PERSPECTIVE / PERSPECTIVE

Inferring age from otolith measurements: a review and a new approach

R.I.C. Chris Francis and Steven E. Campana

Abstract: In 1985, Boehlert (Fish. Bull. **83**: 103–117) suggested that fish age could be estimated from otolith measurements. Since that time, a number of inferential techniques have been proposed and tested in a range of species. A review of these techniques shows that all are subject to at least one of four types of bias. In addition, they all focus on assigning ages to individual fish, whereas the estimation of population parameters (particularly proportions at age) is usually the goal. We propose a new flexible method of inference based on mixture analysis, which avoids these biases and makes better use of the data. We argue that the most appropriate technique for evaluating the performance of these methods is a cost–benefit analysis that compares the cost of the estimated ages with that of the traditional annulus count method. A simulation experiment is used to illustrate both the new method and the cost–benefit analysis.

Résumé : Boehlert a indiqué en 1985 (Fish. Bull. **83**: 103–117) que l'âge des poissons pouvait être déterminé à partir de mesures des otolithes. Depuis lors, plusieurs techniques d'inférence ont été proposées et évaluées sur une gamme d'espèces. Une revue de ces techniques montre que toutes sont soumises à au moins un de quatre types de biais. De plus, toutes cherchent à assigner un âge à des poissons individuels, alors que le but est l'estimation des variables démographiques, en particulier la proportion d'individus à chacun des âges. Nous proposons une nouvelle méthode flexible d'inférence basée sur l'analyse des mélanges qui évite ces biais et qui fait un meilleur usage des données. Nous croyons que la technique la plus appropriée pour évaluer la performance de ces méthodes est une analyse de coûts–bénéfices qui compare le coût des âges estimés avec celui de la méthode traditionnelle du dénombrement des annulus. Une expérience de simulation permet d'illustrer tant la nouvelle méthode que l'analyse coûts–bénéfices.

[Traduit par la Rédaction]

Introduction

Around the world, the ages of close to a million fish are determined each year using otoliths, largely in support of harvest calculations (Campana and Thorrold 2001). Fish age is generally determined after initial preparation of the otolith (such as embedding and thin sectioning) followed by microscopic examination and counts of the annual growth zones (annuli). The preparation process is often time consuming, while the interpretation of the annuli requires skilled technicians. As a result, the process of age determination is reasonably expensive. To minimize time and expense, many agencies take small subsamples of catches or populations for age estimates, producing age–length keys that are used to infer the age composition of the remainder of the catch based on a larger sample of simple length measurements (Kimura 1977).

Although age–length keys rely on the relationship between age and fish length, an alternative approach is to take advantage of the well-documented proportionality between the size of the otolith and both the size and the age of the fish (Templeman and Squires 1956). Although the sizes of the fish and the otolith are correlated, otolith size tends to be somewhat more correlated with fish age than is fish length (Boehlert 1985). Thus, in principle, otolith size can better be used to infer fish age than can fish length. A number of studies have statistically related various measurements of otolith size (e.g., otolith weight, length, area) to the annulus-based age and then used the resulting relationships to estimate the age composition of the remaining, unaged fish (Boehlert 1985; Pawson 1990; Worthington et al. 1995*a*). A common feature shared by this approach and that of the age–length key is that both require two samples: a "calibration" and a "production" sample. The calibration sample (sometimes called the training sample) is used to define a procedure for estimating age, and this procedure is then applied to all fish in the production sample (sometimes called the test sample). The ages of fish are known in the calibration sample but not in the production sample. The motivation for this two-stage

Received 17 October 2003. Accepted 6 February 2004. Published on the NRC Research Press Web site at http://cjfas.nrc.ca on 17 September 2004.
J17802

R.I.C.C. Francis.[1] National Institute of Water and Atmospheric Research, Private Bag 14901, Wellington, New Zealand.
S.E. Campana. Marine Fish Division, Bedford Institute of Oceanography, P.O. 1006, Dartmouth, NS B2Y 4A2, Canada.

[1]Corresponding author (e-mail: c.francis@niwa.co.nz).

 doi: 10.1139/F04-063

approach is simple — the first stage involves expensive annulus-based age determinations, while the second stage does not.

An obvious question is why is so much time and money invested into age determinations? By far the most common products of age determinations are catch proportions at age for use in stock assessments. The second most common outputs would be growth parameters. Thus, the goal is nearly always to estimate the growth or mortality parameters of a fish population, not to estimate the ages of individual fish (Pauly 1987). Indeed, it appears that there is little point in using otolith measurements to assign ages to individual fish because the probability of correct assignment is often quite low. However, the same suite of measurements could be used to provide more accurate estimates of population parameters. Later, we will argue that the literature has placed too much emphasis on individual age estimation and too little on the estimation of population parameters associated with age. One consequence of this is that techniques that directly estimate proportions at age (without assigning ages to individual fish) have been overlooked. Ironically, most of the published techniques that we reviewed assigned individual ages but then went on to calculate proportions at age. In many cases, this has led to inappropriate methods for evaluating the power of using otolith measurements.

In this paper, we start by discussing the observations that have provided a rationale for using otolith measurements to infer age and then describe the various types of bias that can occur. Next, we critically review the published methods for inferring age based on otolith measurements and then present a new approach for directly estimating proportions at age. This new method takes full advantage of the information in both the calibration and production sample but avoids the asymptotic bias that characterizes other methods. We conclude with a review of methods of performance evaluation, suggesting that cost–benefit analyses are a necessary part of any evaluation, as demonstrated by a simple illustrative example.

Throughout this paper, we will mostly treat age as a discrete variable. That is, a reference to fish of age 2 will mean fish from the 2+ age class (unless otherwise stated). There are circumstances when it might be better to think of age as continuous (e.g., when estimating growth parameters using samples gathered throughout the year). However, in the literature that we are reviewing, people were almost always interested in discrete ages only. This is also true throughout fisheries science. Some of the methods that we review below are easily applied to continuous ages but others are not.

In referring to various measurements, we will use the notation W for weight, L for length, w for width, and T and thickness and use a subscript to define what is being measured: "O" for the otolith and "F" for the fish body. The most used of these measurements will be otolith weight (W_O) and fish length (L_F).

Why should otolith measurements predict age well?

In this section, we describe, and comment on, two types of observations on fish growth that provide what Secor and Dean (1989) called "a biological rationale for the use of otolith size and fish size as predictors in age estimation".

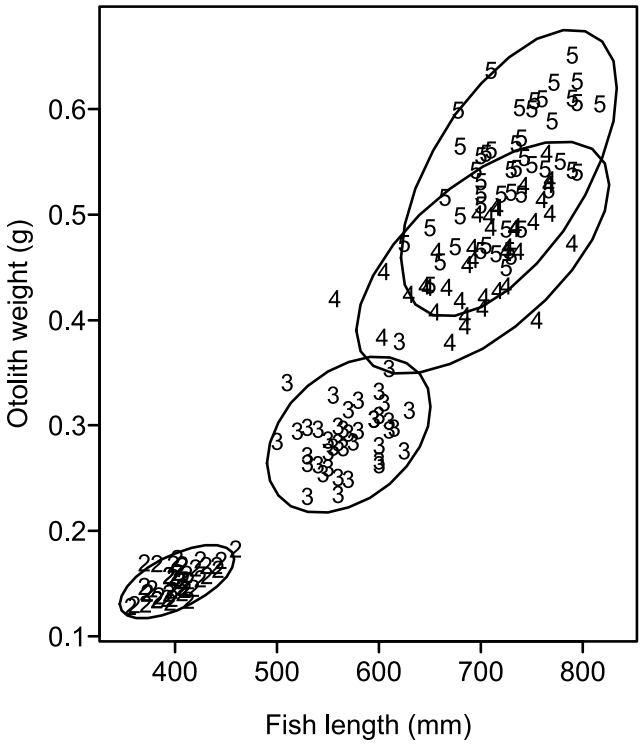
The first type of observation is what we shall call the Templeman–Squire relationship (TSR), which is that among fish of the same length, the older fish tend to have bigger otoliths. The first record of TSR that we are aware of was for haddock (Templeman and Squires 1956), but it has since been reported for many species (e.g., Reznick et al. 1989; Wright et al. 1990; Fossen et al. 2003). In these observations, otolith size was usually measured as W_O, but other otolith dimensions have been used. A common way of characterising TSR is to say that, among fish of the same length, slow growers tend to have heavier otoliths than fast growers. It is important to specify "of the same length" here because when we compare fish of the same age, it is the fast growers (i.e., the large fish) that have the heavier otoliths. That is, within-age-group correlations between W_O and L_F are usually positive (Pawson 1990; Fletcher and Blight 1996; Cardinale et al. 2000).

The second type of observation is that of continuing growth of the otolith. Many authors have noted that as fish grow older, growth in L_F, L_O, and w_O all slow down, but T_O and W_O keep increasing because of continued deposition of material on the medial surface of the otolith (Blacker 1974; Boehlert 1985; Anderson et al. 1992). This growth pattern explains TSR in older fish. It also explains why W_O has been shown to be by far the most important of otolith measurements for inferring age. All studies that we have seen have used either W_O alone or a combination of W_O and other measurements (including, sometimes, body measurements like L_F). When several otolith measurements are compared with age, the highest correlation is usually with W_O (Boehlert 1985; Fossen et al. 2003).

What inference should we take from these observations? The common view seems to be that W_O is a promising potential predictor of age, but this may be going too far. Consider a typical plot of W_O and L_F, which clearly illustrates TSR (Fig. 1). For this population, fish of length 600 mm are very likely to be of age 3 or 4, and we can decide, with reasonable confidence, which age they are if we know their W_O. But it is important to note that we make this inference about the age of the fish using both W_O and L_F, not with W_O alone. Thus, what the above observations suggest is that the combination of W_O and L_F contains more information about age than does L_F alone. It does not necessarily imply that W_O by itself is a good predictor of age. In fact, it does not even imply that W_O is a better predictor than L_F (although this will often be true). To illustrate this point, we need to make a definition.

A separation index is one way of quantifying how well we might be able to infer age. It is defined for each pair of adjacent ages and measures how much overlap there is between them. For example, for the measurement W_O at ages A and $A + 1$, the separation index $S_{A,A+1}$ is defined by $S_{A,A+1} = (\mu_{A+1} - \mu_A)/\sigma_{A,A+1}$, where μ_A is the mean W_O for fish of age A and $\sigma_{A,A+1}$ is the pooled standard deviation of W_O for ages A and $A + 1$. (Of the two slightly different ways that people have calculated $\sigma_{A,A+1}$, $[0.5(\sigma^2_{A+1} + \sigma^2_A)]^{0.5}$ and $0.5(\sigma_{A+1} + \sigma_A)$, the former seems better on theoretical grounds (Snedecor and Cochran 1980), but the difference is probably not often important.) We can convert the separa-

Fig. 1. Otolith weight against fish length for known-age wild Icelandic cod. Each point represents one fish and the plotting symbol identifies its age. The ellipses are 95% confidence regions for bivariate normal distributions fitted to the data for each age.



tion index into an approximate probability of correct age estimation using the formula $P_{\text{correct}} = 2F(S/2) - 1$, where F is the cumulative distribution function of the standard normal (the formula is exact if we assign an age to each fish according to which μ_A its W_O is closest to and if we assume normal distributions with equal variances and no variation with age in either the proportions at age or the separation index). The larger the separation index, the higher P_{correct} is, and thus the better the predictor (Fig. 2).

The separation indices in Table 1 show that W_O is not always a better predictor than L_F (it is not for the lowest ages for either stock). Further, and more importantly, for both cod (*Gadus morhua*) stocks, the two predictors combined are better than either one singly. (The combined separation index is calculated using that linear combination of W_O and L_F that, according to a linear discriminant analysis, best separates the two age groups.) These results support the view of Brander (1974), who said "a two-dimensional separation, using otolith weight as well, may give a better separation of ages than [body] length alone".

Four types of bias

We now define four different ways in which estimated proportions at age can be biased. It should be noted that bias in an estimator is not necessarily a bad thing. For example, a precise estimator with small bias may be preferable to an unbiased but imprecise estimator. However, asymptotic bias (i.e., bias that does not tend to zero when sample sizes become large) is a most undesirable property in an estimator (this is called inconsistency in the statistical literature; Stuart

Fig. 2. Illustration of four values of the separation index S. In each panel, the thin lines denote the distributions (with means μ_1 to μ_4) of some predictor (like otolith weight (W_O) or fish length (L_F)) at ages 1 to 4, and the thick line denotes the combined distribution. P_{correct} is the approximate proportion of ages correctly estimated with the given value of S.

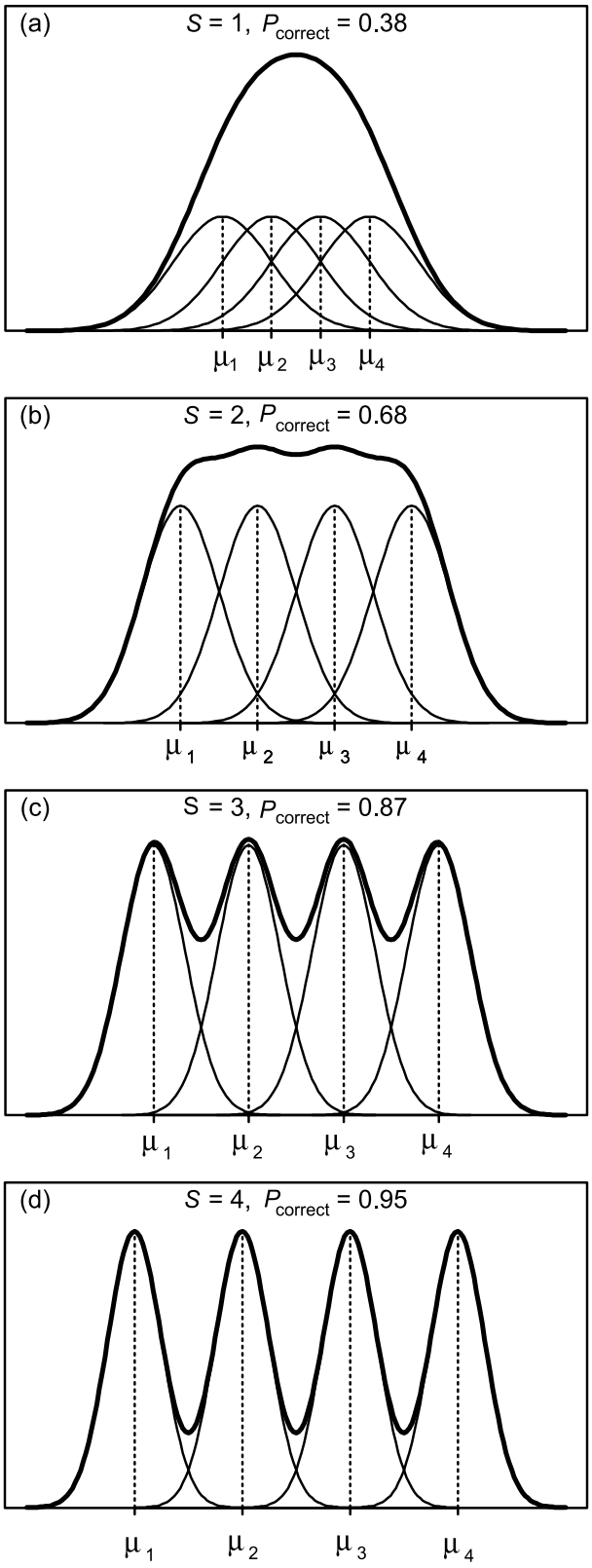


Table 1. Separation indices for the predictors otolith weight (W_O) and fish length (L_F) and the combination W_O and L_F calculated for Irish Sea and Icelandic cod.

		Separation index		
Species	Ages (years)	W_O	L_F	W_O and L_F
Irish Sea cod	1–2	3.14	3.31	4.19
	2–3	2.90	2.31	4.53
	3–4	2.23	1.51	3.66
Icelandic cod	2–3	5.75	5.80	6.71
	3–4	4.32	2.98	4.38
	4–5	1.47	0.52	1.52

Note: The values for Irish Sea cod are from table 6 of Brander (1974); those for Icelandic cod are from the data in Fig. 1.

and Ord 1991). Maximum-likelihood estimators may sometimes be biased but are never asymptotically biased. We will show that all the estimation methods that we review below are subject to asymptotic bias in at least one of our four types.

Before defining these biases, it is useful to describe a simple model, with just one predictor of age (W_O, say). Suppose we have a fish population with ages between 1 and n in which the proportion at age A is p_A and, for fish of age A, W_O is normally distributed with mean μ_A and standard deviation σ_A. From a calibration sample, we can easily make estimates of these parameters: $\hat{p}_A$, $\hat{\mu}_A$, and $\hat{\sigma}_A$. We measure W_O for each fish in a production sample and want to assign an age to each fish in this sample. To do this we need a "cutting rule", which is based on our parameter estimates $\hat{p}_A$, $\hat{\mu}_A$, and $\hat{\sigma}_A$ and which cuts the W_O distribution into n parts and assigns an age to each fish according to which part W_O lies in. Most of the methods that we review use simple cutting rules that are defined by a set of $n - 1$ cutoff points, $c_1,..., c_{n-1}$. A fish is assigned to age 1 if $W_O < c_1$, to age A if $c_A < W_O < c_{A+1}$, and to age n if $W_O > c_{n-1}$.

So what is the "best" cutting rule for our simple model? There are (at least) three answers to this question, depending on how we want to define "best". The "midpoint rule", for which $c_A = 0.5(\hat{\mu}_A + \hat{\mu}_{A+1})$ (Fig. 3*a*), is best only in the sense of being the simplest rule. It is a poor rule because it ignores the other parameters, $\hat{p}_A$ and $\hat{\sigma}_A$. The cutting rule provided by discriminant analysis is best in the sense that each fish is assigned the age that is most likely, given the calibration sample. We call this the MLA (most likely age) rule. For linear discriminant analysis (which assumes homoscedasticity, i.e., $\sigma_A = \sigma$ for all A), this is a simple cutting rule for which c_A can be calculated by solving the equation

$$\hat{p}_A f[(c_A - \hat{\mu}_A)/\hat{\sigma}_{A,A+1}] = \hat{p}_{A+1} f[(c_A - \hat{\mu}_{A+1})/\hat{\sigma}_{A,A+1}]$$

where f is the probability distribution function of the standard normal distribution (we discuss other variants of discriminant analysis below). It has a simple graphical interpretation: the cutting point between two ages is the point at which the two distribution functions intersect (Fig. 3*b*). The MLA rule is known to produce asymptotically biased estimates of proportions at age (McLachlan and Basford 1988), so it will not be the best rule if our aim is to use our assigned ages to estimate these proportions. The bias associated with this rule is our first type of bias, which we call "discriminant bias". It differs from the other types discussed below in that it is associated with a specific cutting rule. A simple alternative rule, which does not produce this

Fig. 3. Illustration of the application of three cutting rules (midpoint, most likely age (MLA), and unbiased proportions at age (UPA)) for otolith weight to a simple population with two age classes (ages 1 and 2 in proportions 0.3 and 0.7, respectively). The curved lines illustrate the assumed distribution of otolith weight for ages 1 (mean = 10, SD = 5) and 2 (mean = 20, SD = 5); the two shaded areas in each panel correspond to fish that are assigned the wrong age by the cutting rule, and the number within each shaded area is the associated percentage of the population (when these numbers are unequal, the proportions at age estimated using the rule will be biased).

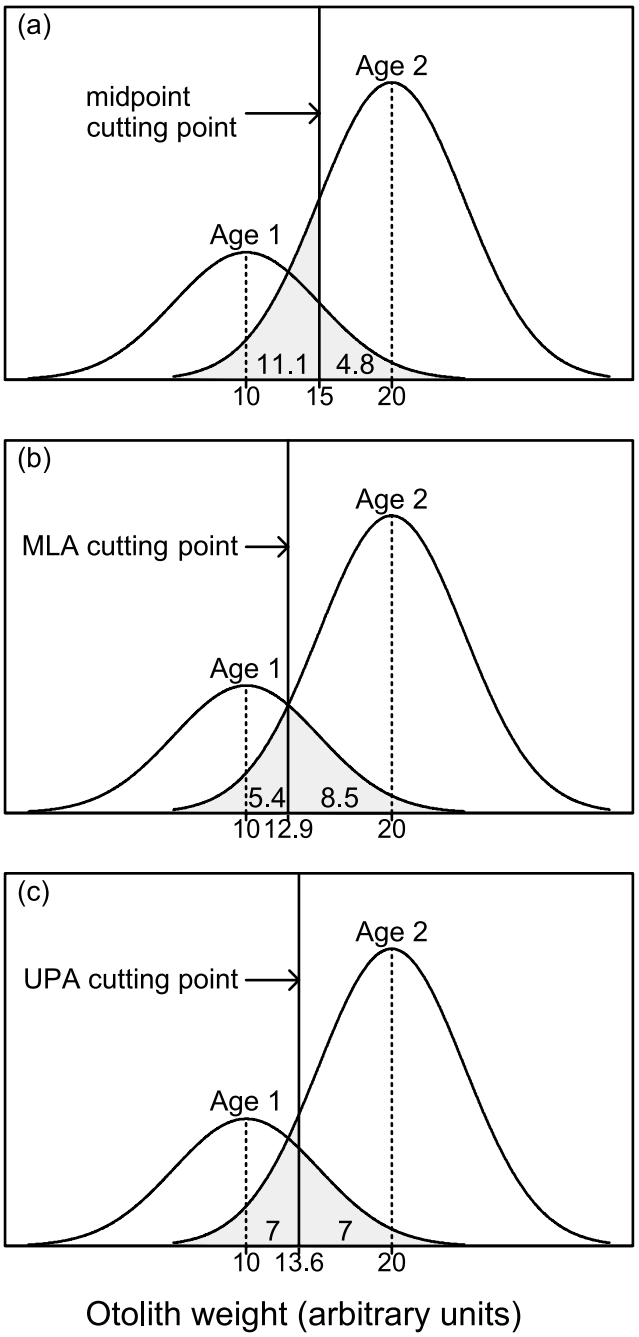


Table 2. Illustration of four types of bias in estimating p_1 (the proportion at age 1 in an artificial population in which there are just two age classes) with proportions at age given by $p_1 = 0.3$ and $p_2 = 0.7$ and the standard deviation of otolith weight at age given by σ_1 and σ_2 (unless otherwise stated, $\sigma_1 = \sigma_2$).

			Expected bias in p_1		
Type of bias	Cutting rule	Cause	$S = 1$	$S = 2$	$S = 3$
Discriminant bias	MLA	Use of MLA rule	–0.13	–0.03	–0.01
Smoothing bias	UPA	Variation in p_A ignored ($p_1 = 0.3$ but assume that $p_1 = p_2$)	0.12	0.06	0.03
Heteroscedastic bias	UPA	Variation in σ_A ignored ($\sigma_2 = 1.5\sigma_1$ but assume that $\sigma_1 = \sigma_2$)	–0.01	–0.03	–0.02
Calibration bias	UPA	p_A in calibration sample differs from that in population ($p_1 = 0.7$ in calibration sample, $p_1 = 0.3$ in population)	0.26	0.13	0.06

Note: The tabulated values are the approximate expected bias in estimates of p_1 for each of three values of S.

bias (at least asymptotically), is what we call the UPA (unbiased proportions at age) rule for which we calculate c_A by solving

$$\sum_{A'=1}^{n} \hat{p}_{A'} F[(c_A - \hat{\mu}_{A'})/\hat{\sigma}_{A'}] = \sum_{A'=1}^{A} \hat{p}_{A'}$$

where F is the cumulative distribution function of the standard normal distribution (Fig. 3*c*). The MLA rule generalizes straightforwardly to the case of multiple predictors, but this is not true (to our knowledge) for the UPA rule.

The other three types of bias are not associated with any particular cutting rule. They will occur if we use the wrong information in constructing our cutting rule. If we ignore variation in p_A (i.e., we make our cutting rule assuming that all p_A are equal when this is not true), we will tend to underestimate strong age groups and overestimate weak ones. This has the effect of smoothing the estimated age frequency, so we call this a "smoothing bias". It is exactly analogous to the bias induced by ageing error. If ageing error is symmetric, a small age class followed by a large age class will tend to be overestimated because the number of the older age class that will be underaged will be greater than the number of the younger age class that will be overaged. If we ignore heteroscedasticity (i.e., assume that all σ_A are equal when this is not true), we generate what we will call "heteroscedastic bias", which overestimates the size of the age groups with the smaller σ_A. Finally, "calibration bias" occurs if the proportions at age in the calibration sample are not representative of those in the population and this is not allowed for in the cutting rule. In this case, the estimated values of p_A will tend to be between the true values and those in the calibration sample.

All four types of bias for the simple example with $n = 2$, $\sigma_1 = \sigma_2$, and $p_1 = 0.3$ (so $p_2 = 0.7$) are illustrated (Table 2). For all bias types except heteroscedastic, the extent of bias depends strongly on the degree of separation and is small when S is large. For example, with a separation index $S = 1$, the expected value of the estimated proportion at age 1 when the MLA rule is used is only 0.17, so the discriminant bias is –0.13 (= 0.17 – 0.30). However, as S increases to 2 and then 3, this bias reduces to –0.03 and then –0.01. Thus, if there is little or no overlap between age groups, these biases will not be serious. Note that the biases in Table 2 are approximate in that they ignore uncertainty in the μ_i and σ_i. They are calculated using

$$\hat{p}_1 = p_1 F[(c_1 - \mu_1)/\sigma_1] + p_2[1 - F((c_1 - \mu_2)/\sigma_2)]$$

Note also that in practice, a mixture of biases may occur. For example, the simplest cutting rule is to use the midpoint between modes, i.e., $c_A = 0.5(\mu_A + \mu_{A+1})$. This will cause both smoothing and heteroscedastic bias if in fact $p_A \neq p_{A+1}$ and $\sigma_A \neq \sigma_{A+1}$.

The model that we have used in this section to describe how W_O varies within a population is called a "mixture model". The distribution of W_O is thought of as a mixture of n normal distributions, one for each age class, with mixing proportions p_A and with each distribution characterized by its parameters μ_A and σ_A. This is a useful model, which we return to below. It is easily extended to higher dimensions (for example, Fig. 1 may be thought of as a pictorial representation of a mixture of bivariate distributions). It can also be generalized by using other distributions in place of the normal.

Published methods of inferring age

The regression method of Boehlert (1985) appears to be the first published technique for estimating age from otolith measurements. Boehlert assembled a suite of potential predictors (including W_O, L_O, w_O, and their squares and cubes and the "interaction variables" W_O/L_O and L_O/w_O) and used forward stepwise linear regression to select those that were the best predictors of age and to construct a regression equation. This equation is constructed from the calibration sample (where ages are known) and then applied to the production sample (in which we know only the values of the predictor variables) to obtain an estimate of age for each fish. Boehlert was implicitly using discrete ages, so the age estimated from the regression was rounded to the nearest integer. However, the regression method could equally well be used with continuous ages.

The regression method is very appealing in its simplicity but has two drawbacks. It will often be necessary to transform predictors and (or) the predictand to obtain linear relationships, and even then, this is likely to achieve only approximate linearity. Some authors have not felt the need for transformations (e.g., Anderson et al. 1992; Ferreira and Russ 1994; Labropoulou and Papaconstantinou 2000); others have used logarithmic (Worthington et al. 1995*a*; Cardinale et al. 2000) or power (Cardinale et al. 2000; Luckhurst et al. 2000) transformations or both. The second, and more serious, drawback is that this method produces asymptotically biased estimates of proportions at age. We illustrate this

using our normal mixture model in the following simple scenario.

We assume a single predictor, W_O, say, which is linearly related to age, A, so the mean of W_O for fish of age A is $a + bA$. It will be convenient to write the standard deviation of W_O for fish of age A as $b\sigma_A$. We will distinguish between three proportions at age: p_A, the true proportion at age A in the fish population of interest; $p_{c,A}$, the expected proportion at age A in the calibration sample; and $p_{p,A}$, the expected proportion of the production sample that will be assigned age A. Thus, we will show bias if we find that $p_{p,A} \neq p_A$. We allow $p_{c,A}$ to differ from p_A because some people construct their calibration samples so as to get a more even spread of fish sizes than would arise from a simple random sample (Worthington et al. 1995*a*; Araya et al. 2001; Pilling et al. 2003). But we will need to assume that however the calibration sample is constructed, all fish of a given age have the same probability of being selected. The production sample is assumed to be a simple random sample from the population, so the expected proportion at age A in this sample is p_A. Given these assumptions, we can derive approximate formulae for $p_{p,A}$ (Appendix A). (These formulae are large-sample approximations, which become closer and closer to being exact as the size of the calibration sample tends to infinity).

Evaluation of these formulae for some specific scenarios shows that the regression method, as used by Boehlert (1985), produces three of the types of bias described above plus another one. In these scenarios, we will assume homoscedasticity unless otherwise stated. This means that the separation index for W_O is independent of age and given by $S = 1/\sigma$. We will also assume that all fish in the population are known to be of age between 1 and 5, inclusive (which means, for example, that if the regression estimate of age is 5.9, this would be rounded down to 5 rather than up to 6). Four scenarios were considered, each chosen to illustrate a different type of bias (Table 3). When all age classes are the same size, we see that estimated proportions at age are biased down for the younger and older ages and up for the middle ages and that the extent of bias increases as the separation index decreases (Fig. 4*a*). Bias is small when age groups are well separated ($S = 3$) but substantial when separation is poor ($S = 1$). This sort of bias, which we will call "centric bias" (because bias is towards the centre of the age distribution), is well known in regression situations. When Y is regressed on X, estimates of Y for large (small) values of X are negatively (positively) biased, and the extent of bias increases as the correlation between X and Y decreases. It is more difficult to illustrate our three other types of bias because we cannot avoid centric bias when using this regression method. In each of scenarios 2 to 4, variant 1 (which is identical to variant 2 of scenario 1) involves only centric bias, and variants 2 and 3 involve this bias plus increasing amounts of the bias that is illustrated by the scenario. The results suggest that, for Boehlert's regression method, smoothing bias (Fig. 4*b*) is likely to be quite significant, calibration bias (Fig. 4*c*) may be minor, and heteroscedastic bias (Fig. 4*d*) may be of intermediate severity.

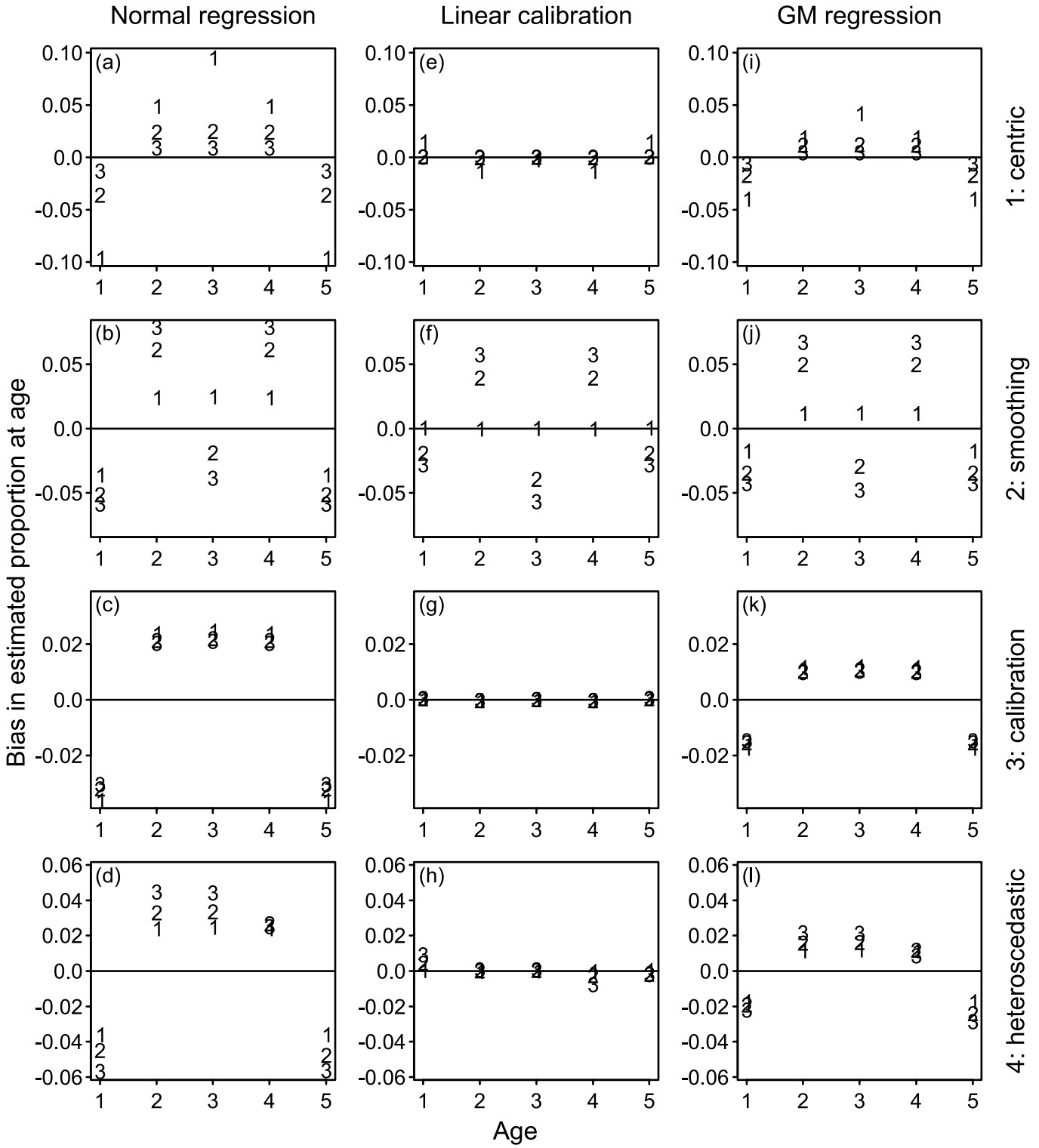
In situations where there is only a single predictor (W_O, say), some authors have queried whether it is appropriate to use the usual predictive regression of A on W_O. Two alternatives have been proposed. Worthington et al. (1995*a*) suggested that it would be better to use the opposite regression. The usual theoretical justification for using this approach (which is called linear calibration) does not seem to apply here. Stuart and Ord (1991) pointed out that regressing X on Y provides maximum-likelihood estimation of the regression coefficients conditioned on the values of X. This suggests that linear calibration might be the theoretically correct regression if the calibration sample were constructed by, say, choosing at random 20 fish from each of a selected set of age classes. But this is not a common method of creating a calibration sample (because the age of fish is not generally known when this sample is selected). Whatever the theoretical justification, this approach does appear to produce much less bias. Smoothing bias is reduced (Fig. 4*f*), centric and heteroscedastic biases are almost nonexistent (Figs. 4*e* and 4*h*), and there is no calibration bias (Fig. 4*g*). The second alternative to the usual regression of A on W_O is to use the geometric mean (GM) regression (Ricker 1973). Pilling et al. (2003) suggested that this was preferable to the normal regression model because both age and its predictor are likely to be measured with error. With GM regression, we do not need to ask whether we should regress W_O on A, or vice versa, because we get exactly the same results with both regressions. The bias from this method is (at least in our examples) intermediate between that for standard regression and linear calibration (Figs. 4*i*–4*l*). The two alternative regressions (linear calibration and GM) seem limited in that they allow only one predictor variable. This could perhaps

Table 3. Four scenarios used in illustrating bias in the regression method in Fig. 4.

Scenario	Type of bias illustrated	Variant assumptions
1	Centric	1: $S = 1$
		2: $S = 2$
		3: $S = 3$
2	Smoothing	1: $p_A = p_{c,A}$ = (0.20, 0.20, 0.20, 0.20, 0.20)
		2: $p_A = p_{c,A}$ = (0.25, 0.13, 0.25, 0.13, 0.25)
		3: $p_A = p_{c,A}$ = (0.27, 0.09, 0.27, 0.09, 0.27)
3	Calibration	1: $p_{c,A}$ = (0.20, 0.20, 0.20, 0.20, 0.20)
		2: $p_{c,A}$ = (0.25, 0.13, 0.25, 0.13, 0.25)
		3: $p_{c,A}$ = (0.27, 0.09, 0.27, 0.09, 0.27)
4	Heteroscedastic	1: σ_A = (0.50, 0.50, 0.50, 0.50, 0.50)
		2: σ_A = (0.50, 0.54, 0.57, 0.61, 0.65)
		3: σ_A = (0.50, 0.57, 0.65, 0.73, 0.80)

Note: Each scenario illustrates one type of bias and has three variants. Note that variant 2 in scenario 1 is identical to variant 1 in scenarios 2 to 4. Unless otherwise specified, p_A = (0.20, 0.20, 0.20, 0.20, 0.20), $p_{c,A}$ = (0.20, 0.20, 0.20, 0.20, 0.20), σ_A is independent of age, and $S = 2$ for all ages.

Fig. 4. Illustration of bias in estimation of proportions at age using three regression methods ((*a–d*) normal regression, (*e–h*) linear calibration, and (*i–l*) GM regression) and scenarios illustrating four types of bias (centric (top row), smoothing (second row), calibration (third row), and heteroscedastic (bottom row)). In each panel, the plotting symbols 1, 2, and 3 show the bias for the three variants within each scenario.



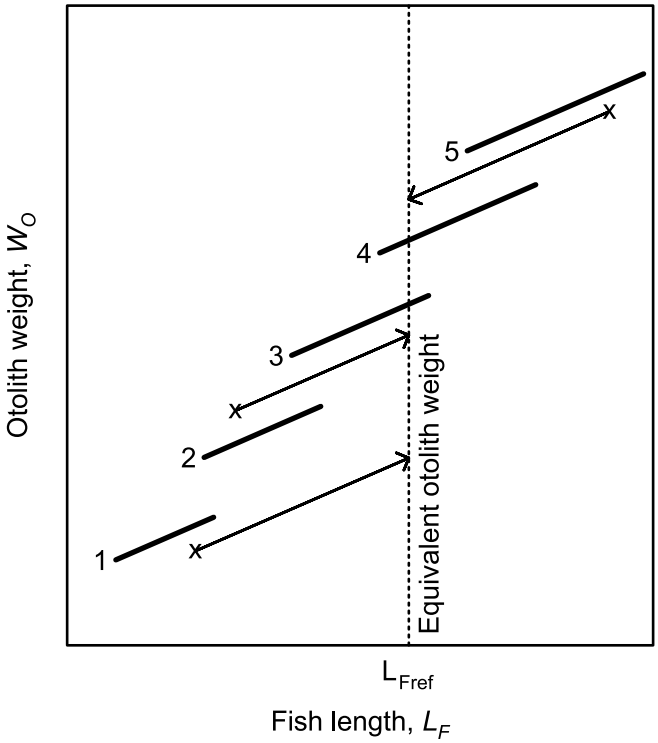
be overcome by a two-stage procedure: first, use Boehlert's (1985) approach to find the best linear combination of predictors and then use this combination as a single predictor in linear calibration or GM regression. However, it would be better to find a method of inferring age that was not subject to smoothing bias.

Pawson (1990) proposed a new method of estimating age from W_O and L_F. This requires the assumption that when W_O

is regressed on L_F for each age group, the regression slope, b, is the same for all age groups. The first step is to calculate "equivalent" otolith weights, W_{Oe}, for each fish using some common reference length, L_{Fref}, via the formula $W_{Oe} = b(L_{Fref} - L_F) + W_O$. This calculation can be thought of as a projection in the W_O–L_F plane along lines of slope b (Fig. 5). For the sardine data analysed by Pawson, the age groups show very little overlap when the calibration sample is plotted as W_{Oe} against L_F (fig. 5 in Pawson (1990)). Thus, for this species, ages could be assigned unequivocally to most fish in a production sample by making a W_{Oe}–L_F plot. There would be only a few fish for which there was some doubt (those that lay in an area of overlap between age groups in the calibration sample W_{Oe}–L_F plot). (Note that it does not matter what value of L_{Fref} is used (as long as the same value is used for all fish); changing to a different value is equivalent to applying a linear transformation to W_{Oe}, which does not change the degree of overlap in the W_{Oe}–L_F plot.) There are two problems with this method. First, it will have limited use because it requires the strong assumption that the slope b is the same for all age groups; although there may be cases where this assumption holds, two data sets that we have examined (including that in Fig. 1) show this slope increasing with age. Second, it is not specified exactly what cutting rule should be used to deal with any overlap between age groups (which was slight for Pawson's sardines but may be substantial for other species). Pawson referred to the construction of W_{Oe}–L_F keys, but the details of this method are unclear. As we have shown in the previous section, the choice of cutting rule is important. Until we know what cutting rule is proposed, we cannot evaluate Pawson's method any further.

A third method of age inference, modal analysis, was used by Fletcher (1995). This differs from all other proposed methods in that it uses no age data and thus does not require a calibration sample. The problem addressed is the same as has been considered by many authors who have estimated proportions at age from multiple L_F samples. The only difference is that Fletcher used W_O in place of L_F. Otolith weights were measured from random samples taken at 3-month intervals from catches in a west Australian pilchard fishery. The histogram of W_O for each sample showed a series of modes, and these modes were seen to move to the right from sample to sample in a way that suggested that each mode was associated with an age class. The modal decomposition software MIX (MacDonald and Green 1988) was used to find the location (i.e., the mean) of each mode, the position of modes was averaged between years, and an age class was assigned to individual fish according to which W_O mode it was closest to. Estimated ages were then aggregated to calculate proportions at age in the catch. Although the idea of inferring ages from multiple W_O samples is promising, an alternative to the analytical method used by Fletcher would provide more accurate results. We have seen above that choosing the midpoint between modes invites smoothing and heteroscedastic bias. Averaging modal positions between years will cause calibration bias if there is any between-year variation in otolith growth. In addition, there is no need to assign ages to individual fish because MIX directly estimates proportions at age (and these, rather than individual ages, were a stated objective of this study). A

Fig. 5. Illustration of the calculation of equivalent otolith weight (W_{Oe}) in the method of Pawson (1990). The parallel thick solid lines are from the regression of otolith weight (W_O) on fish length (L_F) for each age from 1 to 5; the arrows illustrate how points representing individual fish (×) are projected onto the broken line ($L_F = L_{Fref}$) with the height of the resultant point being the value of W_{Oe} for the fish.



weakness of MIX is that it analyses each histogram separately and thus cannot use information from the adjacent samples to help locate modes and estimate their spread. Adapting a tool like MULTIFAN (Fournier et al. 1990), which analyses multiple L_F histograms, could overcome this weakness.

The final method is discriminant analysis, or the MLA rule (see above), which Fletcher and Blight (1996) applied to W_O–L_F data for pilchards. This technique has the merit of being easily applied to multiple predictors and not requiring linearity between predictors and predictand. However, it is important to specify which of several variants of discriminant analysis is used and what "prior" assumptions are made about the p_A. The simplest, and most common, variant, usually called linear discriminant analysis, assumes normality and homoscedasticity. Quadratic discriminant analysis drops the latter assumption, and nonparametric discriminant analysis drops the former. Two common "prior" assumptions are the uniform prior ($p_A = 1/n$) and priors equal to the proportions in the calibration sample. In the present context, the best results will be obtained if the calibration sample is a simple random sample from the population and the latter choice of prior is used. Otherwise, calibration bias will occur. Given appropriate assumptions, discriminant analysis should avoid centric, smoothing, and heteroscedastic bias. However, if the objective is to estimate proportions at age (rather than assign ages), it will be subject

to discriminant bias (see above). When there is only one predictor, this may be overcome by the use of the UPA rule. An alternative approach is to apply a bias correction to the MLA rule estimate of p_A. We call this the confusion matrix (CM) estimator of p_A because it uses the so-called CM, whose ijth term is the estimated probability that a fish in the ith age class will be assigned to the jth age class by the MLA rule (see section 4.3 of McLachlan and Basford (1988) for more about this estimator).

Two general points can be made about all of the methods reviewed in this section. First, all methods assigned ages to individual fish. This is intrinsic to the regression and discriminant analysis methods but was not necessary with modal analysis. Given that the aim in many studies is to estimate proportions at age, it would seem sensible to consider methods that do this directly rather than via assigned ages. Second, in those methods that use a calibration sample, inference is a two-stage procedure: devise a rule from the calibration sample and then apply it to the production sample. This means that these methods cannot use any information from the production sample in formulating their rule. In the next section, we propose a new method that directly estimates proportions at age and that does this in one step, combining information from both samples.

A new method of inferring age: mixture analysis

We return to our mixture model but describe it first in a more general form. We assume that a vector $\mathbf{X}$ is associated with every fish in a population. Vector $\mathbf{X}$ may contain any otolith or body measurements, such as W_O, L_O, or L_F, or any transformations (e.g., $\log(W_O)$) or functions (e.g., W_O/L_O) of these measurements. For fish of a given age, A, the distribution of $\mathbf{X}$ is described by the density function $g(\mathbf{X}; \boldsymbol{\theta}_A)$ for some unknown vector of parameters, $\boldsymbol{\theta}_A$. The proportion of fish of age A is p_A. As above, we have two random samples from the population: the calibration sample, containing measurements and age for each of n_c fish, ($\mathbf{X}_i$, A_i, $i = 1,..., n_c$), and the production sample, containing just measurements for each of n_p fish, ($\mathbf{X}_j$, $j = 1,..., n_p$). To start, we will assume that both samples are simple random samples, but we will later discuss ways in which this assumption could be modified.

Given these assumptions, we can estimate the p_A (and the $\boldsymbol{\theta}_A$) by maximum likelihood (Stuart and Ord 1991). The log-likelihood of the parameters (p_A,$\boldsymbol{\theta}_A$) given the observations is

$$\lambda = \lambda_c + \lambda_p$$
$$= \Sigma_i \log[p_{A_i} g(\mathbf{X}_i; \theta_{A_i})] + \Sigma_j \log[\Sigma_A p_A g(\mathbf{X}_j; \boldsymbol{\theta}_A)]$$

so estimation is simply a matter of searching for the values of (p_A,$\boldsymbol{\theta}_A$) that maximize λ (the terms λ_c and λ_p are the log-likelihood components associated with the calibration and production samples, respectively). If we also wish to assign ages to individual fish in the production sample, we can do this by assigning the most probable age, just as is done in discriminant analysis. That is, we assign the jth fish to the age A for which $p_A g(\mathbf{X}_j;\boldsymbol{\theta}_A)$ is largest. However, we should not then estimate proportions at age using these assigned ages because, as with the MLA rule, these estimates will be biased.

In the example that we have used above, $\mathbf{X}$ has just one member, W_O, $\boldsymbol{\theta}_A$ is the pair (μ_A,σ_A), and $g(\mathbf{X};\boldsymbol{\theta}_A)$ is the normal density function:

$$g(\mathbf{X};\boldsymbol{\theta}_A) = g(W_O; \mu_A, \sigma_A) = \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left[-\frac{(W_O - \mu_A)^2}{2\sigma_A^2}\right]$$

In its most general form, mixture analysis can present some technical problems (McLachlan and Basford 1988). The likelihood function may be unbounded or have multiple maxima (so that maximum-likelihood estimation is not possible), and there can be difficulty in deciding how many groups there are in the mixture. These problems are avoided in the present application as long as the calibration sample is large enough to contain all of the age groups present in the production sample. When there are few enough age classes that the calibration sample can be expected to include multiple observations for each age class, the number of groups in the mixture will sensibly be set to the number of age classes in this sample. When there is a large number of ages (e.g., Boehlert 1985), not all may be observed in the calibration sample, so a mixture with components spanning the range of ages in this sample might be reasonable, and applying some constraints on the parameters (see below) should avoid estimation difficulties.

When there is no calibration sample and only one normally distributed measurement, the mixture analysis method described here is precisely that used in the program MIX (MacDonald and Green 1988). That method was substantially and elegantly extended by MULTIFAN (Fournier et al. 1990), which made the likelihood more robust and enhanced estimation by allowing the simultaneous analysis of several simple random samples collected at different times (so that the way that length modes shifted over time could be used in the estimation). We make no further comment for the situation where there is no calibration sample except to remind readers that we might do better if our predictor $\mathbf{X}$ were multivariate. In particular, as observed by Brander (1974) (see quote above), the pair (L_F,W_O) may be a better predictor of age than either W_O or L_F alone.

There are a number of reasons to recommend the mixture analysis approach. First, if the aim is to estimate proportions at age, this method will avoid all the asymptotic biases mentioned above because maximum-likelihood methods are known to be asymptotically unbiased (Stuart and Ord 1991). Second, whether we are estimating p_A or assigning ages, this method seems likely to make better use of the information in the production sample. For example, when $\mathbf{X}$ is multivariate normal, both this method and discriminant analysis obtain estimates of the $\boldsymbol{\theta}_A$, but the mixture estimate ought to be superior because it uses information from both samples, whereas the discriminant analysis estimate uses only the calibration sample. A better estimate of the $\boldsymbol{\theta}_A$ should lead to a better estimate of the p_A.

The mixture analysis method is also flexible. If the parameter vector $\boldsymbol{\theta}_A$ has length m, then we can estimate a full set of $n(m + 1)$ parameters or, if the data seem to warrant it, we can reduce the number of parameters by using appropriate constraints. For example, for the one-dimensional normal

example that we used in Fig. 4, we have the constraint $\mu_A = a + bA$, which reduces the number of parameters to be estimated from $3n$ to $2n + 2$, and the additional constraint $\sigma_A = \sigma$ reduces it further to $n + 3$. Such constraints are particularly recommended when $\mathbf{X}$ is multivariate. If, for example, $\mathbf{X} = (W_O, L_O, L_F)$ and this is multivariate normal, there would, in the absence of any constraints, be $7n$ parameters to estimate (3 SDs, three correlations, and p_A for each age class). Linear constraints on the standard deviations and (or) correlations would often be sensible. This is particularly true when the number of calibration observations per age class is small. Likelihood ratio tests can be used to determine whether these constraints are justified. There is no need to assume linearity, as was required in the regression methods. If, for instance, the relationship between mean W_O and age is quadratic, we simply change our constraint to $\mu_A = a + bA + cA^2$. We may choose to transform some of our measurements (e.g., use $\log(W_O)$ rather than W_O), but this need not be done to obtain linearity. Other reasons for such transformations are to normalize a distribution or to obtain homoscedasticity (so we can assume that $\sigma_A = \sigma$).

Extensions to the mixture analysis method

The log-likelihood function given above assumes that the calibration sample is a simple random sample of the population. This is not always possible or convenient. There are two alternative sample structures for which it is easy to modify the λ_c term in the log-likelihood function. If the calibration sample is random at age (i.e., all fish of the same age have the same probability of being included in the sample), then the form $\lambda_c = \Sigma_i \log[g(\mathbf{X}_i; \boldsymbol{\theta}_{A_i})]$ should be used. This form could also be used when the calibration sample is a simple random sample but is selected from a different population (different, that is, from the population from which the production sample is selected and about which we wish to make age inferences). However, we must be able to assume that the distribution of $\mathbf{X}$ at each age is the same in both populations. Suppose, for instance, that we know that the distributions of W_O at each age do not vary significantly from year to year. Then, if we are using only W_O as a predictor, we could use a calibration sample from an earlier year. This sample would then be representative of the current population (from which we draw the production sample) in terms of W_O but not in terms of p_A, so the random-at-age form of λ_c could be used. Some authors prefer to use a length-stratified calibration sample (e.g., Araya et al. (2001) selected 10 fish from each 1-cm length class). This can be accommodated as long as the sample is random at length (i.e., all fish of the same length have the same probability of being included in the sample) and fish length is included in the vector of measurements $\mathbf{X}$. If we take the example $\mathbf{X} = (L_F, W_O)$, then λ_c should be given by

$$\lambda_c = \Sigma_i \log[p_{Ai}\, g(L_{Fi}, W_{Oi}; \boldsymbol{\theta}_{Ai}) / \Sigma_A p_A h(L_{Fi}; \boldsymbol{\theta}_A)]$$

where $h(L_F; \boldsymbol{\theta}_A)$ is the density function of L_F for fish of age A, which is a marginal density of $g(L_F, W_O; \boldsymbol{\theta}_A)$.

We can also modify λ_c to deal with ageing error in the calibration sample as long as we can provide a misclassification matrix $\mathbf{M}$ to characterize this error. The element $\mathbf{M}_{A,A'}$ of this matrix denotes the probability that a fish of true age A is given age A' (so the rows of this matrix must sum to 1). This matrix can be constructed using replicate age estimates. With ageing error, we have

$$\lambda_c = \Sigma_i \log[\Sigma_A p_A \mathbf{M}_{A,A_i}\, g(\mathbf{X}_i; \boldsymbol{\theta}_A)]$$

Estimation for the mixture-analysis model has been described in terms of maximum likelihood. However, it would be relatively straightforward to convert to Bayesian estimation (Berger 1980). This requires the user to supply prior distributions for every model parameter. Also, estimation is more complex because what is required is a joint posterior distribution for all parameters rather than just a single value. However, there are several widely available software packages such as BUGS (http://www.mrc-bsu.cam.ac.uk/bugs/) and AD Model Builder (http://otter-rsch.com/admodel.htm) that facilitate Bayesian estimation. The specification of prior distributions is often difficult. However, it might be simpler if the model were to be used year after year with the same fish stock. In this situation, it would be sensible to use the posterior distributions for the $\boldsymbol{\theta}_A$ from one year as the prior distributions in the following year (but of course, it would not be sensible to do this for p_A).

It is common to estimate proportions at age in a fishery catch using an age–length key (Kimura 1977). A small sample of age and length data is used to convert estimated proportions at length in a catch, say, to proportions at age. An obvious question to ask is whether otolith measurements could be used in place of annulus counts in this setting. In particular, can the mixture analysis model be extended for use here? The answer is yes, and this extension will be described in a separate paper.

Mixture analysis can also be adapted to the situation where the prime reason for inferring age is to estimate growth parameters rather than proportions at age. All that is required is to include L_F in the vector of measurements $\mathbf{X}$ and impose the constraint that mean lengths at age follow a growth curve. The parameters of that growth curve are then estimated directly in the mixture analysis method.

Evaluating performance

In this section, we discuss approaches for evaluating the performance of methods of inferring age from otolith measurements. We will consider both types of inference (assigning ages and estimating proportions at age) and begin by reviewing the literature.

The most common statistic that has been used to measure how well ages have been assigned to individual fish is the probability of correct age estimation, P_{correct}, which is sometimes expressed as the percentage correctly classified (Pawson 1990; Worthington et al. 1995*a*; Pilling et al. 2003). When a calibration sample is used, this compares "true" ages (usually from annulus counts) with assigned ages; in a modal analysis, it could be calculated using the estimated means and standard deviations for each mode and normal distribution theory. An alternative statistic, useful when ageing error is proportional to true age, is the coefficient of variation of the ageing error (Worthington et al. 1995*a*; Cardinale et al. 2000; Pilling et al. 2003). In almost all of the studies that we reviewed, the statistics presented gave an optimistic view of estimation performance because they were calculated using

the same sample that was used to calibrate the cutting rule. For an unbiased estimate of either of these statistics, the simplest approach is to use two calibration samples. The cutting rule is calculated using the first sample, which is then applied to the second sample, and the statistic is calculated from the difference between "true" and assigned ages in the second sample. A better, but more complicated, alternative is to use cross-validation with a single calibration sample (Finn et al. 1997).

There are two other statistics that provide a rough guide to how well ages are estimated. The first is the separation index, S, whose approximate relationship to P_{correct} has been discussed above. The separation index is useful in that it provides an objective measure of the amount of overlap between adjacent modes. For example, a visual assessment of histograms by Fletcher (1991) suggested "little overlap of otolith weight between age classes", but separation indices calculated from the means and standard deviations in his table 2 range from 1.7 to 5.7, with median 2.7. The second statistic, relevant only for the regression method, is the residual standard deviation, s_{res}. The bigger s_{res} is, the poorer the estimation. To show how this statistic relates to P_{correct}, we take an example from Boehlert (1985). In his table 3, he presents a regression for estimating whole-otolith age from W_{O} and w_{O} for females and calculates $s_{\text{res}} = 4.15$ years. This means that if we take a group of female fish, all of which have exactly the same values of W_{O} and w_{O}, the expected standard deviation of their ages will be 4.15 years. If the regression estimate of the age of these fish is unbiased, they will all be estimated to be of age equal to their mean age, A_{mn}. Only those fish whose true age differs from A_{mn} by less than 0.5 year will be assigned the correct age so $P_{\text{correct}} \approx F(0.5/4.15) - F(-0.5/4.15) = 0.096$ (that is, only 9.6% of fish are aged correctly, assuming normality and ignoring bias). If there is only a single regression predictor, we can get some idea of estimation performance from a data plot. For example, from fig. 4 of Luckhurst et al. (2000), we can see that a fish with $W_{\text{O}} = 550$ mg could be any age between about 10 and 17 years, so we cannot expect a high P_{correct} for this species.

Many authors appear to believe that, in a regression context, high R^2 values imply good estimation performance, but it is easy to show that this is not so. If we consider the simplest case of the mixture model used in Fig. 4 (i.e., assuming homoscedasticity and no variation in proportions at age), the equation for R^2 given in Appendix A simplifies to $R^2 = 1/(1 + 12/(S^2(n^2 - 1)))$, where $S = 1/\sigma$ is the separation index between adjacent age classes and n is the number of age classes. So, if $S = 3$, R^2 can be as low as 0.69 for $n = 2$ but increase to 0.997 for $n = 20$. But for this simple model, the probability of correct age estimation $P_{\text{correct}} = 0.87$, regardless of the number of age classes n (see above). Thus, R^2 is clearly not closely related to how well we estimate age for this model. Another indication of how little R^2 means is given by the regression results of Boehlert (1985). Amongst the 12 regression models he presents in his tables 3–5 and 10–12, there are three for which $s_{\text{res}} = 2.8$ (which implies $P_{\text{correct}} \approx 0.14$), but the R^2 values for these regressions vary between 0.70 and 0.92. Where R^2 can be useful is in comparing predictors for the same predictand (which will be age or some transform of age).

Worthington et al. (1995*a*) recognized that R^2 was not an appropriate measure of estimation performance but suggested that "the ratio of the mean squares due to the regression and the residual...is a more appropriate index of the potential of otolith weight to estimate age". As they pointed out, this ratio is just the F statistic used to test the null hypothesis that the regression slope is zero (Draper and Smith 1981). This is not a good measure of estimation performance because its value can be made as large as we like (i.e., the associated P value can be made as small as we like) simply by taking a large enough sample from the population. All it tells us is how confident we can be that the correlation between otolith weight (say) and age is nonzero.

A common way of evaluating estimation performance has been with what we might call comparative methods where the comparison made is with the conventional ageing method (annulus counts). The logic behind these methods seems to be that if the otolith measurement method can be shown to be comparable with or better than the annulus count method, then it is preferable because it is cheaper. We first describe four different comparative methods and then return to the logic behind them. The first two methods compare differences between repeated annulus counts (either within-reader, between-reader, or between-agency differences) with those between annulus counts and ages inferred from otolith measurements. Thus, they compare within-annulus-method error with the between-methods error. The first method, used by Boehlert (1985), compares mean differences, which are a measure of bias (although in describing this comparison, Boehlert referred to variability rather than bias). Other studies have compared variability, measured either by P_{correct} (Pawson 1990; Fletcher and Blight 1996) or an error coefficient of variation (Worthington et al. 1995*a*). The other two methods compare estimates, rather than errors, and do so with statistical tests. Boehlert did this for estimates of individual ages using a paired t test; other authors have used a Kolmogorov–Smirnov test to compare estimated proportions at age (Worthington et al. 1995*a*; Cardinale et al. 2000; Pilling et al. 2003).

It seems to us that all of these methods have missed the point. To start with, the focus has mostly been on the assignment of ages to individual fish. As we have said above, the reason for age estimation is almost always to estimate population parameters, usually proportions at age. If people are going to go to the trouble and expense of devising an alternative method of inferring age, they are likely to be doing so for use in routine (probably annual) estimation of population parameters. Thus, when they are evaluating these alternative methods, their focus ought to be on determining which method will produce the best parameter estimates. Once we focus on parameter estimates, cost and sample sizes become important. Many authors have noted that otolith measurement is much cheaper than annulus counting. For example, Boehlert (1985) gave processing rates of 6–8/h for annulus counts using sectioned otoliths and 40/h for otolith measurements. Thus, for a given cost, we have a choice between using annulus counts from a medium-sized sample or otolith measurements from a large production sample together with both annulus counts and otolith measurements from a small calibration sample. The key question is, which will give better parameter estimates? In other words, to decide be-

tween otolith measurement and annulus count methods, we need a cost–benefit analysis.

The study of Worthington et al. (1995*b*) appears to be the only one that has attempted to evaluate an otolith measurement method of inferring age using a cost–benefit analysis. They simulated data based on two populations of *Pomacentrus moluccensis* and showed, among other things, that an annulus count sample of 200 fish would provide less accurate estimates of proportions at age than an otolith-weight sample of 500 (assuming a coefficient of variation of ageing error of 5% for annulus counts and 10% for an otolith weight method). Because the processing time for annulus counts is 5–10 times that for otolith weights, they concluded that the otolith weight method is a more cost-effective, and thus preferable, tool for estimating proportions at age for this species. Despite the overall advantages of this approach, there are some additional considerations that would have been useful. First, it is not clear whether all costs have been considered. The costs are defined as relative processing times and do not appear to include sample collection costs (it presumably costs more to collect 500 otoliths than to collect only 200). Also, the otolith weight method requires a calibration sample, whose cost does not appear to have been allowed for, and the error coefficient of variation of this method must depend on the sizes of both the calibration and production samples. Second, the modelling of errors for the otolith weight method was perhaps too simple. Estimation errors are a mixture of bias and imprecision, and different estimation methods will produce a different mixture. The authors' approach assumes that the two methods have the same mixture. This might be important because the mixture of bias and imprecision associated with an estimator determines how quickly performance improves as sample sizes increase. These details aside, we believe that a cost–benefit analysis similar to that adopted by Worthington et al. (1995*b*) is the only valid way to decide whether otolith measurement methods are worth using instead of the traditional annulus counts.

A simple simulation experiment

We present a simple experiment, with simulated data, that illustrates both the application of the mixture analysis model and the cost–benefit analyses recommended in the preceding section. The simulated data came from a mixture model with one predictor, W_O, and just two ages (i.e., $n = 2$). We assumed that W_O had mean values $\mu_1 = 10$ and $\mu_2 = 20$ and standard deviations $\sigma_1 = 5$ and $\sigma_2 = 8$ and that the proportions at age were $p_1 = 0.3$ and $p_2 = 0.7$. Five hundred data sets were simulated from this population; in each, there was a calibration sample of size 50 and a production sample of size 250 (both simple random samples). For each data set, we obtained seven estimates of p_1. The first five used methods reviewed above (ordinary regression, linear calibration, GM regression, MLA (quadratic discriminant analysis), and the UPA rule). Each of these methods assigns an age to each fish in the production sample, and our estimate of p_1 was the proportion of 1-year-olds in the combined calibration and production samples. (Note that it is sensible to combine the calibration and production samples to estimate p_1 because the former sample is simple random; had it been only random at age or random at length, we would estimate p_1 from the proportion of assigned 1-year-olds in the production sample alone.) Our sixth estimate used the CM method, which assigns ages using the MLA rule but then adjusts the estimate of p_1 for bias. The final estimate came from the mixture analysis model (this is a direct estimate in that it does not involve assigning ages to individual fish).

For each method, we calculated three performance measures. The most important of these, the root mean square error (RMSE), is defined as $[(1/500)\Sigma_k(\hat{p}_{1,k} - p_1)^2]^{0.5}$, where the summation is over the 500 estimates of $p_1(\hat{p}_{1,1}, \hat{p}_{1,2}, \ldots, \hat{p}_{1,500})$. This measures how close the estimated p_1 is to the true value, on average, so the smaller RMSE, the better. The other two measures are bias $= (1/500)\Sigma_k(\hat{p}_{1,k} - p_1)$ and precision $= [(1/500)\Sigma_k(\hat{p}_{1,k} - \bar{p}_1)^2]^{0.5}$. The three measures are related by the equation $\text{RMSE}^2 = \text{bias}^2 + \text{precision}^2$. The last two measures are useful to help us distinguish between estimators that are poor (i.e., have high RMSE) because they are biased but precise from those that are unbiased but imprecise. Approximate 95% confidence intervals were calculated for all three measures (see Appendix B).

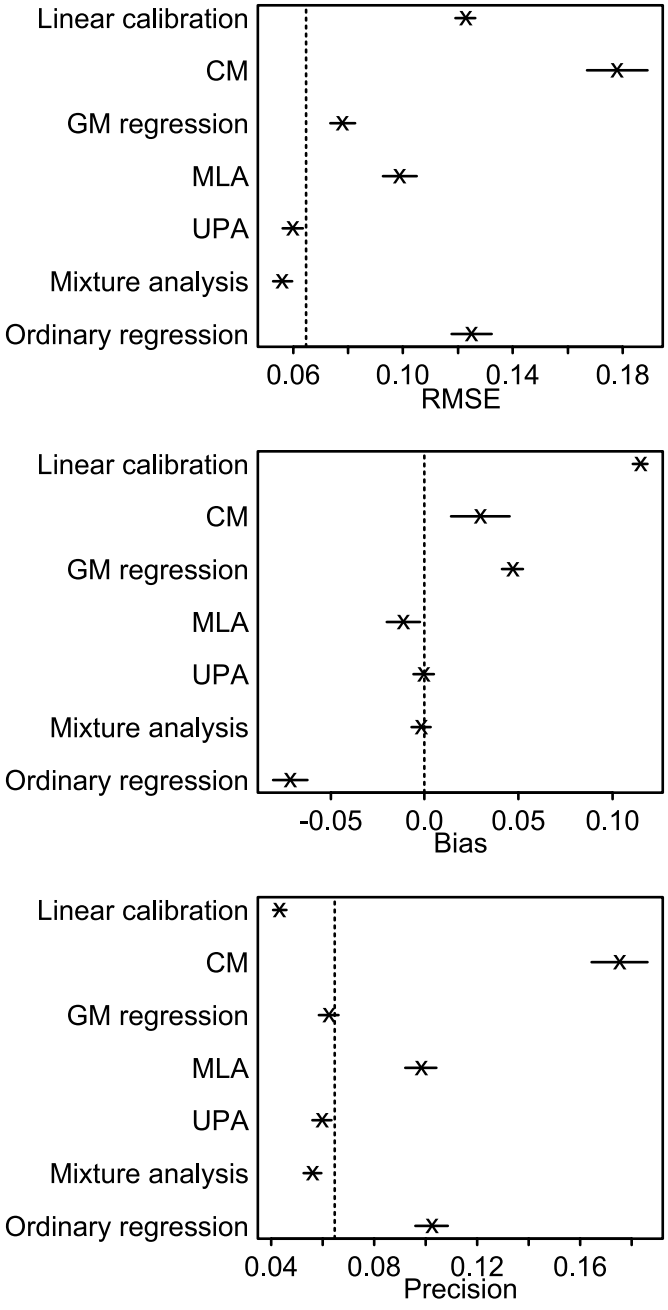
A useful benchmark for comparison is the estimate of p_1 we would get if we simply ignored the W_O data and calculated the proportion of 1-year-olds in the calibration sample. From binomial theory, we know this estimator has bias = 0 and RMSE = precision $= [p_1(1 - p_1)/50]^{0.5} = 0.0648$.

The results from this experiment (Fig. 6) show that, for this simple example, only mixture analysis and UPA give better results than we would get if we ignored the W_O data, and all methods but these two show significant bias (i.e., the 95% confidence intervals for the estimated bias do not include zero). Linear calibration is the most precise of all methods considered but also shows the greatest bias. The RMSE confidence intervals for UPA and mixture analysis overlap, so we cannot, on the basis of this statistic, say which is better. However, because both methods were applied to the same data sets, we can make a paired comparison, which is more powerful. We counted how many times out of 500 the mixture analysis estimate was closer to the true value. Under the null hypothesis of no difference between the two methods, this count should have a binomial distribution with parameters 500 and 0.5. The observed count was 291, so we can reject the null hypothesis and declare the mixture analysis method superior for this simple example (two-sided test, $P = 0.0002$).

Our judgement of these methods would be quite different if our main aim had been to assign ages to individual fish (rather than estimate proportions at age) and we chose P_{correct} as our performance measure. By this measure, UPA is the best method (for our simple example), although there is not a great deal to choose from between the different methods (Fig. 7).

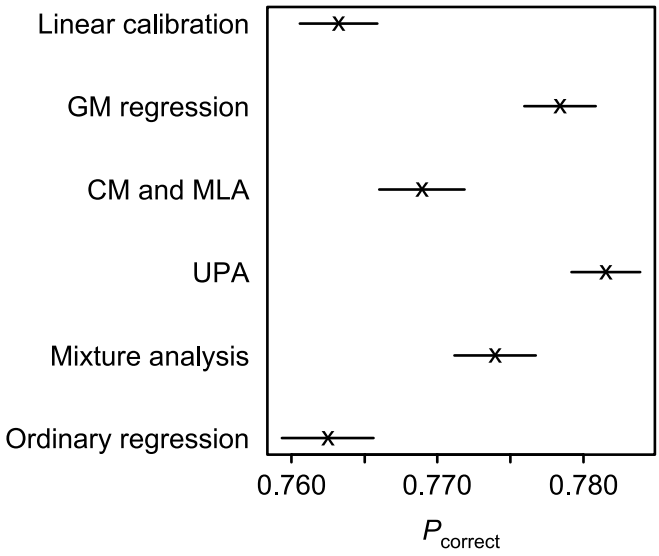
Now to a cost–benefit analysis. Suppose the per-otolith costs of collecting, weighing, and ageing fish are C_{collect}, C_{weigh}, and C_{age}, respectively (these costs could reflect not only the time taken for each of these operations but also the different levels of skill required, which affects salaries). For the sample sizes used in our experiment, the total cost is given by $C_1 = 50(C_{\text{collect}} + C_{\text{weigh}} + C_{\text{age}}) + 250(C_{\text{collect}} + C_{\text{weigh}})$. The RMSE obtained using the mixture analysis method was 0.056. To have achieved the same RMSE without W_O data would have required a sample of 67 (= 0.3 ×

Fig. 6. Estimated performance measures (RMSE, bias, and precision) for seven methods of inferring proportions at age using annulus count and otolith weight (W_O) data from a simple two-age population. Horizontal lines are approximate 95% confidence intervals for each performance measure; vertical broken lines are the performance measures that would be obtained if the W_O data were ignored. CM, confusion matrix; GM, geometric mean; MLA, most likely age; UPA, unbiased proportions at age; RMSE, root mean square error.



$0.7/0.056^2$) otoliths, which would cost $C_2 = 67(C_{collect} + C_{age})$. It would be worthwhile using the mixture analysis method only if $C_1 < C_2$. For the UPA method, the analogous calculations lead to an equivalent sample size of 58 (= $0.3 \times 0.7/0.060^2$) otoliths, so this method would be worthwhile only if $C_1 < 58(C_{collect} + C_{age})$.

Fig. 7. Estimates of $P_{correct}$ (proportion of fish assigned the correct age , ×) with 95% confidence intervals (i.e., ±2 SE shown by horizontal lines) from the same simulation experiment as for Fig. 5. CM, confusion matrix; GM, geometric mean; MLA, most likely age; UPA, unbiased proportions at age.



We stress that we intend this example to be illustrative rather than definitive. It would be quite wrong to draw general conclusions about the relative merits of the different age inference methods on the basis of such a limited experiment. Nevertheless, our results do give some support to the view, expressed above, that the mixture analysis method should be superior to the other methods because it is (at least asymptotically) unbiased and makes better use of the production sample than do all of the other methods. The equivalent sample sizes calculated for our cost–benefit analysis might seem surprisingly small. They imply, for example, that our production sample of 250 otolith weights contains only as much information as 17 (= 67 – 50) additional annulus counts (assuming that we use the mixture analysis method). The reason for this is that the separation index in our example is low (2.1), which implies that the W_O data cannot contain much information about age. When we repeated the above experiment with a higher separation index (shifting μ_2 to 25 so $S = 3.2$), we obtained a lower RMSE (0.046), which meant that the production sample contained as much information as 49 additional annulus counts. So, the higher the separation index, the more likely it will be that it is worthwhile to use W_O data in inferring age.

Discussion

Given the large number of otoliths that are read every year, usually with a high per-otolith cost, there is potential for great cost savings if otolith measurements can be used to infer age. How much money can be saved, if any, will vary from fish stock to fish stock depending on the specific form of otolith (and body) growth. For each stock, there are four questions that we must address before we can decide whether there are savings to be made. We will discuss each question separately, referring both to the literature that we have reviewed and to the method suggested above.

The first question is "what do we want age data for?" We believe that the most common answers to this question are "to estimate proportions at age" and "to estimate growth parameters", but many published studies seem to have assumed the answer "to assign ages to individual fish". The question is important because it affects the way we compare our otolith measurement method with the annulus count method. For example, we have shown in our example that the use of P_{correct} as a performance measure will produce misleading results if our aim is to estimate proportions at age.

The answer to our second question, "which are the best measurements to use as predictors?", will clearly vary from stock to stock. The literature shows that otolith weight (W_O) is a prime candidate, but there will often be gains to be made from using multiple predictors. The figures in Table 1 show that we should not ignore fish length (L_F) as a potential predictor (to be used in conjunction with other predictors, such as W_O). Some methods of inference (e.g., linear calibration and GM regression) are limited by allowing only one predictor.

We have provided some information towards an answer to our third question, "which is the best method of inference?" When the aim is to estimate proportions at age, we have shown that all published methods are subject to asymptotic biases of various sorts and that there are some grounds for preferring the mixture analysis method. Nevertheless, only a thorough investigation will establish which is the best method in a particular situation. Several methods have been proposed for inferring age from otolith measurements, but there have been no studies comparing alternative methods for a specific fish stock.

Our fourth question, "how should we evaluate these methods of inference?", is the only one that seems to us to have just one answer. Only a cost–benefit analysis will show whether a proposed method is superior to counting annuli. To be acceptable, the proposed method must provide "better" estimates of the desired quantities (be they proportions at age, growth parameters, or simply ages) for the same cost as the annulus count method (or equally good estimates for a lower cost). Exactly how we define "better" will depend on what we are estimating. For our simple example, RMSE in the estimate of p_1 seemed to be an appropriate measure. This could be extended for the case when there are more than two age classes by making the summation in the formula for RMSE to be over age classes as well as data sets. When our focus is on assigning ages to individuals, we might use measures such as P_{correct} or average percentage error. A simulation experiment, like that described above (but much more comprehensive), is an ideal way to carry out the cost–benefit analysis. While artificial data provide a useful way to investigate general properties of estimators, we can make useful conclusions about a specific fish stock only by basing our simulations on data from that stock. For instance, we might simulate more realistic data sets by bootstrapping (i.e., selecting at random, with replacement) the data in Fig. 1. With such data, we can easily simulate the effect of changes in proportions at age by altering the probability of selecting fish of different ages. Ideally, these simulated data should include error in annulus counts (as inferred from replicate annulus counts), where this is significant. When the reason for inferring age is to provide inputs to a stock assessment, it may be possible to extend the simulation experiment to include running the stock assessment model with inferred ages from the simulated data. In this case, our performance measures for evaluating the alternative methods of inference would measure how well we estimated key assessment outputs (e.g., current exploitation rates).

Many authors have followed Boehlert (1985) in describing age inference methods based on otolith measurements as being "objective". This seems to be inaccurate. Most otolith measurements are rightly labelled objective, particularly in contrast with the subjectivity of annulus counts. However, all methods that use a calibration sample depend on both objective measurements and subjective counts and so cannot be considered as objective. As Worthington et al. (1995*b*) noted, these methods cannot be expected to produce better estimates of age than the annulus counts that are used to calibrate them; even the age readings of skilled age readers are not completely reproducible, underlining the presence of random ageing error (Campana 2001). It may be reasonable to use the word "objective" when no calibration sample is used, but even here, there is often an element of subjectivity (e.g., in deciding how many modes, or age classes, to fit).

Several authors have mentioned the need for regular recalibration (Boehlert 1985; Worthington et al. 1995*b*; Pilling et al. 2003). That is, we should be cautious about using the same calibration sample over and over again. There are two ways in which a calibration sample might be inappropriate for use with a particular population. First, its proportions at age may not be representative of that population. For some methods, this will cause what we have called calibration bias, but other methods (e.g., mixture analysis) can allow for this lack of representativeness and thus avoid this bias. Second, the relationship between age and the chosen predictors may be different from that in the target population. For example, the mean otolith weight of fish for a given age may differ between the calibration sample and the target population. This sort of variation has been demonstrated in the spatial domain (Anderson et al. 1992; Worthington et al. 1995*a*; Pilling et al. 2003), so we should be particularly cautious about using a calibration sample in an area other than that in which it was collected. Whether temporal variation is as much of a problem has not been as widely investigated (at least for otolith measurements), but it would be prudent to recalibrate each year until this has been shown to be unnecessary. Pilling et al. (2003) found no significant difference in the otolith weight – age relationships in two samples of *Lethrinus mahensa* collected from the same location 2 years apart. There are many examples of annual variations in mean length at age, so annual calibration samples are likely to be needed when body length is one of the predictors. The need for recalibration means that we cannot avoid completely the costs involved in maintaining the expertise required for annulus count age estimation. However, the number of people who must have that skill might be reduced, as might the time they spend in counting annuli for each species.

An interesting question is whether there is an advantage to using a length-stratified calibration sample. An argument in favour of this sample structure is that it is easier to achieve, from a logistical point of view, than a simple random sample. This may be why length-stratified samples are so widely used for age–length keys despite the finding by Kimura

(1977) that simple random samples produce better estimates of proportions at age. An argument in favour of simple random calibration samples in the present context is that they provide two types of information — proportions at age and the age–predictor relationship — whereas length-stratified samples provide only the latter type. Also, some methods of inference (e.g., mixture analysis) can be adjusted for length-stratified samples, whereas most cannot.

In conclusion, we believe that there is clear scope to reduce the cost of fisheries research by inferring age from otolith measurements. However, it is important that we are aware of what we are trying to achieve (e.g., assigning ages to individual fish or estimating proportions at age) and that we carry out cost–benefit analyses for each fish stock to find the best method of inference and to ensure that there are real cost-savings to be made.

Acknowledgments

We are grateful to Peer Doering-Arjes for the use of the data in Fig. 1 and to Dave Gilbert, Graham Pilling, and an anonymous referee for reviewing earlier versions of this paper.

References

Anderson, J.R., Morison, A.K., and Ray, D.J. 1992. Age and growth of Murray cod, *Maccullochella peelii* (Perciformes: Percichthyidae), in the lower Murray–Darling Basin, Australia, from thin-sectioned otoliths. Aust. J. Mar. Freshw. Res. **43**: 983–1013.

Araya, M., Cubillos, L.A., Guzmán, M., Peñailillo, J., and Sepúlveda, A. 2001. Evidence of a relationship between age and otolith weight in the Chilean jack mackerel, *Trachurus symmetricus murphyi* (Nichols). Fish. Res. **51**: 17–26.

Berger, J.O. 1980. Statistical decision theory and Bayesian analysis. 2nd ed. Springer-Verlag, New York.

Blacker, R.W. 1974. Recent advances in otolith studies. *In* Sea fisheries research. *Edited by* Harden Jones, FR. John Wiley & Sons, New York. pp. 67–90.

Boehlert, G.W. 1985. Using objective criteria and multiple regression models for age determination in fishes. Fish. Bull. **83**: 103–117.

Brander, K. 1974. The effects of age-reading errors on the statistical reliability of marine fishery modelling. *In* The ageing of fish. *Edited by* T.B. Bagenal. Unwin Brothers Limited, Old Woking, Surrey, UK. pp. 181–191.

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.

Campana, S.E., and Thorrold, S.R. 2001. Otoliths, increments, and elements: keys to a comprehensive understanding of fish populations? Can. J. Fish. Aquat. Sci. **58**: 30–38.

Cardinale, M., Arrhenius, F., and Johnsson, B. 2000. Potential use of otolith weight for the determination of age-structure of Baltic cod (*Gadus morhua*) and plaice (*Pleuronectes platessa*). Fish. Res. **45**: 239–252.

Draper, N.R., and Smith, H. 1981. Applied regression analysis. 2nd ed. John Wiley & Sons, New York.

Ferreira, B.P., and Russ, G.R. 1994. Age validation and estimation of growth rate of the coral trout, *Plectropomus leopardus* (Lacepede 1802), from Lizard Island, Northern Great Barrier Reef. Fish. Bull. **92**: 46–57.

Finn, J.E., Burger, C.V., and Holland-Bartels, L. 1997. Discrimination among populations of sockeye salmon fry with Fourier analysis of otolith banding patterns formed during incubation. Trans. Am. Fish. Soc. **126**: 559–578.

Fletcher, W.J. 1991. A test of the relationship between otolith weight and age for the pilchard *Sardinops neopilchardus*. Can. J. Fish. Aquat. Sci. **48**: 35–38.

Fletcher, W.J. 1995. Application of the otolith weight–age relationship for the pilchard, *Sardinops sagax neopilchardus*. Can. J. Fish. Aquat. Sci. **52**: 657–664.

Fletcher, W.J., and Blight, S.J. 1996. Validity of using translucent zones of otoliths to age the pilchard *Sardinops sagax neopilchardus* from Albany, western Australia. Mar. Freshw. Res. **47**: 617–624.

Fossen, I., Albert, O.T., and Nilssen, E.M. 2003. Improving the precision of ageing assessments for long rough dab by using digitized pictures and otolith measurements. Fish. Res. **60**: 53–64.

Fournier, D.A., Sibert, J.R., Majkowski, J., and Hampton, J. 1990. MULTIFAN a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*). Can. J. Fish. Aquat. Sci. **47**: 301–317.

Johnson, N.L., and Kotz, S. 1970. Continuous univariate distributions — 2. John Wiley & Sons, New York.

Kimura, D.K. 1977. Statistical assessment of the age–length key. J. Fish. Res. Board Can. **34**: 317–324.

Labropoulou, M., and Papaconstantinou, C. 2000. Comparison of otolith growth and somatic growth in two macrourid fishes. Fish. Res. **46**: 177–188.

Luckhurst, B.E., Dean, J.M., and Reichert, M. 2000. Age, growth and reproduction of the lane snapper *Lutjanus synagris* (Pisces: Lutjanidae) at Bermuda. Mar. Ecol. Prog. Ser. **203**: 255–261.

MacDonald, P.D.M., and Green, P.E.J. 1988. User's guide to program MIX: an interactive program for fitting mixtures of distributions. Icthus Data Systems, Hamilton, Ont.

McLachlan, G.J., and Basford, K.E. 1988. Mixture models: inference and applications to clustering. Marcel Dekker, New York.

Pauly, D. 1987. Application of information on age and growth of fish to fishery management. *In* Age and growth of fish. *Edited by* R.C. Summerfelt and G.E. Hall. Iowa State University Press, Ames, Iowa. pp. 495–506.

Pawson, M.G. 1990. Using otolith weight to age fish. J. Fish Biol. **36**: 521–531.

Pilling, G.M., Grandcourt, E.M., and Kirkwood, G.P. 2003. The utility of otolith weight as a predictor of age in the emperor *Lethrinus mahsena* and other tropical fish species. Fish. Res. **60**: 493–506.

Reznick, D., Lindbeck, E., and Bryga, H. 1989. Slower growth results in large otoliths: an experimental test with guppies (*Peocilia reticulata*). Can. J. Fish. Aquat. Sci. **46**: 108–112.

Ricker, W.E. 1973. Linear regressions in fishery research. J. Fish. Res. Board Can. **30**: 409–434.

Secor, D.H., and Dean, J.M. 1989. Somatic growth effects on the otolith – fish size relationship in young pond-reared striped bass, *Morone saxatilis*. Can. J. Fish. Aquat. Sci. **46**: 113–121.

Snedecor, G.W., and Cochran, W.G. 1980. Statistical methods. 7th ed. Iowa State University Press, Ames, Iowa.

Stuart, A., and Ord, J.K. 1991. Kendall's advanced theory of statistics. Vol. 2. Classical inference and relationship. 5th ed. Charles Griffin and Company Limited, London, U.K.

Templeman, W., and Squires, H.J. 1956. Relationship of otolith lengths and weights in the haddock *Melanogrammus aeglefinus* (L.) to the rate of growth of the fish. J. Fish. Res. Board Can. **13**: 467–487.

Worthington, D.G., Doherty, P.J., and Fowler, A.J. 1995*a*. Variation in the relationship between otolith weight and age: implications

for the estimation of age of two tropical damselfish (*Pomacentrus moluccensis* and *P. wardi*). Can. J. Fish. Aquat. Sci. **52**: 233–242.

Worthington, D.G., Fowler, A.J., and Doherty, P.J. 1995*b*. Determining the most efficient method of age determination for estimating the age structure of a fish population. Can. J. Fish. Aquat. Sci. **52**: 2320–2326.

Wright, P.J., Metcalfe, N.B., and Thorpe, J.E. 1990. Otolith and somatic growth rates in Atlantic salmon parr, *Salmo salar* L: evidence against coupling. J. Fish Biol. **36**: 241–249.

Appendix A. Bias in the regression model

In this appendix, we derive the equations used in Fig. 4. The text preceding a description of the results in this figure includes the assumptions on which these equations are based.

If we define $\overline{W}_O$ and $\overline{A}$ to be the mean otolith measurement and age in the calibration sample, then the age, $\hat{A}$, predicted by the regression method is given by $\hat{A} = \overline{A} + \beta(W_O - \overline{W}_O)$, where the formula for β depends on which regression model is used:

$$\beta = \begin{cases} E\{(W_O - \overline{W}_O)(A - \overline{A})\} \big/ E\{(W_O - \overline{W}_O)^2\} & \text{for standard regression} \\ E\{(A - \overline{A})^2\} \big/ E\{(W_O - \overline{W}_O)(A - \overline{A})\} & \text{for linear calibration} \\ \left[E\{(A - \overline{A})^2\} \big/ E\{(W_O - \overline{W}_O)^2\}\right]^{0.5} & \text{for GM regression} \end{cases}$$

where $E\{X\}$ denotes the expectation, or mean, of X.

Because our equations are approximations for a large calibration sample, we can assume that there is no appreciable difference between the actual sample means $\overline{W}_O$ and $\overline{A}$ and their expected values. With this assumption, it is easy to show that $\overline{A} = \Sigma_A p_{c,A} A$, $\overline{W}_O = a + b\overline{A}$, $E\{(A - \overline{A})^2\} = \Sigma_A p_{c,A}(A - \overline{A})^2 = V_A$, $E\{(W_O - \overline{W}_O)^2\} = b^2(V_A + \overline{V}_O)$, and $E\{(W_O - \overline{W}_O)(A - \overline{A})\} = bV_A$, where $\overline{V}_O = \Sigma_A p_{c,A} \sigma_A^2$. We note in passing that the proportion of variance explained by the regression is given by

$$R^2 = \frac{[E\{(W_O - \overline{W}_O)(A - \overline{A})\}]^2}{E\{(W_O - \overline{W}_O)^2\}E\{(A - \overline{A})^2\}} = \frac{1}{1 + \overline{V}_O / V_A}$$

A fish from the production sample will be assigned to age A_0 if $A_0 - 0.5 \le \hat{A} < A_0 + 0.5$, which is the same as $f(A_0 - 0.5) \le W_O < f(A_0 + 0.5)$, where $f(x) = a + b\overline{A} + (x - \overline{A})/\beta$. Therefore, the proportion of fish in this sample that is assigned age A_0 is given by

$$\begin{aligned} p_{p,A_0} &= P[f(A_0 - 0.5) \le W_O < f(A_0 + 0.5)] \\ &= \sum\nolimits_A p_{t,A} P_A [f(A_0 - 0.5) \le W_O < f(A_0 + 0.5)] \\ &= \sum\nolimits_A [g_A(A_0 + 0.5) - g_A(A_0 - 0.5)] \end{aligned}$$

where P_A is the probability for fish of age A, and

$$g_A(x) = p_A F\left[\frac{f(x) - a - bA}{b\sigma_A}\right]$$

$$= p_A F\left[\frac{b(\overline{A} - A) + (x - \overline{A})/\beta}{b\sigma_A}\right]$$

$$= \begin{cases} p_A F\left[\dfrac{x(V_A + \overline{V}_O) - \overline{A}\,\overline{V}_O - AV_A}{\sigma_A V_A}\right] & \text{for standard regression} \\ p_A F\left[\dfrac{x - A}{\sigma_A}\right] & \text{for linear calibration} \\ p_A F\left[\dfrac{(x - \overline{A})(V_A + \overline{V}_O)^{0.5} + (\overline{A} - A)V_A^{0.5}}{\sigma_A V_A^{0.5}}\right] & \text{for GM regression} \end{cases}$$

and F is the cumulative distribution function of the standard normal distribution.

It is of interest to note that $p_{p,A}$ does not depend at all on the parameters a and b for any of the regression models.

Appendix B. Approximate confidence intervals

In this appendix, we describe the method of calculating the approximate 95% confidence intervals in Fig. 6. These were based on the assumption that our estimator of p_1 is approximately normally distributed with mean $\tilde{p}_1$ and standard deviation σ ($\tilde{p}_1$ will differ from the true value p_1 if the estimator is biased). With this assumption, an approximate interval for bias is straightforwardly (bias – 2s, bias + 2s), where s is the estimated standard error of the bias, $s \approx [\Sigma_k(\hat{p}_{1,k} - \bar{p}_1)^2/(N(N-1))]^{0.5}$, $\bar{p}_1$ is the mean of $\hat{p}_{1,k}$, and $N = 500$ is the number of simulated data sets. Further, precision^2 is distributed according to σ^2/N times a χ^2 distribution with $N - 1$ degrees of freedom. Therefore, a 95% confidence interval for precision is given by

$$\left(\sqrt{(\chi^2_{N-1,0.025}\sigma^2/N)},\ \sqrt{(\chi^2_{N-1,0.975}\sigma^2/N)}\right)$$

where $X^2_{N-1,P}$ is the Pth quantile of the χ^2 distribution with $N - 1$ degrees of freedom. In estimating this interval, we replace σ^2/N by s.

With our normality assumption, RMSE^2 is distributed according to σ^2/N times a noncentral χ^2 distribution with N degrees of freedom and noncentrality parameter $\lambda = N(\tilde{p}_1 - p_1)^2/\sigma^2$. This distribution may be approximated by $(\sigma^2/N)(cX + b)$, where X is a χ^2 distribution with f degrees of freedom, $c = (n + 3\lambda)/(n + 2\lambda)$, $b = -\lambda^2/(n + 3\lambda)$, and $f = n + \lambda^2(3n + 8\lambda)/(n + 3\lambda)^2$ (this is Pearson's approximation; see Johnson and Kotz 1970, p. 139). Therefore, an approximate 95% confidence interval for RMSE is given by

$$\left(\sqrt{(\sigma^2/N)(c\chi^2_{f,0.025} + b)},\ \sqrt{(\sigma^2/N)(c\chi^2_{f,0.975} + b)}\right)$$

In estimating this interval, we again replace σ^2/N by s and $(\tilde{p}_1 - p_1)$ by the estimated bias.