Nonparametric estimates of age misclassification from paired readings

William G. Clark

Abstract: With a large enough sample, age misclassification probabilities can be estimated nonparametrically by fitting the observed distribution of differences between paired readings. Application of this procedure to Pacific halibut data reveals that the unsigned deviations of single readings from the mode follow a geometric distribution, leading to a simple specification of misclassification probabilities. The potential for bias in this and other published procedures is discussed.

Résumé : Dans un échantillon assez grand, on peut estimer par des méthodes non paramétriques les probabilités de classification erronée des âges en ajustant les distributions observées des différences entre des lectures appariées. L'application de cette procédure à des données sur le flétan du Pacifique indique que les déviations (sans signe) de lectures uniques tirées du mode suivent une distribution géométrique, ce qui mène à une simple spécification des probabilités de classification erronée. Les probabilités d'erreur dans cette procédure et dans d'autres présentées dans la littérature font l'objet d'une discussion.

[Traduit par la Rédaction]

Introduction

Most stock assessments are done by fitting an age-structured model to a series of age-structured fishery and survey data, the latter obtained by collecting samples of fish from commercial and survey catches and determining their ages by counting the annuli (annual rings) formed on a chosen structure, in most cases the sagittal otolith or earstone. Forsberg (2001) provides a detailed account of how otoliths are processed and read at the International Pacific Halibut Commission (IPHC) to estimate the age composition of catches. Campana (2001) reviews methods used to validate age readings and measure precision.

Age readers strive to follow consistent and objective rules for distinguishing annuli from other marks on an otolith, but there is still some judgement involved and, as a result, some variability in readings of a single otolith, both within and among readers (Kimura and Lyons 1991). A single age reading can therefore be regarded as a random variable drawn from a probability distribution. The mode of this distribution occurs at the age that would be chosen most often if the same otolith were read many times by experienced readers following the established protocol, and it is therefore by definition the correct age. It may or may not be the true age of the fish — that depends on whether the protocol is in fact accurate — but it is the correct result of applying the protocol. For clarity, in this paper, this modal age will be called the "canonical age" to stress the point that it is correct relative to the rules in use, not relative to the true age of a fish. When a different age is assigned, the otolith is said to be misclassified.

Otoliths are generally read only once, with a fraction read twice or more for quality control. Estimated age compositions therefore consist of a mixture of correct ages and misclassifications, and that leads to errors in estimates of abundance at age when a stock assessment is done, even if the age-reading protocol is accurate (Reeves 2003). If the misclassification probabilities are known, however, the assessment model can be so constructed that it will predict their effect on the observed age compositions and thereby correctly estimate the true underlying abundance at age.

For various reasons, it is not practical to assemble a large number of independent readings for each of even a small sample of otoliths, so one cannot directly observe the form of the distribution of readings or determine empirically what is the canonical age of a given otolith. What is usually available instead is a small number of readings for each of a large sample of otoliths, often the paired quality control readings done on some proportion of the total. It is common practice to estimate misclassification probabilities by fitting one or another assumed parametric model to data of this sort (e.g., Richards et al. 1992), but as explained below, some of the usual model fits can be misleading. This paper describes a nonparametric estimation method based on the observable distribution of differences between paired readings and its application to data on Pacific halibut (*Hippoglossus stenolepis*).

Received 12 August 2003. Accepted 20 July 2004. Published on the NRC Research Press Web site at http://cjfas.nrc.ca on 15 December 2004.
J17701

W.G. Clark. International Pacific Halibut Commission, P.O. Box 95009, Seattle, WA 98145, USA (e-mail: bill@iphc.washington.edu).

Materials and methods

Until the early 1990s, Pacific halibut otoliths were aged by surface reading wherein whole otoliths are examined at 5–50× under reflected light. It is now known that this method underestimates the age of older fish and so has been replaced by the break-and-burn method (Forsberg 2001). The canonical age of an otolith read by the surface method is therefore biased, but it is still the canonical surface age in that it is the modal age assigned by experienced readers following that protocol. The stock assessment model can be written to account for both bias and misclassification when predicting the age compositions estimated by surface readings.

The data available for estimating the misclassification probabilities consist of a large number (80 963) of paired surface readings, because in many years, all otoliths were read twice. The assigned ages range from 4 to 36, but the first age with a substantial number of readings is 6, and after 20, the numbers are small (Fig. 1).

The approach in this paper will be to derive and test a nonparametric method of estimating the misclassification probabilities from paired readings and then report the nonparametric estimates for the Pacific halibut surface readings. Some comparisons between this procedure and the usual parametric fits will be reported in the discussion section.

Derivation of the estimation method

Let $f_A(a)$ denote the probability distribution of a single reading of an otolith of canonical age A, and let $v = a - A$ denote the deviation of a single reading from the canonical age. The distribution of v is therefore the same as that of a, but with the mode at zero rather than at A. Let $f_A(v)$ denote this shifted distribution.

Suppose we have two independent readings (a_1, a_2) of every otolith in a large sample of otoliths of a given canonical age A. Such a sample could never be assembled in practice because the canonical age of an otolith cannot be observed, but the notion of such a sample will facilitate the derivation. Let d denote the signed difference between the paired readings, or $d = a_1 - a_2 = v_1 - v_2$, and let $g_A(d)$ denote the distribution of d. This distribution is determined by $f_A(v)$, and the elements of $f_A(v)$ can be estimated by predicting and fitting the sample distribution of d. For example, consider the probability of observing a difference $d = -2$ between the first and second readings of a particular otolith. If the deviation v of a single reading from the mode ranges from $v_{\min}$ (a negative number) to $v_{\max}$, then this probability is

$$g_A(-2) = P_A(v_2 = v_1 + 2) = \sum_{v_{\min}}^{v_{\max}-2} f_A(v) f_A(v+2)$$

The probability of observing $d = +2$ is the sum of the products of the same pairs, and likewise for other $d \neq 0$, so $g_A(d)$ is symmetric even if $f_A(v)$ is not. The frequency distribution of the absolute differences therefore contains all of the information in the sample, and the expected frequencies are

$$g_A(0) = \sum_{v_{\min}}^{v_{\max}} [f_A(v)]^2$$

and

$$g_A(|d|) = 2 \sum_{v_{\min}}^{v_{\max}-d} f_A(v) f_A(v+d) \quad \text{for } |d| > 0$$

In mathematical terms, $g_A(d)$ is the autocorrelation function of $f_A(v)$ (Press et al. 1988). Estimating the latter from the former is therefore a matter of inverting an autocorrelation function. There appears to be no discussion of this problem in the literature, but in extensive trials with randomly generated test distributions $f_A(v)$ and the corresponding computed distributions $g_A(d)$, it was always possible to locate the inverse numerically. The only restriction placed on the test distributions was that they decrease monotonically on either side of the mode.

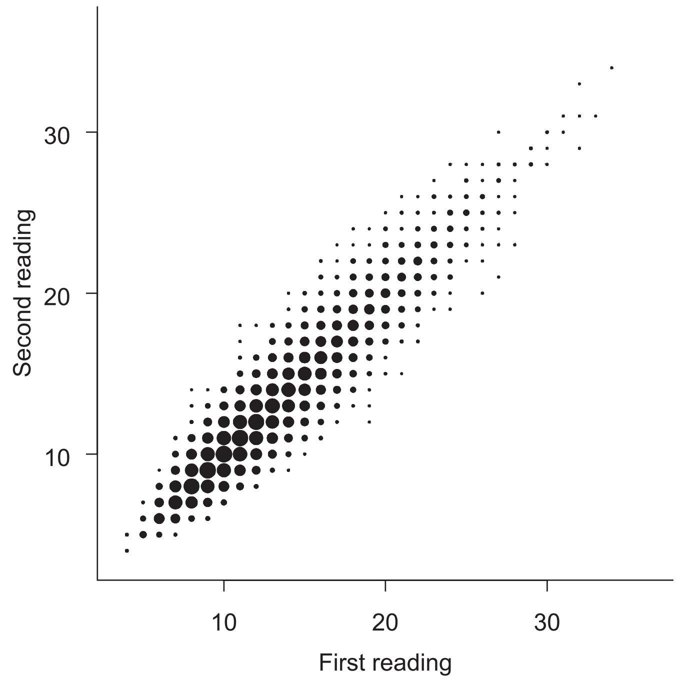


Fig. 1. Joint distribution of the paired readings. The size of each point is proportional to the logarithm of the number of readings.

The trials showed that with deterministic data it is possible to recover any distribution of deviations $f_A(v)$ by numerically fitting the equations above to the observed frequencies of unsigned sample differences $g_A(d)$. However, there are two serious drawbacks to this procedure. The first is that any given $f_A(v)$ and its reverse will generate the same $g_A(d)$. If $f_A(v)$ is asymmetric, the search algorithm will locate one or the other isomer of $f_A(v)$, depending on the starting value of the parameter vector. The second drawback is that the estimates become quite variable in the presence of even moderate sampling error.

As a practical matter, therefore, it is only possible to obtain unique estimates if $f_A(v)$ is assumed to be symmetric. The symmetric estimates are also more stable. If $f_A(v)$ is not symmetric, the estimates obtained in this way are approximate averages, e.g., $\hat{f}_A(-2) \equiv \hat{f}_A(+2)$ is approximately the average of $f_A(-2)$ and $f_A(+2)$. The same fitting procedure can be used, with obvious differences in how the parameter vector is specified. (Detailed instructions for the symmetric case are provided in Appendix A.)

Up to this point, the derivation has referred to a hypothetical sample of paired readings of otoliths of the same canonical age. Real data always consist of an unknown mixture of canonical ages, so they cannot be grouped by canonical age to estimate $f_A(v)$ for each A. Instead they can be grouped by mean assigned age, i.e., the mean of the two actual readings of each otolith. (The mean assigned age of the two readings is a sample value, and it can be higher or lower than the ca-

Fig. 2. Results of 100 simulation trials of the nonparametric estimates, computed with the readings grouped by canonical age and by mean assigned age. The total sample size was 80 000 in each trial.

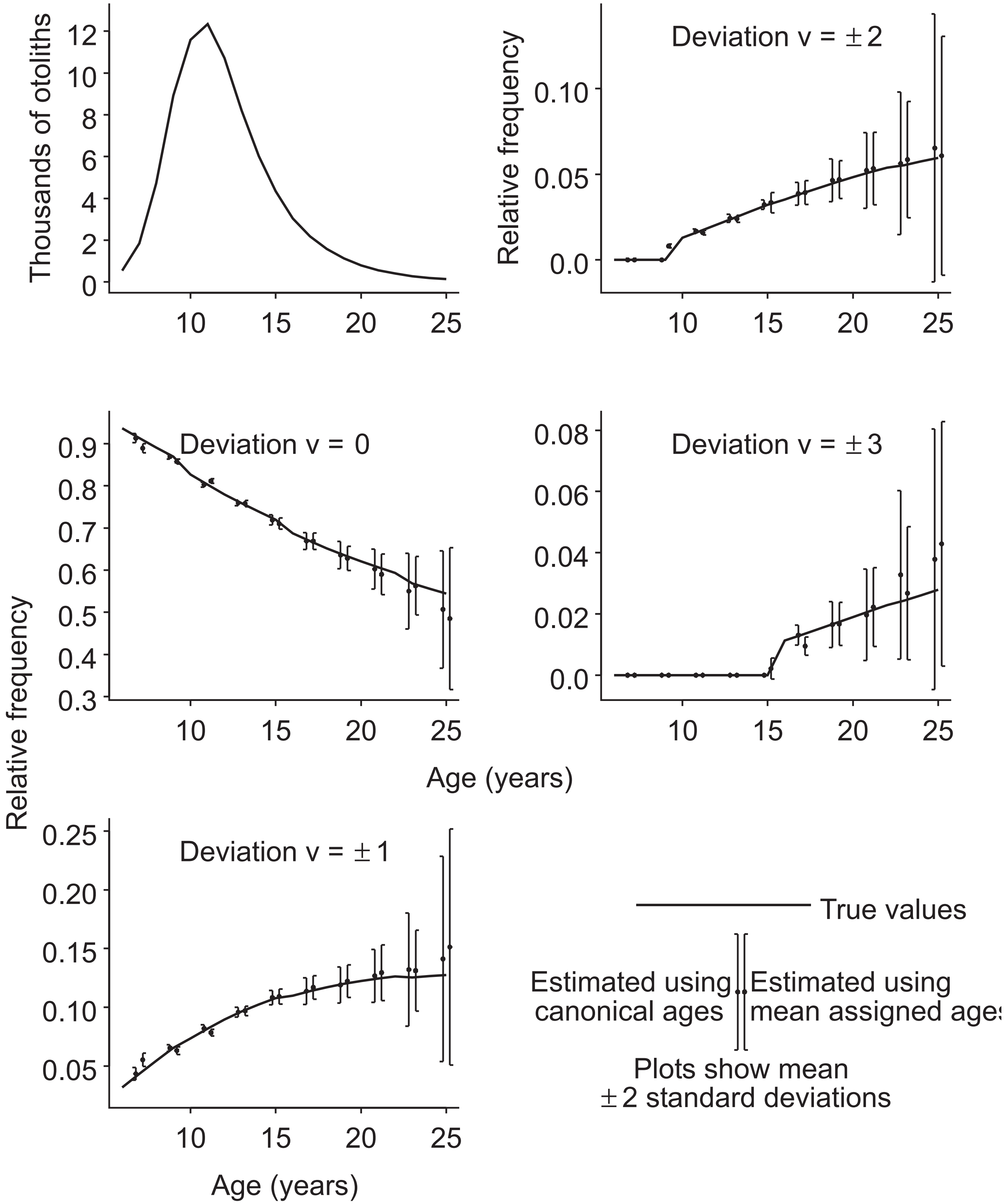


Fig. 3. Observed (points) and predicted (lines) distributions of differences between paired readings of Pacific halibut (*Hippoglossus stenolepis*) otoliths, plotted by mean assigned age.

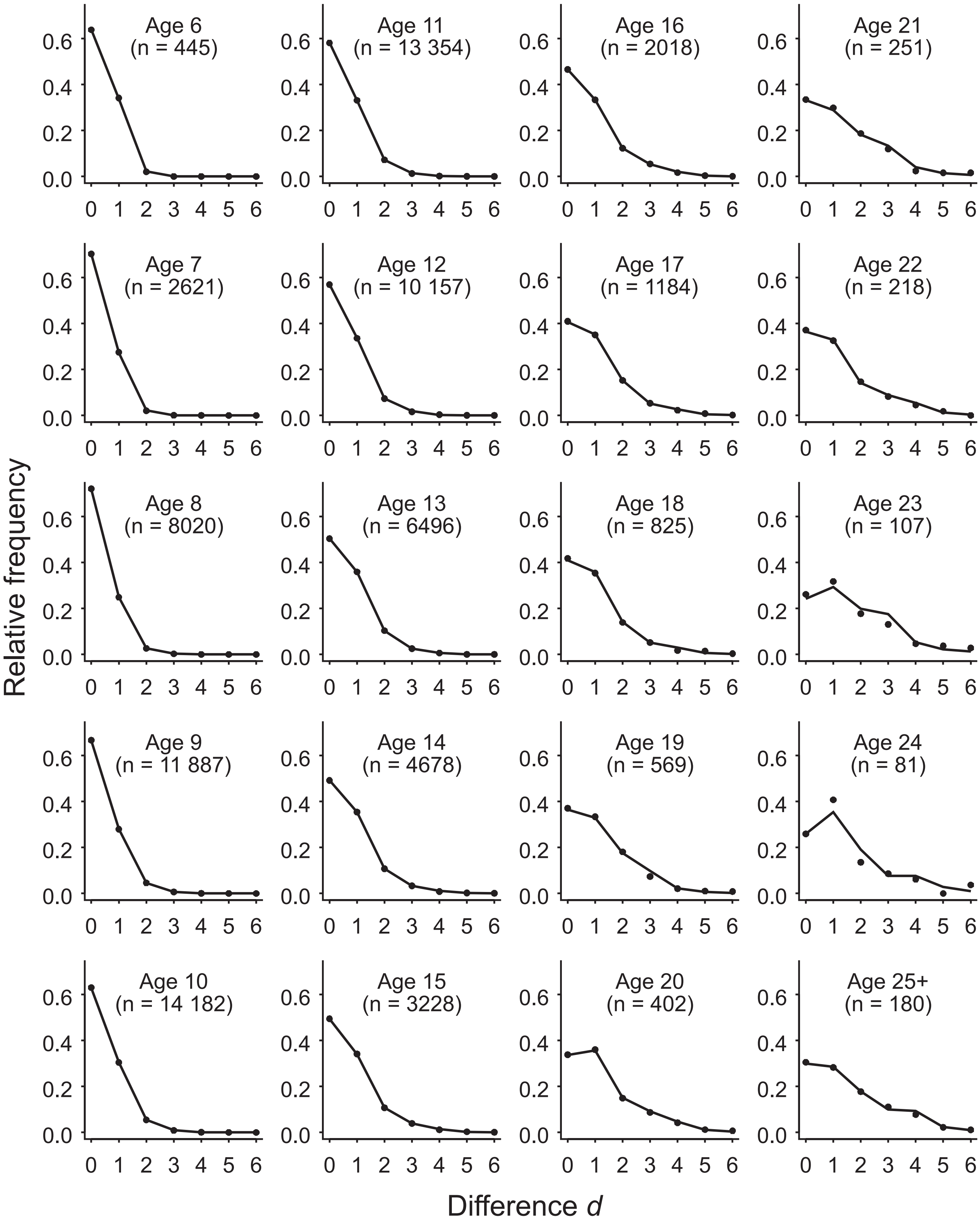
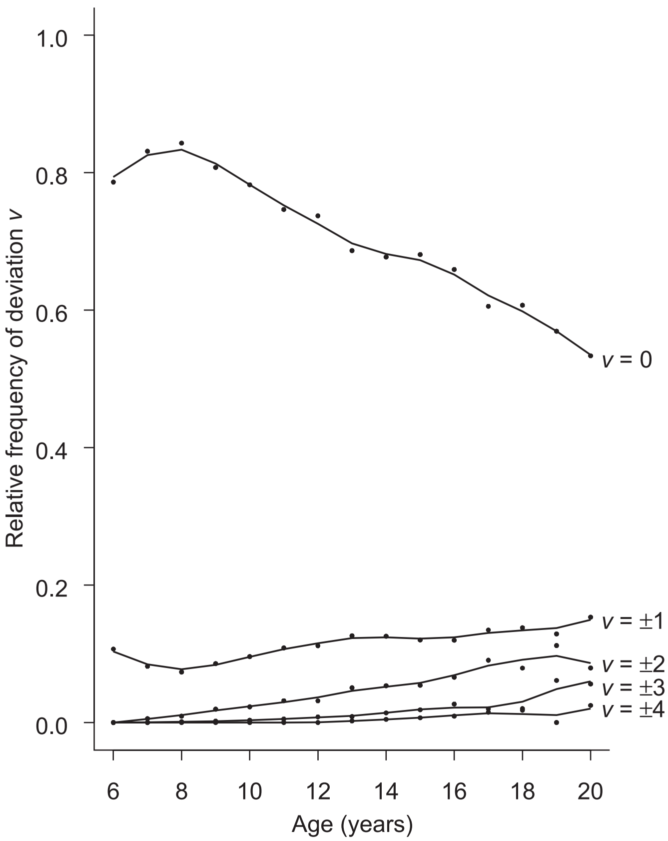


Fig. 4. Estimates of the relative frequencies of deviations of individual readings from the canonical age $f_A(v)$ computed from the Pacific halibut (*Hippoglossus stenolepis*) data, plotted against mean assigned age. The lines are data smoothers.



nonical age of the otolith, which is the mode that would be observed if the otolith were read many times.) In real data, the otoliths of mean assigned age $\bar{a} = A$ will consist of a mixture of canonical ages in the vicinity of A. It seems reasonable to suspect that the distribution of differences $g_{\bar{a}}(d)$ would be very similar to $g_A(d)$, so $\hat{f}_{\bar{a}}(v)$ would be a good estimate of $f_A(v)$, but that needs to be demonstrated. Simulations were conducted for that purpose.

Simulation tests

There were two questions to be investigated by simulation. First, what are the inherent bias and variance of the estimates when computed with paired readings grouped by canonical age (as can be done with simulated but not real data)? Second, what is the effect of grouping the data by mean assigned age?

A simulation model for this purpose was constructed with mortality and selectivity parameters appropriate to Pacific halibut. No attempt was made to model the effect of variable year-class strength on the estimates because the real Pacific halibut data come from several years, so the relative abundance at age in the data is already averaged over strong and weak year classes. Each trial consisted of drawing a sample of size 80 000 from the canonical age composition of the catch and then generating two random readings of each otolith in the sample according to the estimated misclassification probabilities reported in the results section below. Misclassification estimates were computed with the data grouped by both canonical age and mean assigned age. (For the latter grouping, an mean assigned age that fell midway between two successive ages was rounded up or down with equal probability.) Whenever an estimate failed to provide a very good fit to the simulated distribution of sample differences, the search was repeated several times from randomly chosen starting points to check for multiple maxima.

Estimates computed using the canonical ages were unbiased and quite precise for the more numerous age groups, but they became quite variable and slightly biased when sample size dropped into the low hundreds among the oldest age groups in the simulation (Fig. 2). A few instances of multiple maxima also appeared among these groups. Estimates computed using the mean assigned ages were significantly biased statistically ($t > 10$ for many ages), but as can be seen from the graphs, the size of the bias is negligible. The mean assigned ages produce virtually the same estimates as the canonical ages in this case.

Results

Estimates of misclassification probabilities were computed for each mean assigned age from 6 through 25 in the Pacific halibut data. Both symmetric and asymmetric estimates of $f_A(v)$ were computed to check for evidence of asymmetry. For the youngest fish, the best estimates were asymmetric, but from age 10 onward, both procedures produced the same symmetric estimates.

The symmetric estimates predict very well the sample distributions of differences between paired readings through age 20 (Fig. 3). At the smaller sample sizes after age 20, the fits deteriorate, just as in the simulations. The estimated probability that an age reading will be the canonical age decreases from about 0.8 at age 6 to about 0.5 at age 20, whereas the probabilities of deviations of all sizes increase gradually with age (Fig. 4).

The computed age-specific estimates are satisfactory through age 20 but not for older fish. One way to handle age groups with small sample sizes is to smooth the sample frequencies of differences $g_A(0)$, $g_A(1)$, … over ages and then use the smoothed frequencies to estimate $f_A(v)$ at each age.

A simpler solution is available for Pacific halibut because all of the well-determined estimates of $f_A(v)$ have the same simple form. A plot of the estimated log frequencies of unsigned deviations against the deviations is linear at every age (Fig. 5), meaning that the unsigned deviations follow the geometric distribution $f_A(|v|) = p_A\, q_A^{|v|}$, where $q_A = 1 - p_A$. The signed deviations are therefore distributed as $f_A(0) = p_A$ and $f_A(v \neq 0) = (p_A\, q_A^{|v|})/2$.

The value of p_A for any age can be inferred from any well-determined summary statistic of the distribution of readings at that age, e.g., variance. Simulations show that for the Pacific halibut example, the variance of the difference between paired readings is virtually the same whether calculated with the data grouped by canonical age or by mean assigned age. The general trend of the variance of the differences among older age groups (Fig. 6) is clear even though the sample sizes are too small to permit age-specific estimates of $f_A(v)$, and the variance of single readings about the canonical age is just half the variance of the differences. Because of the two

Fig. 5. The logarithms of the nonparametric estimates of the frequencies of unsigned deviations $f(|v|)$ plotted against the deviations. The lines are linear regressions.

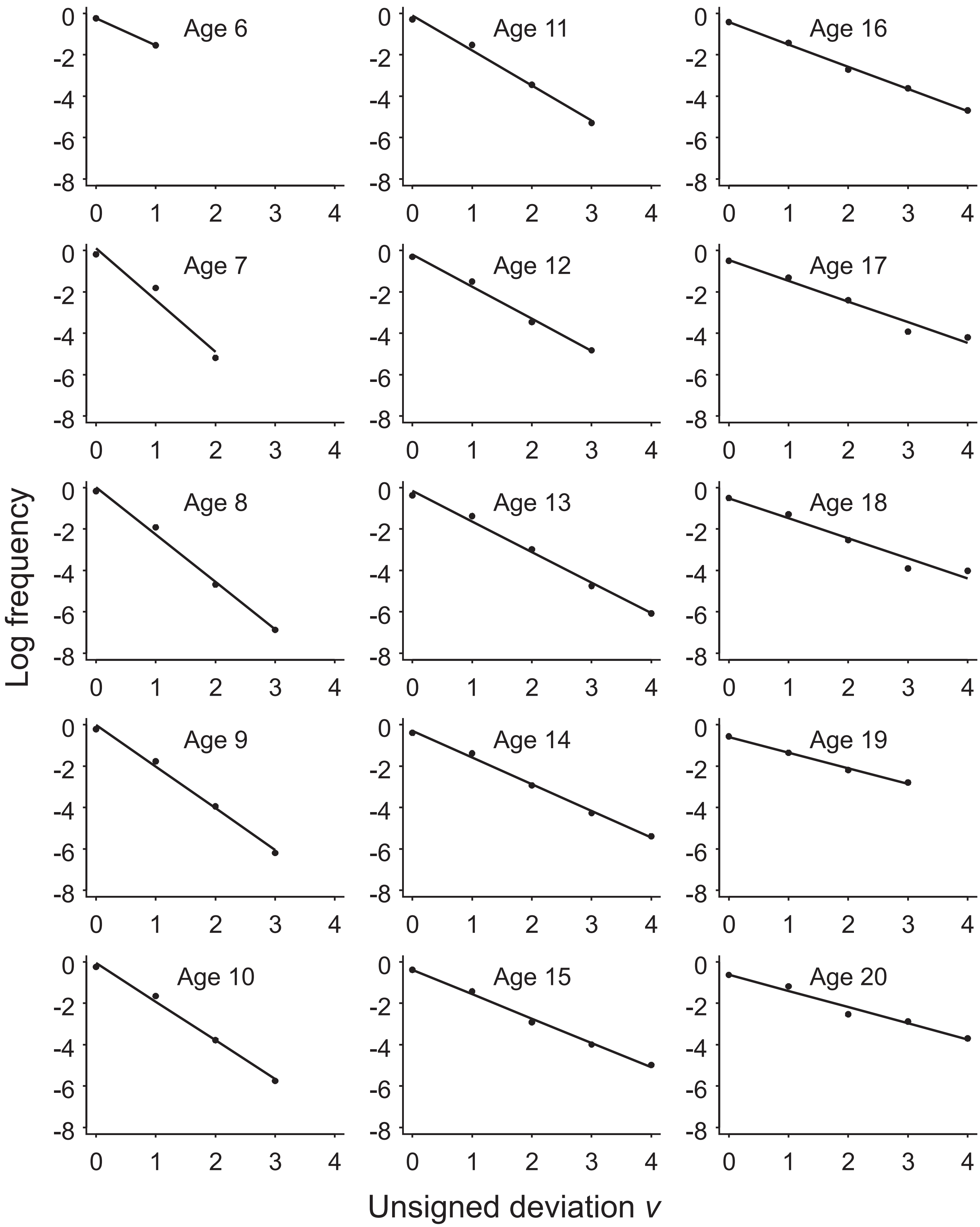
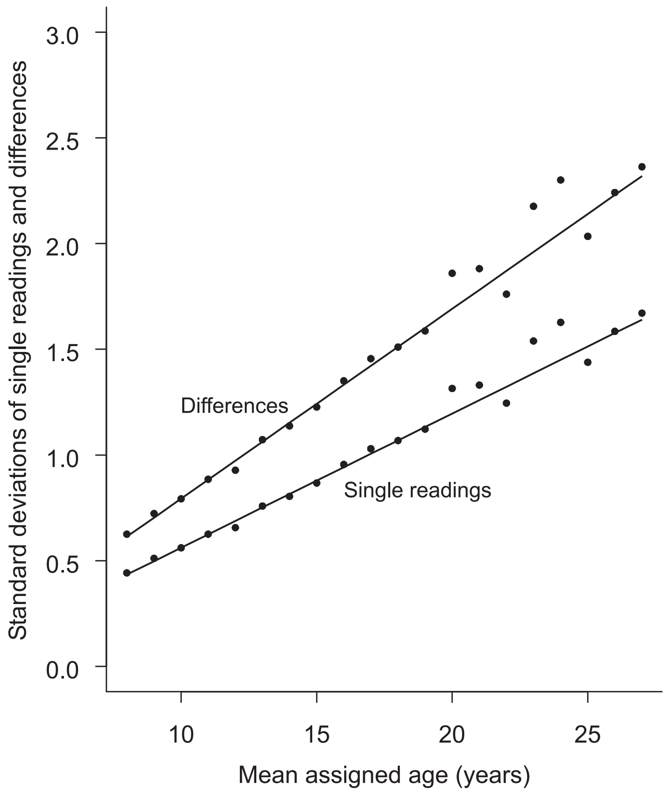


Fig. 6. The standard deviation of the difference between paired readings at each mean assigned age, and the inferred standard deviation of single readings about the canonical age.



sidedness of the distribution, the variance of the signed deviations is equal to the second moment of the geometric distribution, or (omitting the subscript A for simplicity)

$$\sigma_v^2 = (q + q^2)/p^2$$

whence after some algebra

$$p = \frac{-3 + \sqrt{1 + 8\sigma_v^2}}{2(\sigma_v^2 - 1)}$$

(The value of p at $\sigma_v^2 = 1$ is 2/3.) The standard deviations of single readings (plotted in Fig. 6) all lie on a straight line, so the misclassification probabilities for all ages of Pacific halibut can be specified concisely by saying that the unsigned deviations are geometric and the standard deviation of the signed deviations at canonical age A is $\sigma_v = -0.112 + 0.0668A$. These are the probabilities used in the simulations.

Discussion

This paper has presented a method of estimating age misclassification probabilities without making any assumption about the form of the distribution of readings (apart from symmetry). It was motivated by the large number of paired surface readings available for Pacific halibut, but it could be applied to cases with more than two readings per otolith by simply generating all possible pairs to construct the sample of differences.

A potential weakness of the method is its reliance on grouping the data by mean assigned age rather than by (unknown) canonical age to compute the estimates. This does result in bias, the size of which will depend on the actual variation with age of both abundance and misclassification probabilities. The latter can be expected to vary smoothly, so the real concern is unsmooth variation in abundance. When the data are derived from several years of sampling, the sample age compositions should be smooth and the bias should be demonstrably negligible, as in the case of Pacific halibut. A single year's data from a stock with large variations in year-class strength would be questionable, but the question could and should be answered by simulation.

A more serious limitation of the method is its need for relatively large samples — at least a few hundred pairs of readings for each estimate and preferably at least 500. One way to increase the sample size available for each age-specific estimate is to use a moving data window of 3 or more years to compute the estimate for the central age. This will reduce the variance of the estimates but increase the bias caused by grouping. If that bias is negligible to begin with, there may be little harm done.

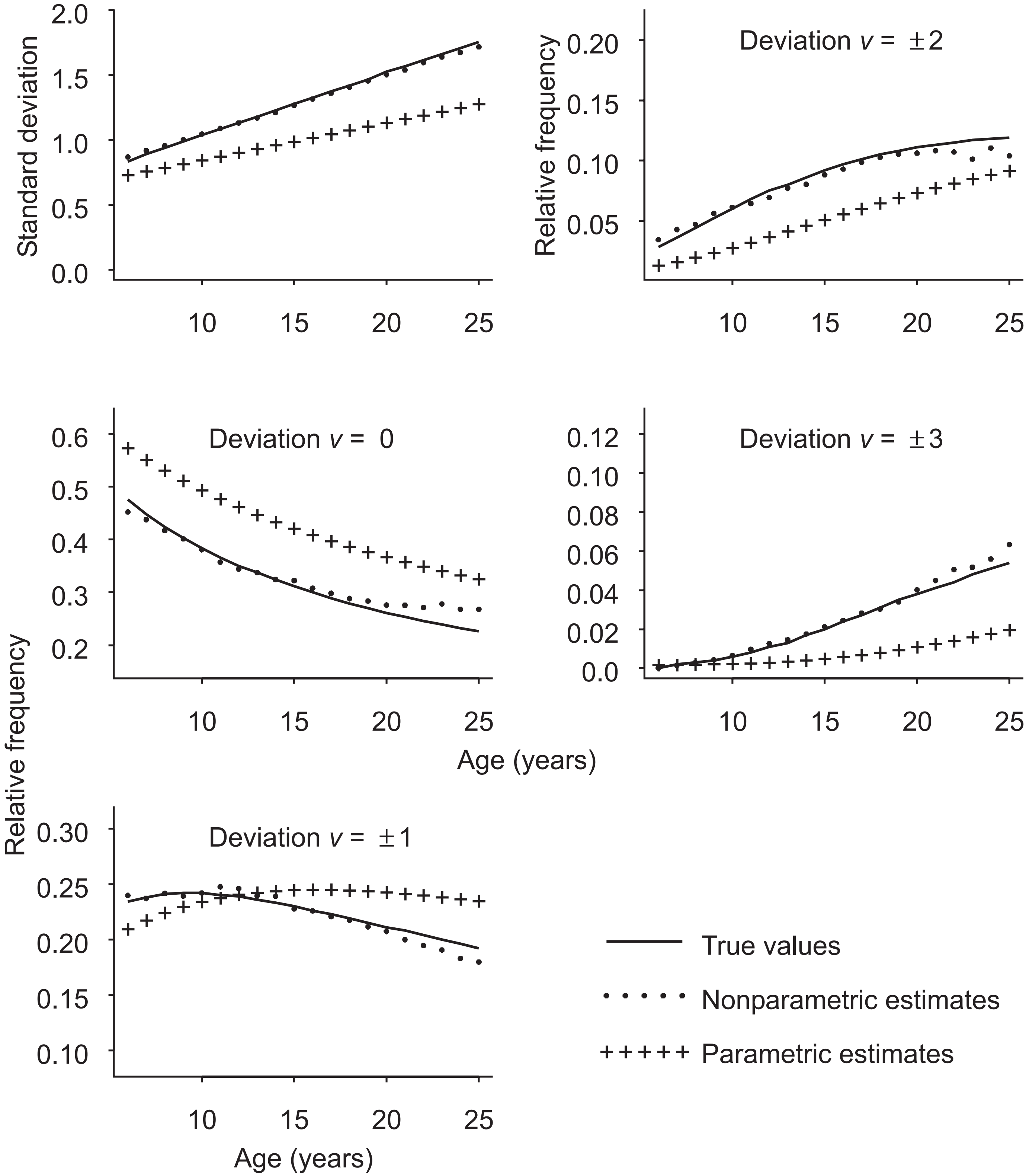
In some cases, it may be possible to identify the form of $f_A(v)$ and the way in which it varies with age in the more numerous age groups from nonparametric estimates and thereby choose a parsimonious parametric model that can be fitted to the entire data set. This was true of the IPHC surface readings and is also true for IPHC break-and-burn readings where the available number of paired readings is much smaller. (The break-and-burn readings are also geometric.)

If sample sizes at all ages are too small for nonparametric estimates, one is obliged to assume a parametric model for $f_A(v)$ at each age and fit that, or to fit alternative candidates and choose the best one according to standard model selection criteria. There is no way to know to what extent the estimates are biased by model misspecification.

Richards et al. (1992) provide a detailed account of how to compute maximum-likelihood estimates of the parameters of an assumed misclassification model and, if desired, associated estimates of the canonical age distribution of the sample or the canonical age of each fish in the sample. It is a thorough treatment of the topic except for its lack of any consideration of possible bias, which can be present in maximum-likelihood estimates even when the model is correctly specified. For example, one of the procedures outlined in the paper, and applied to paired readings by Heifetz et al. (1999), is to estimate the canonical age of each fish and use a discrete version of the normal density function to model misclassification. The procedure is easy to apply because with a normal density the maximum-likelihood estimate of the canonical age is simply the mean assigned age, and the associated likelihood is quite simple. But when applied to paired readings, it greatly underestimates their variability because the procedure essentially locates the maximum-likelihood estimates of the mean and variance of the readings of each otolith. In the case of the variance that is $\Sigma(x - \bar{x})^2/n$ rather than the usual unbiased sampling estimate $\Sigma(x - \bar{x})^2/(n-1)$. When n is large, the difference is unimportant. When the data are paired readings and n = 2, the difference is substantial. (Richards et al. (1992) had data with n = 6 readings of each otolith.)

This behavior is illustrated (Fig. 7) with simulated paired readings that actually follow the discrete version of the nor-

Fig. 7. Parametric and nonparametric estimates of the distribution of deviations of single readings computed with simulated data that follow the “normal model” described in Richards et al. (1992). Each plotted point is the mean of 100 simulated trials with a sample size of 80 000 paired readings each.



mal density: $f(v) \propto \exp(-v^2/(2\sigma^2))$. The standard deviation of this distribution is equal to the parameter σ for $\sigma \geq 0.6$. Parametric estimates of the standard deviation and the misclassification probabilities were computed with likelihood 4.3 in table 4 of Richards et al. (1992), and nonparametric estimates were computed using the method developed in this paper. Even though the model is correctly specified, almost all of the parametric estimates are very biased. Meanwhile the nonparametric estimates perform quite well without benefit of any knowledge or assumption concerning the form of the distribution.

Heifetz et al. (1999), in fact, knew the true ages of the fish in their sample because they had been marked early in life and later recaptured. The ages assigned by two readers showed considerable variability about the known ages but no bias, so the known age could be taken to be the canonical age and the authors could fit the model to known misclassification frequencies. They did that, and they also applied the procedure from Richards et al. (1992) to the paired readings using the mean assigned age instead of the known age. They were puzzled that the latter estimates of misclassification probabilities were so much lower than the former and attributed the difference to some unknown peculiarity of the paired readings, but it is more likely that the reason was simply bias in the estimation procedure. This example illustrates the importance of conducting simulation trials to verify that a numerical estimation procedure is really working correctly before applying it to real data (Hilborn and Mangel 1997).

Acknowledgements

The large number of age readings described in this paper were done by many readers in the International Pacific Halibut Commission's Age Reading Unit, supervised over the years by Edward Best, Calvin Blood, and presently Joan Forsberg. Dan Kimura and Bruce Leaman kindly read an earlier draft and made a number of suggestions that greatly improved the exposition. Geoffrey Evans' signed review of the submitted paper pointed out a number of places where there was still room for improvement.

References

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.

Forsberg, J.E. 2001. Aging manual for Pacific halibut: procedures and methods used at the International Pacific Halibut Commission. Int. Pac. Halibut Comm. Tech. Rep. No. 46.

Heifetz, J., Anderl, D., Maloney, N.E., and Rutecki, T.L. 1999. Age validation and analysis of ageing error from marked and recaptured sablefish, *Anoplopoma fimbria.* Fish. Bull. **97**: 256–263.

Hilborn, R., and Mangel, M. 1997. The ecological detective. Princeton University Press, Princeton, N.J.

Kimura, D.K., and Lyons, J.J. 1991. Between-reader bias and variability in the age determination process. Fish. Bull. **89**: 53–60.

Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. 1988. Numerical recipes in C. Cambridge University Press, Cambridge.

Reeves, S.A. 2003. A simulation study of the implications of agereading errors for stock assessment and management advice. ICES J. Mar. Sci. **60**: 314–328.

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.

Appendix A. Computing instructions.

Given paired readings of a number of otoliths from fish of canonical age A or in that vicinity (e.g., fish of mean assigned age $\overline{a} = A$), the aim is to estimate the distribution $f_A(v)$ of deviations of age readings from the canonical age. This can be done numerically by predicting and fitting the observed distribution of differences of paired readings $g_A(d)$. The following instructions detail how to tabulate the data, handle the parameter vector, and code the objective function. For simplicity, the subscript A is omitted. Any general-purpose minimizer can be used to locate the estimates.

1. Tabulate the frequencies of absolute differences of paired readings of all otoliths of mean assigned age A (or range of mean assigned ages centered at A). These absolute differences will range from 0 through, say, $d_{\max}$. Denote these frequencies n_0, n_1, n_2,…, $n_{d_{\max}}$.
2. Set the largest absolute deviation (of a single reading from the canonical age) $v_{\max}$ to $d_{\max}/2$ rounded up, e.g., if $d_{\max}$ is 5 or 6, $v_{\max}$ is 3 and $v_{\min}$ is –3.
3. The parameters to be estimated are the probabilities of deviations (from the canonical age) $f(0), f(1),\ldots, f(v_{\max} - 1)$, with the understanding that $f(-v) = f(v)$. The last value $f(v_{\max}) = f(v_{\min})$ is not estimated as a parameter because the probabilities must sum to one. A simple way to handle the working vector of estimates $\hat{f}(v)$ is to specify a parameter vector of unconstrained real numbers $x_0,\ldots, x_{v_{\max}-1}$ and run them through the logistic function $y = 1/(1 + e^{-x})$ so that they are all confined to (0,1). Call those numbers $y_0,\ldots, y_{v_{\max}-1}$ and fill in $\hat{f}(v)$ as follows:

$$\hat{f}(0) = y_0$$

$$\text{remainder} = 0.5(1 - \hat{f}(0))$$

$$\text{for } v \text{ in } 1 \text{ to } (v_{\max} - 1)$$

$$\{\text{prop} = y_v \cdot \text{remainder}$$

$$\hat{f}(v) = \hat{f}(-v) = \text{prop}$$

$$\text{remainder} = \text{remainder} - \text{prop}\}$$

$$\hat{f}(v_{\max}) = \hat{f}(-v_{\max}) = \text{remainder}$$

4. With this value of $\hat{f}(v)$, calculate the predicted distribution of differences $\hat{g}(d)$

$$\hat{g}(0) = \sum_{v_{\min}}^{v_{\max}} [\hat{f}(v)]^2$$

and

$$\hat{g}(|d|) = 2 \sum_{v_{\min}}^{v_{\max}-d} \hat{f}(v)\hat{f}(v+d) \quad \text{for } |d| > 0$$

5. Compute the multinomial log likelihood of this predicted value as

$$\log L = \text{constant} + \sum_{d=0}^{d_{\max}} n_d \log(\hat{g}(d))$$

With a general-purpose minimizer, the objective function should return –log$L$ so as to maximize log L.