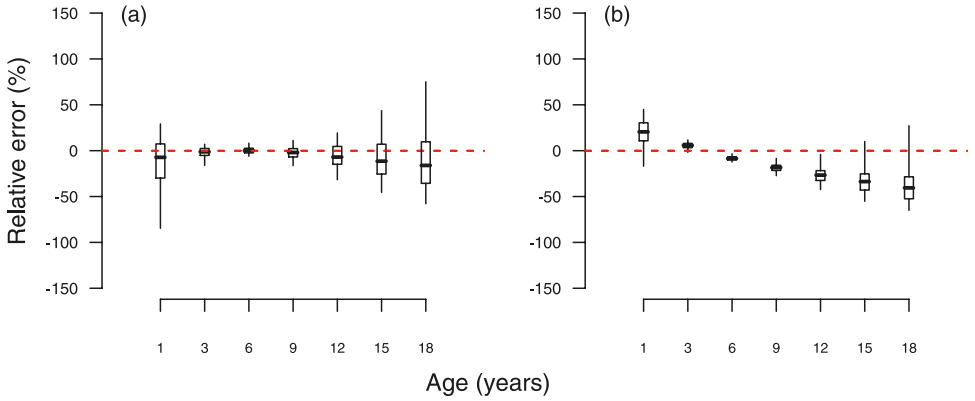
Although the approach outlined above can deal with cases in which there are several readers and the data available to quantify age-reading error consist of numerous data sets in which some ageing structures are read by various subsets of the readers, this can lead to substantial computational requirements owing to the need to estimate many nuisance parameters (essentially the underlying age structure of the sample for each subset of readers). It is therefore beneficial to have a design for how double (and triple) reads are conducted and perhaps to limit the number of readers to ensure that adequate sample sizes are available for each reader. In contrast, having fewer readers may mask the ability to detect ageing bias, especially if the readers are all part of the same laboratory and hence for whom age-reading error may be correlated. It is standard practice within the CAF that only one reader is responsible for ageing a particular species each year once initial ageing protocols have been established. This should be the aim of any ageing facility that provides long-term production ageing data, although it should also be recognized that this is rarely possible due to staff turnover.

The number of double-aged animals needed to obtain relatively precise and unbiased estimates of age-reading error depends on the life history of the species (and the extent of variation in age-reading error). For example, very long-lived species would require large sample sizes. An additional problem with long-lived species is that the number of nuisance parameters depends on the number of age classes postulated to be included in the sample, so application of the estimator to long-lived species leads to fewer degrees of freedom for the same sample size, with a consequential reduction in precision. In fact, it may be appropriate to apply a rule-of-thumb such as there should be minimally a sample of 10 for each age in the sample. Finally, there are occasionally very old animals in the sample. These "loners" will impact the number of estimable parameters of the age-reading error model, and applications of the method have generally excluded the upper 1% of the oldest ages. In principle, the difficulties associated with "loners" and long-lived species could be overcome by implementing the estimation method within a random effects context. Unfortunately, the nuisance parameters (which would constitute the random effects in a random effects model) enter the model in a highly nonlinear way, rendering the problem of maximizing the likelihood function computationally prohibitive at present. It is possible that Bayesian methods (based, for example, on the WinBUGs platform) could be used to overcome the computational concerns associated with a nonlinear random effects implementation of the method.

A variety of techniques, both numerical and graphical (e.g., Beamish and Fournier 1981; Campana 2001; Power et al. 2006), have been developed to examine age-reading error. Some of these (e.g., the "percent agreement" method) are now known to be inappropriate in many contexts (Campana et al. 1995), whereas others (e.g., computing an age-reading error coefficient of variation) are based on a specific model of age-reading error (in this case, that age estimates are random and unbiased with a constant coefficient of variation). In the past, age-reading error matrices based on age-bias tables have formed the basis for assessments (e.g., Smith and Punt 1998; Punt et al. 2001). The estimator outlined above implicitly includes several of the previous methods for quantifying age-reading error as special cases. The basic approach (i.e., eqs. 3 and 4) can be extended to include a wider set of models without negatively impacting the basic analytical structure. This flexibility, combined with the use of model selection methods such as BIC, permits the selection of an appropriate model for age-reading error for each specific case.

Clark (2004) outlined a nonparametric approach for quan-

Fig. 12. Distributions of percent relative error for age-reading error standard deviations based on (*a*) the method in this paper and (*b*) that of Heifetz et al. (1998) in which the "true" age for an otolith is set equal to the average of its two age estimates. The sample size for this figure is 500.



tifying ageing precision. This approach, being nonparametric, is very flexible. However, to provide precise estimates, it requires very large sample sizes, much larger than those that are typically available in actual applications. In addition, the method only provides unique estimates if the distribution for age-reading error is assumed to be symmetric. The original method of Richards et al. (1992) did not treat the true ages as random variables but rather set them to the modal value across age estimates. Similarly, Heifetz et al. (1998) set the true age equal to the average age estimate. The method of this paper is compared with that applied by Heifetz et al. (1998) for a case in which there is one age reader and it is known that the age reader is unbiased (i.e., variant 1; Fig. 12). There is a tendency for both methods to provide negatively biased estimates of age-reading standard error, but the approach of this paper is much closer to being unbiased than that of Heifetz et al. (1998).

Although the method of Richards et al. (1992), as modified in this paper, provides a better basis for estimating the entries of age-reading error matrices, simpler methods such as APE, coefficients of variation, and age-bias plots continue to play an important role in ageing programs. Specifically, these methods provide a rapid (and easily understood) way to ensure quality control in ageing laboratories, such as Australia's CAF, and there is no reason that they should not continue to be used for this purpose. However, the number of rereads needed to apply these methods, given their largely qualitative use, can be much less than is needed to provide reliable estimates of the entries of age-reading error matrices. Adoption of methods such as that outlined in this paper to provide the types of input needed for stock assessment purposes implies (generally) an increase in the number of rereads conducted routinely, although the number of rereads could be minimized by selecting otoliths for rereading to ensure a relatively uniform coverage of the ages for which measures of age-reading error are most needed. Fairly regular calculation of age-reading error matrices using the method in this paper will also provide a stronger (and more quantitative) basis to evaluate whether the ageing procedures remain consistent over time. Finally, consideration needs to be given to the trade-off between conducting double reads to enable age-reading error matrices to be estimated and conducting primary reads to ensure that the data necessary for stock assessment purposes (e.g., the construction of catch age compositions and age–length keys) are of sufficient precision. In principle, this trade-off could be evaluated by means of simulations that capture the entire management system, including data collection schemes, data analysis (such as construction of age–length keys and ageing error matrices), and application of stock assessment methods and harvest control rules.

Acknowledgements

This work was jointly supported by FRDC project 2002/095, the Department of Primary Industries, Victoria, and CSIRO Marine and Atmospheric Research. Support for A.E.P. was also provided by National Marine Fisheries Service (NMFS) grant NA07FE0473. Rich Little and Geoff Tuck (CMAR), Ian Stewart (NMFS, NOAA), Paul Breen (NIWA), and two anonymous reviewers are thanked for their comments on an earlier version of this manuscript. Geoff Tuck (CMAR) is also thanked for providing the data on which Fig. 11 is based.

References

Beamish, R.J., and Fournier, D.A. 1981. A method for comparing the precision of a set of age determinations. Can. J. Fish. Aquat. Sci. **38**: 982–983. doi:10.1139/f81-132.

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., and Smith, M.H. 2003. CASAL (C++ algorithmic stock assessment laboratory): CASAL user manual v2.01-2003/8/01. NIWA Tech. Rep. No. 124. NIWA, Private Bag 14901, Wellington, N.Z.

Burnham, K.P., and Anderson, D.R. 2002. Model selection and inference: a practical information-theoretic approach. 2nd ed. Springer-Verlag, New York.

Campana, S.E. 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. J. Fish Biol. **59**: 197–242. doi:10.1111/j.1095-8649.2001.tb00127.x.

Campana, S.E., Annand, M.C., and McMillan, J.I. 1995. Graphical and statistical methods for determing the consistency of age determinations. Trans. Am. Fish. Soc. **124**: 131–138. doi:10.1577/1548-8659(1995)124<0131:GASMFD>2.3.CO;2.

Clark, W.G. 2004. Nonparametric estimates of age misclassification from paired readings. Can. J. Fish. Aquat. Sci. **61**: 1881–1889. doi:10.1139/f04-122.

Fournier, D., and Archibald, C.P. 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. **39**: 1195–1207. doi:10.1139/f82-157.

Francis, R.I.C.C., Paul, L.J., and Mulligan, K.P. 1992. Aging of adult snapper (*Pagrus auratus*) from otolith ring counts: validation by tagging and oxytetracycline injection. Aust. J. Mar. Freshw. Res. **43**: 1069–1089. doi:10.1071/MF9921069.

Heifetz, J., Anderi, D., Maloney, N.E., and Rutecki, T.L. 1998. Age validation and analysis of ageing error from marked and recaptured sablefish, *Anoplopama fimbria*. Fish. Bull. (Washington, D.C.), **97**: 256–263.

Hilborn, R., Maunder, M., Parma, A., Ernst, B., Payne, J., and Starr, P. 2003. Coleraine: a generalized age-structured stock assessment manual. User's manual version 2.0. SAFS-UW-0116, University of Washington, Seattle, Wash.

Kalish, J.M., Johnstone, J.M., Smith, D.C., Morison, A.K., and Robertson, S.G. 1997. Application of the bomb radiocarbon chronometer to the validation of blue grenadier *Macruronus novaezelandiae*. Mar. Biol. (Berl.), **128**: 557–563. doi:10.1007/s002270050121.

Kimura, D.K., Lyons, J.J., MacLellan, S.E., and Geotz, B.J. 1992. Effects of year-class strength on age determination. Aust. J. Mar. Freshw. Res. **43**: 1221–1228. doi:10.1071/MF9921221.

Methot, R.D. 2000. Technical description of the Stock Synthesis Assessment Program. NOAA Tech. Memo. NMFS-NWFSC-43, Seattle, Wash.

Methot, R. 2007. User manual for the integrated analysis program Stock Synthesis 2 (SS2). Model version 2.00c (updated 26 March 2007). Northwest Fisheries Science Center, NOAA, National Marine Fisheries Service, 2725 Montlake Blvd. E., Seattle, WA 98112, USA.

Morison, A.K., Robertson, S.G., and Smith, D.C. 1998. An integrated system for production fish ageing: image analysis and quality assurance. N. Am. J. Fish. Manag. **18**: 587–598. doi:10.1577/1548-8675(1998)018<0587:AISFPF>2.0.CO;2.

Power, G.R., King, P.A., Kelly, C.J., McGrath, D., Mullins, E., and Gullaksen, O. 2006. Precision and bias in the age determination of blue whiting, *Micromesistius poutassou* (Risso, 1810), within and between age-readers. Fish. Res. **80**: 312–321. doi:10.1016/j.fishres.2006.03.031.

Punt, A.E., and Smith, D.C. 2001. Assessments of species in the Australian south east fishery can be sensitive to the method used to convert from size- to age-composition data. Mar. Freshw. Res. **52**: 683–690. doi:10.1071/MF99129.

Punt, A.E., Smith, D.C., Thomson, R.B., Haddon, M., He, X., and Lyle, J.M. 2001. Stock assessment of the blue grenadier *Macruronus novaezelandiae* resource off south-eastern Australia. Mar. Freshw. Res. **52**: 701–717. doi:10.1071/MF99136.

Reeves, S.A. 2003. A simulation study of the implications of age-reading errors for stock assessment and management advice. ICES J. Mar. Sci. **60**: 314–328. doi:10.1016/S1054-3139(03)00011-0.

Richards, L.J., Schnute, J.T., Kronlund, A.R., and Beamish, R.J. 1992. Statistical models for the analysis of ageing error. Can. J. Fish. Aquat. Sci. **49**: 1801–1815. doi:10.1139/f92-200.

Schwarz, G. 1978. Estimating the dimension of a model. Ann. Statist. **6**: 461–464. doi:10.1214/aos/1176344136.

Shepherd, J.G. 1999. Extended survivors analysis: an improved method for the analysis of catch-at-age data and abundance indices. ICES J. Mar. Sci. **56**: 584–591. doi:10.1006/jmsc.1999.0498.

Smith, A.D.M., and Punt, A.E. 1998. Stock assessment of gemfish (*Rexea solandri*) in eastern Australia using maximum likelihood and Bayesian methods. *In* Fisheries stock assessment models. *Edited by* T.J. Quinn II, F. Funk, J. Heifetz, J.N. Ianelli, J.E. Powers, J.F. Schweigert, P.J. Sullivan, and C.I. Zhang. Alaska Sea Grant College Program, AK-SG-98–01, Juneau, Alaska. pp. 245–286.

Smith, D.C. 1994. Spotted warehou, *Seriolella punctata*. *In* The south east fishery. *Edited by* R.D.J. Tilzey. Bureau of Resource Sciences, Australian Government Publishing Service, Canberra, ACT, Australia. pp. 179–188.

Smith, D.C., Fenton, G.E., Robertson, S.G., and Short, S.A. 1995. Age determination and growth of orange roughy (*Hoplostethus atlanticus*): a comparison of annuli counts with radiometric ageing. Can. J. Fish. Aquat. Sci. **52**: 391–401. doi:10.1139/f95-041.

Thomson, R., and He, X. 2001. Modelling the population dynamics of high priority SEF species. Final report to the Fisheries Research and Development Corporation. FRDC Project 1997/115, CSIRO Marine Research, Hobart, Tasmania, Australia.