Validation of fish ageing methods should involve bias estimation rather than hypothesis testing: a proposed approach for bomb radiocarbon validations

R.I.C. Chris Francis, Steven E. Campana, and Helen L. Neil

Abstract: The need to validate methods of ageing fish is widely accepted and several approaches to validation have been used. Most validations are essentially informal tests, using graphical methods, of the null hypothesis of zero bias in the age estimates. It is argued that it would be more useful to estimate a confidence interval for this bias. This would provide both a quantitative measure of the strength of the validation and a means of formalising the hypothesis test. A method of estimating this confidence interval is proposed for validations based on bomb radiocarbon, and this is illustrated using data for bluenose (*Hyperoglyphe antarctica*) and haddock (*Melanogrammus aeglefinus*).

Résumé : La nécessité de valider les méthodes de détermination de l'âge chez les poissons est universellement reconnue et plusieurs méthodologies de validation ont été utilisées. La plupart des validations sont essentiellement des tests informels, à l'aide de méthodes graphiques, de l'hypothèse nulle voulant qu'il n'y ait aucun biais dans les estimations de l'âge. On a proposé qu'il serait plus utile d'estimer un intervalle de confiance pour ce biais. Cela fournirait à la fois une mesure quantitative de la force de la validation et une manière de formaliser le test d'hypothèse. Nous proposons une méthode pour estimer cet intervalle de confiance pour les validations qui est basée sur le radiocarbone des essais nucléaires et nous l'illustrons avec des données sur la rouffe antarctique (*Hyperoglyphe antarctica*) et l'aiglefin (*Melanogrammus aeglefinus*).

[Traduit par la Rédaction]

Introduction

The ageing of large numbers of fish is a fundamental part of the assessment, and thus management, of many fisheries throughout the world (Morison et al. 1998). The methods used (the counting of annual rings in otoliths, scales, or other hard parts) are simple in principle but often difficult in practice, requiring great skill in preparing samples for counting and in distinguishing annual rings from other marks. Inaccurate age estimates could compromise stock assessments, so it is important that ageing methods be validated (Beamish and McFarlane 1983; Campana 2001). Moreover, validations must be carried out for each new species and sometimes even for different stocks of the same species, because some aspects of successful ageing procedures can be very species-specific. A variety of validation methods have been used, including marginal increment analysis (Cappo et al. 2000), chemical marking of otoliths at tagging (Fowler 1990), radiochemical dating (Andrews et al. 2009), and bomb radiocarbon (Kalish 1993).

In this paper, we first argue that age validation assesses the possibility of bias in age estimates and that most validations are effectively informal tests of the null hypothesis of no bias. We suggest that validations would be more useful if they focussed on the estimation of bias rather than on hypothesis testing, and propose a method of achieving this for validations based on bomb radiocarbon. The method is illustrated using data for bluenose (*Hyperoglyphe antarctica*) and haddock (*Melanogrammus aeglefinus*).

What is needed from an age validation?

Loosely speaking, what is needed from a validation of an ageing method for a particular fish species (or stock) is a determination of whether the age estimates produced are, on average, approximately correct. The two phrases "on average" and "approximately" are important here. The former is necessary because even with the most rigorously controlled ageing procedures, it is common to find differences among repeated age estimates for the same fish. Thus we know that not all age estimates will be correct. All that we can hope is that they will be correct on average. That is to say, they will be unbiased: there will not be a consistent tendency to over- or under-estimate age. Note that our focus in age validations

Received 1 December 2009. Accepted 28 May 2010. Published on the NRC Research Press Web site at cjfas.nrc.ca on 10 August 2010. J21542

Paper handled by Associate Editor Bronwyn Gillanders.

R.I.C.C. Francis[1] and H.L. Neil. National Institute of Water & Atmospheric Research Ltd. (NIWA), 301 Evans Bay Parade, Greta Point, Private Bag 14901, Wellington, New Zealand.
S.E. Campana. Fisheries and Oceans Canada, Population Ecology Division, Bedford Institute of Oceanography, P.O. Box 1006, Dartmouth, NS B2Y 4A2, Canada.

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

Fig. 1. An example of age validation using otolith marginal increments: monthly mean percent marginal increment; vertical lines represent ±1 standard error (SE) and the numbers above them are sample sizes (replotted from fig. 5 in Franks et al. 1999).

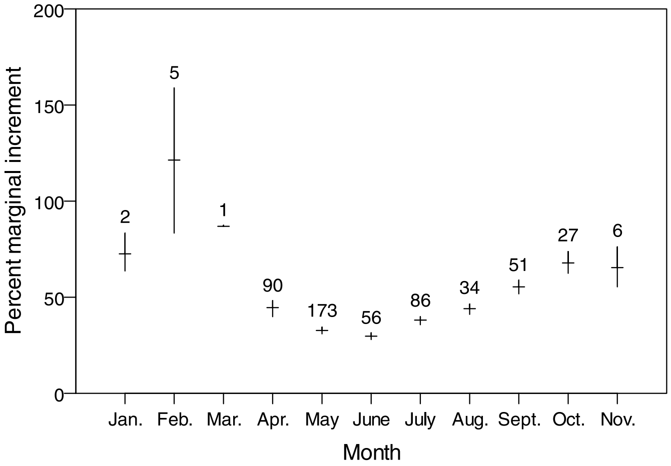
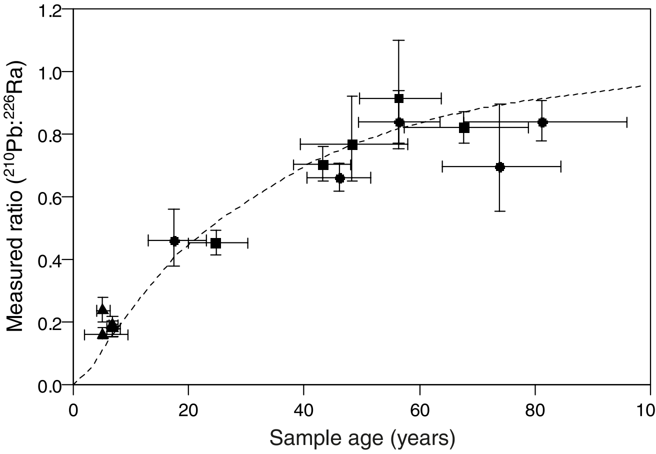


Fig. 2. An example of age validation using radiometrics: the ratio of ^{210}Pb to ^{226}Ra plotted against sample age, with horizontal and vertical error bars, and the expected ingrowth curve (broken line) (replotted from fig. 4 in Stevens et al. 2004). Triangles, juvenile age groups; circles, female age groups; squares, male age groups.



is on bias in the ageing method; we are not concerned with the precision of estimated ages (i.e., how similar repeated readings of the same otolith are). The word “approximately” is necessary because we can never expect to say that an ageing method is completely without bias. To do so would be to claim to have proven the null hypothesis. However, although a validation cannot say that an ageing method is unbiased, it might allow us to say how large any bias might be.

Most ageing studies implicitly treat validation as an exercise in hypothesis testing. The null hypothesis apparently being tested (but not usually explicitly stated) is that the ageing method is unbiased. The two possible conclusions from most validation studies — that the method was, or was not, validated — are equivalent to saying that this null hypothesis was not, or was, rejected. Testing of this hypothesis is typically informal (i.e., nonquantitative) and often graphically based. Consider two otolith-based examples typical of the recent literature. In both of these, the validation is achieved by showing that the plotted data fit some expected pattern. In the first example (Fig. 1), a plot of mean marginal increment against month was treated as providing validation because it demonstrated an annual cycle, with a new annulus apparently forming in the middle of the year (when the marginal increment is smallest). In the second (Fig. 2), the validation depended on the fact that plotted ratios of ^{210}Pb to ^{226}Ra were found to lie near a predicted curve. In what follows, we are not questioning the correctness of these two validations; we wish only to draw attention to typical methods of inference in age validations.

There are two drawbacks to this type of graphical validation. First, it is informal: no level of statistical significance is associated with the acceptance or rejection of the null hypothesis, and no objective criteria are provided to help decide how close the plotted data need to be to the expected pattern to avoid rejecting the null hypothesis (and thus failing to validate the ageing method). For example, it is not clear how distant the points in Fig. 2 could have been from the broken line before the validation would have failed. What would have been concluded if most points were near the line but a few were well away from it? In Fig. 1, the marginal increment varied from about 30% in June to about 100% in January. Would the validation have failed if the variation had only been from 45% to 70%, or 55% to 65%, or if the marginal increment for September had been substantially lower than that for August? In these examples (as in most published validations), the conclusion that the ageing method was validated was purely subjective. The second drawback is that the graphical approach provides no quantitative measure of any bias that might have been present in the ageing method. Clearly, a very small ageing bias (say 1%) could not be ruled out in either of our examples. But can we be confident that there was not a bias of 10% (or 20%, or 30%) in these ageing methods? Another way of expressing this second drawback is to say that the graphical approach provides no quantitative measure of the strength of a validation. It seems reasonable to conclude that the strength of a radiometric validation (like that in Fig. 2) is related in some way to the closeness of the sample points to the line, but there is no obvious quantitative measure of that strength.

We suggest that what is most needed from an age validation is an estimated 95% confidence interval for bias, rather than a test of the hypothesis of no bias. The width of this interval would provide a measure of the strength of the validation: a strong validation would produce a narrow confidence interval centred near zero, e.g., (–4%, +6%), and a weak validation would produce a wide interval, e.g., (–25%, +15%). This would take us away from an implausible black-and-white world view in which ageing methods could be classified only as good (validated), bad (not validated), or unknown (unvalidated) to one in which shades of grey are acknowledged in the form of strong and weak validations. This confidence interval would also allow us to formalise the associated hypothesis test in that the null hypothesis would be rejected (at the 5% significance level) if the 95% confidence interval excluded 0%.

In the remainder of this paper we propose a method of estimating such a confidence interval when the validation is based on bomb radiocarbon data. Initially, we will assume that any ageing bias is proportional to the true age, and thus may be expressed as a percentage. For example, a bias of +10% means that, on average, the estimated age is 10% higher than the true age. In a later section, we will discuss why we made this assumption and how our approach to validation could be modified to accommodate other models of bias. Before presenting our proposed approach, we will describe two data sets that will be used to illustrate it.

Two data sets

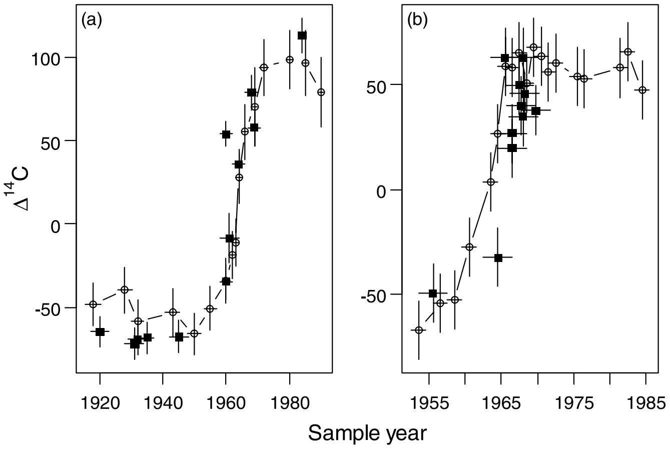
Bomb radiocarbon validation of an ageing method for a particular species requires a pair of data sets: (*i*) the test data set includes estimated ages and the associated radiocarbon values for this species and (*ii*) the reference data set includes known ages and radiocarbon values for another species (or, sometimes, the same species). We will use two such test–reference pairs. In the first (Fig. 3*a*), the test data are for bluenose (Table 1, with details in Appendix A) and the reference data set is that for snapper (*Pagrus auratus*) presented by Kalish (1993; table 1), with ageing-error estimates derived for the present study (see Appendix B) and assumed sample (core) age 0. In the second pair (Fig. 3*b*), the test data are for adult haddock (the test data in table 1 of Campana 1997, with sample (core) age 0.5 years and assumed ageing error SE 1 year) and the reference data set is the Northwest Atlantic (NWA) set developed, using juvenile haddock, redfish (*Sebastes* spp.), and yellowtail flounder (*Limanda ferruginea*) of known age, by Campana et al. (2008) (Table 2).

Both pairs of data sets show, at least approximately, the pattern expected for a successful validation based on bomb radiocarbon (i.e., the pattern to which the data must conform if the ageing method is to be validated). This pattern is associated with the atmospheric testing of nuclear weapons in the 1950s and 1960s, which produced a rapid rise in the concentration of atmospheric ^{14}C (measured as $\Delta^{14}C$). Both test and reference data sets are expected to show a similar rapid rise in $\Delta^{14}C$ beginning in the same year. The timing of this rise will often be somewhat later than that for the atmosphere because of the time taken for ^{14}C to migrate from the atmosphere to the fish's environment (i.e., the adjacent water mass) and from there into the otolith. Subsequent $\Delta^{14}C$ levels in the aquatic environment decline gradually, due not to radioactive decay (the half-life of radiocarbon is 5730 years), but to gradual sequestration of the carbon out of the water column (e.g., into the sediments or deeper water). Both of our example data sets show low levels of ^{14}C before the late 1950s and a sudden rise in the 1960s (Fig. 3). What is crucial to bomb radiocarbon validations is that the test and reference data sets should agree during the period of rapid increase of ^{14}C. For that reason, our proposed approach to validation will ignore data before and after that period (see below).

The effect of ageing bias on bomb radiocarbon plots

Because validation concerns the existence (or extent) of bias in age estimates, it is of interest to ask what effect any such bias would have on plots such as Fig. 3. The effect, as was shown by Kastelle et al. (2008), is to shift the test data points horizontally. This is illustrated in Fig. 4, which shows how we would change our plotting of the test data if we believed that the ages were biased. Note that because bias is assumed to be proportional to age, the amount of horizontal shift is different for different points. For example, fish BNS2 and BNS13 both have sample year 1960 (though they were sampled at very different ages in different years; see Table 1), but when replotted after correction for a 20% bias, they are shifted to 1964 and 1967, respectively (see Fig. 4*c*).

Fig. 3. Two examples of test–reference pairs of data sets (solid squares are test points; open circles are reference points) used in bomb radiocarbon age validation: (*a*) bluenose (*Hyperoglyphe antarctica*) – snapper (*Pagrus auratus*) (test–reference), and (*b*) haddock (*Melanogrammus aeglefinus*) – Northwest Atlantic (NWA). For each plotted point, sample year is the estimated date of formation of the otolith material that was analysed to produce the $\Delta^{14}C$ value, and error bars are ± 2 SE.



We can use plots such as Fig. 4 to make informal inferences about what levels of ageing bias are plausible. For example, we can probably rule out a bias of –20% for bluenose because that results in most of the test data points from the period of increasing ^{14}C being well to the left of the reference line (Fig. 4*a*). A bias of +20% is not so easy to rule out. Although this puts almost all test points to the right of the reference line, they are not far from the line (Fig. 4*c*). In fact, on average, they are about as close to the reference line (but on the other side of it) as those in Fig. 4*b*. Thus we might infer that biases of 0% and +20% are approximately equally plausible for bluenose ageing and that our best estimate of ageing bias for this species might be near +10%.

Our proposed approach provides a way to formalise such inferences. For example, it will allow us to determine how likely it is that the pattern shown in Fig. 4*b* (with most data points just to the left of the reference line) could occur by chance (because of random sampling errors in $\Delta^{14}C$ and age) if the ageing were unbiased. If we were to find that such a pattern was very unlikely to occur by chance, we would reject the null hypothesis of no bias.

Proposed approach to bomb radiocarbon age validation

A fundamental assumption underlying our approach is that the test and reference species occupy the same, or sim-

Table 1. Radiocarbon and age data for core samples from bluenose (*Hyperoglyphe antarctica*) otoliths (for details of sample collection and preparation, see Appendix A).

Sample	Catch year	Fish age (years)	SE of age[a] (years)	Sample age[b] (years)	Δ^{14}C	SE of Δ^{14}C	Sample year[c]
B01	1986	55	1.5	0	–71.8	4.7	1931
B05	1980	45	1.0	0	–68.3	4.8	1935
B08	1980	48	1.5	0	–69.3	4.7	1932
B10	1980	35	1.5	0	–67.7	4.9	1945
B14	1980	60	1.5	0	–64.5	4.6	1920
BNS2	1980	22	1.0	2	–34.5	5.4	1960
BNS5	1980	13	1.0	2	57.7	5.5	1969
BNS6	1980	21	1.5	2	–8.6	7.5	1961
BNS7	1980	18	1.0	2	35.6	4.4	1964
BNS8	1980	14	1.0	2	78.7	5.2	1968
BNS13	1999	41	1.0	2	54.0	3.6	1960
BNS15	1985	3	0.5	2	113.0	5.3	1984

[a] Estimated informally by the otolith reader as half of the maximum likely error.

[b] Estimated age of the fish at which 50% of the material in the core sample from the otolith was formed.

[c] Estimated year of formation of the otolith core, calculated as catch year – fish age + sample age.

Table 2. Radiocarbon and age data for the Northwest Atlantic (NWA) reference data set of Campana et al. (2008).

Catch year	Fish age (years)	SE of age (years)	Sample age (years)	Δ^{14}C	SE of Δ^{14}C
1954	1	0.5	0.5	–67.3	7
1957	1	0.5	0.5	–54.5	7
1959	1	0.5	0.5	–53	7
1961	1	0.5	0.5	–27.8	7
1964	1	0.5	0.5	3.6	7
1965	1	0.5	0.5	26.4	7
1966	1	0.5	0.5	58.9	7
1967	1	0.5	0.5	58.2	7
1968	1	0.5	0.5	65.5	7
1969	1	0.5	0.5	50.6	7
1970	1	0.5	0.5	68	7
1971	1	0.5	0.5	63.7	7
1972	1	0.5	0.5	55.9	7
1973	1	0.5	0.5	60.4	7
1976	1	0.5	0.5	53.9	7
1977	1	0.5	0.5	52.8	7
1982	1	0.5	0.5	58	7
1983	1	0.5	0.5	65.9	7
1985	1	0.5	0.5	47.6	7

ilar, environments, so that the carbon incorporated into the otoliths of the two species in the same year will contain the same proportion of ^{14}C. We discuss below what can be done when this "same environment" assumption does not hold.

The first step in our proposed approach is to restrict both data sets. The main aim of this restriction is to include only those points for which Δ^{14}C is rapidly increasing because, as pointed out above, a bomb radiocarbon validation depends on agreement between these parts of the test and data sets (e.g., for the bluenose–snapper data set, we considered only those points for which the sample year lay between 1955 and 1972, inclusive). In the years before and after this period of rapid increase, Δ^{14}C values change too slowly to allow for effective discrimination in the year of deposition. A secondary requirement of this restriction is that it must exclude any test point whose Δ^{14}C value lies outside the range of the Δ^{14}C values in the restricted set of reference points (because we cannot say whether such a test point is to the left or the right of the reference line).

In describing our proposed approach, we will focus on the main concepts rather than the details (which are given in Appendix C), illustrating these concepts using our two data sets. First, we define a statistic, h, which measures the horizontal displacement of the test data relative to the reference line in plots such as Fig. 4. The definition is simple: h = median(h_i), where h_i is the horizontal distance of the ith test data point from the reference curve. It is important that h is defined in terms of horizontal distance (rather than shortest distance or vertical distance) because our aim is to detect ageing bias, and this bias causes a horizontal displacement. The distance is measured in units of standard error so as to give less weight to data points with larger uncertainty (see Appendix C). If the points are mostly to the left (or the right) of the line, h will be negative (or positive). For example, the calculated values of h for the three panels of Fig. 4 are –5.27, –0.97, and 1.76 for assumed biases of –20%, 0%, and +20%, respectively (Table 3 extends this calculation to further levels of assumed bias).

We can use h to refine the simple inferences described above. For example, our best estimate of bias should be that for which h = 0 (because this means that the test data are exactly centred on the reference line). Interpolating from the values in Table 3, we see that the best estimate of bias for the bluenose ageing method is about +6%. We can also see that, contrary to what was suggested above, a bias of 0% is actually more plausible than one of +20%, because its associated value of h (–0.97) is closer to zero than that for 20% bias (1.76).

To be able to say how plausible a particular level of bias might be, we need to know how far h could deviate from 0 as a result of sampling error. We quantify this using a simulation experiment. For the purposes of this experiment, we assume that a line fitted to the reference data set represents the "truth" for both the reference and test species. Thus, in

Fig. 4. The effect of an assumed ageing bias of (*a*) –20%, (*b*) 0%, or (*c*) +20% on the relationship between the bluenose (*Hyperoglyphe antarctica*) test data (solid squares) and a line joining the (snapper (*Pagrus auratus*)) reference data (line). The reference line is the same in each panel; for panels (*a*) and (*c*), the sample year for each test data point was recalculated (using the formula in footnote *c* of Table 1) after the tabulated fish age was corrected for the assumed bias (by dividing by 0.8 or 1.2 for a bias of –20% or 20%, respectively); in panel (*b*), the test data points are the same as in Fig. 3*a*.

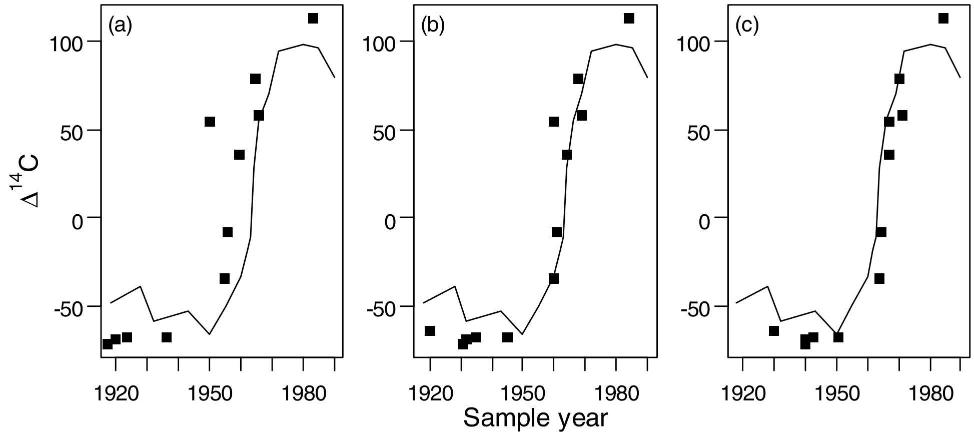


Table 3. The effect of assumed bias on the statistic *h*, which measures the extent to which the bluenose (*Hyperoglyphe antarctica*) test data points tend to be to the left (*h* < 0) or right (*h* > 0) of the snapper (*Pagrus auratus*) reference curve in plots like those in Fig. 4.

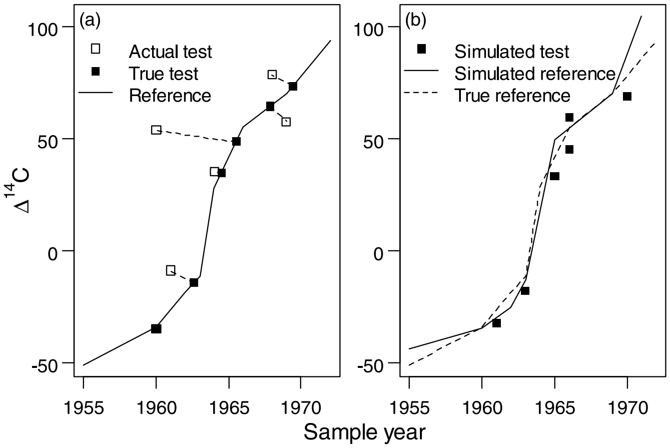
	Assumed bias (%)								
	–20	–15	–10	–5	0	+5	+10	+15	+20
h	–5.27	–3.93	–2.64	–1.81	–0.97	–0.14	0.69	1.03	1.76

this experiment, the “true” positions of all our test and reference data points lie on this line, although the observed positions deviate from the line because of measurement error (in both age and Δ^{14}C). The “true” position of each test and reference data point is assumed to be the point on the reference line that is “closest” to the actual point (“closest” is defined in terms of the standard errors for age and Δ^{14}C — see Appendix C). For the bluenose–snapper data, the “truth” assumed for the simulation was represented by the line and solid squares in Fig. 5*a*.

Five-thousand test–reference data sets are simulated by adding normal random errors to the “true” points. The size of the added observation errors is determined by the standard errors for the real data (e.g., those given in Table 1). Each simulated data set may be thought of as an example of what could have been observed given the assumed “truth”; we show an example of a simulated data set for bluenose–snapper (Fig. 5*b*). Note that the simulated reference line in this example differs slightly from the “true” one (because of measurement errors), and by chance, all but one of the simulated test data points lie to the right of this line. The statistic *h* is calculated for each simulated data set, and the resulting 5000 values of *h* ranged from –3.2 to 3.2, with 95% of them lying between –1.51 and 1.55 (Fig. 6). This tells us that if bluenose ageing is unbiased, we can be 95% confident that the value of *h* that we calculate from our data will lie between –1.51 and 1.55.

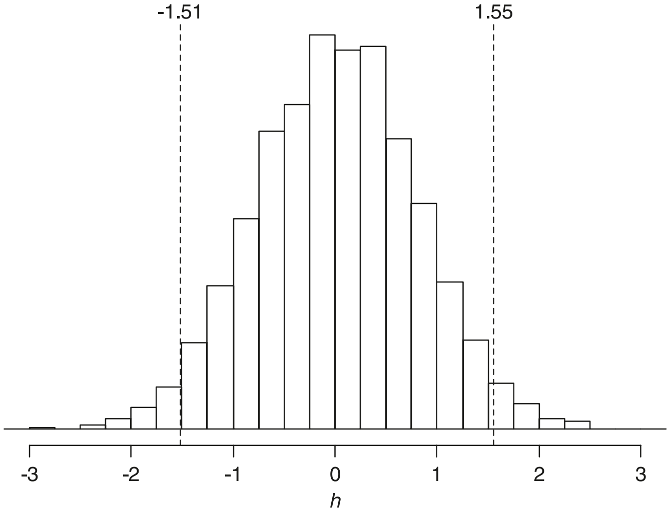
Our final step involves converting this 95% confidence interval for *h* into a confidence interval for bias. We do this for our bluenose–snapper example simply by interpolating to find the values of bias that correspond to *h* values of –1.51 and 1.55 (Table 3). These are –3% and +19%, respectively (to the nearest whole percentage point), so our 95% confidence interval for bias is (–3%, +19%).

Fig. 5. Illustration of the simulation experiment used to calculate a 95% confidence interval for ageing bias for bluenose (*Hyperoglyphe antarctica*): (*a*) the “true” data, derived from Fig. 3*a* (the reference line was fitted to the snapper (*Pagrus auratus*) data points), restricted to the years during which Δ^{14}C increased rapidly; also shown is how each “true” test point (solid square) in the simulation experiment is the point on this line that is closest to an actual test point (open square); (*b*) simulated data generated by adding random errors to the “true” data (to aid comparison, the true reference line from panel (*a*) is added to this panel as a broken line).



The logic of this final step may not be immediately clear, so here is a rationale for it. Suppose we want to evaluate the possibility that bluenose ageing has a bias of, say, +20%. If

Fig. 6. Histogram of values of the statistic *h* for each of 5000 simulated bluenose (*Hyperoglyphe antarctica*) – snapper (*Pagrus auratus*) data sets. Vertical broken lines show a 95% confidence interval for *h* (defined by the 0.025 and 0.975 quantiles of the set of *h* values).



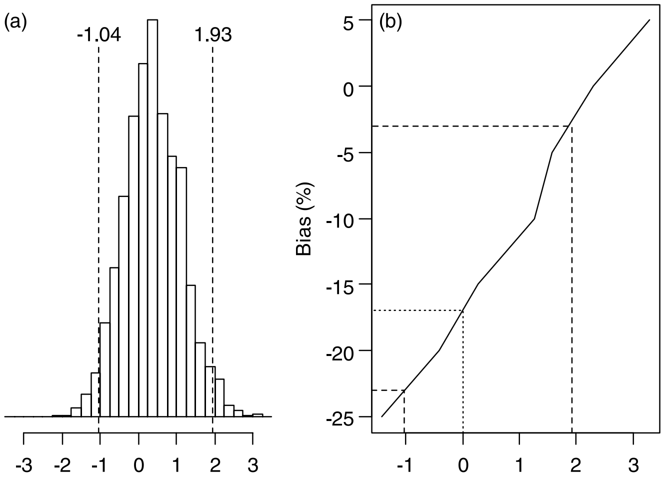
we believed that the bias was +20%, we would replot our data after correcting for this bias, and this would produce Fig. 4*c*. If our belief were right, then the corrected ages would be unbiased. Now, our simulation experiment has demonstrated that, if the ages are unbiased, we can be 95% confident that the *h* value for this plot will lie in the interval (–1.51, 1.55). Because the *h* value for Fig. 4*c* is 1.76, which is outside this confidence interval, we can reject (with 95% confidence) the possibility that the bias is as high as +20%. Repeating this argument for other possible values of bias leads us to the bias confidence interval (–3%, +19%).

The fact that this bias confidence interval is so wide is mainly due to of the small sample sizes in the restricted data sets (eight reference points and six test points), but also depends on the size of the standard errors. For example, had the snapper data been measured without error (in either age or Δ^{14}C), the estimated bias confidence interval for bluenose would have been reduced from (–3%, +19%) to (–1%, +16%).

Application of our approach to the haddock–NWA data set (restricted to sample years 1953–1970) produced 95% confidence intervals of (–1.04, 1.93) for *h* and (–23%, –3%) for bias (Fig. 7). Thus we reject the null hypothesis of no ageing bias for haddock. Of course, this rejection is conditional on the assumption that ageing bias is proportional to the true age (see below for a discussion of other bias models). It is not surprising that the best estimate of bias, –17%, is negative, as the majority of the haddock data points lie to the right of the reference data (Fig. 3*b*).

What did seem surprising was that the distribution of *h* for haddock–NWA was not centred near zero (Fig. 7*a*). The reasons for this are quite complex. Recall that the simulation experiment that produced this distribution started by associating a "true" test data point, lying on the reference line, with each observed point (as illustrated for bluenose–snapper in Fig. 5*a*). Then, for each of these "true" test data points, 5000 simulated points were generated by adding random observation error. Overall, we would expect about 50% of these simulated points to fall to the right of the reference line. However, the actual percentage that falls to the right depends on the shape of the reference line near each "true" point. The percentage is higher if this shape is convex downwards and lower if it is convex upwards. By chance, it happened that most of the "true" haddock test points were in the former category. Thus, simulated test points tended to be to the right of the reference line, which caused the mean *h* value to be positive.

Fig. 7. Some steps in the calculation of a bias 95% confidence interval using the haddock (*Melanogrammus aeglefinus*) – Northwest Atlantic (NWA) data sets: (*a*) the distribution of *h* from the simulation experiment, with broken lines showing the 95% confidence interval for *h* (–1.04, 1.93); and (*b*) the calculation of the bias confidence interval (–23%, –3%) (the solid line shows the calculated relationship between *h* and bias; the broken lines connect the two confidence intervals; and the dotted line shows how the best estimate of bias (–17%) is that corresponding to *h* = 0).



Some weaknesses of the proposed approach

It must be acknowledged that our proposed approach has several weaknesses, which are exacerbated by the small sample sizes typical of bomb radiocarbon validation data sets. Although our approach is intended to be as objective as possible, we have not been able to avoid all subjective, or arbitrary, decisions. One such decision is the range of sample years to include when restricting the data set to those points for which Δ^{14}C is rapidly increasing. The combination of measurement error (in ages and Δ^{14}C) and small sample sizes means that there is no clear-cut point at which Δ^{14}C levels change from being more or less stable to rapidly increasing (or vice versa). Fortunately, the bluenose–snapper analysis is not sensitive to the range of sample years used (Table 4*a*). Another arbitrary decision was the choice of a set of bias values at which to evaluate *h*. This matters because the relationship between bias and *h* is not completely smooth (a consequence of small sample sizes). However, it makes little difference to our estimated bias confidence interval if we use a bias step size of 1% or 2%, rather than the 5% used in Table 3 (Table 4*b*). Also, the results of any simulation experiment will depend, to some extent, on the random number seed (which determines the sequence of random numbers that is used to generate simu-

Table 4. Effect, on the estimated lower and upper bounds of a 95% confidence interval for bias in bluenose (*Hyperoglyphe antarctica*) ageing, of changing some arbitrary decisions. The original estimated bounds are given, together with estimates using (*a*) different ranges of sample years for the restricted data sets (the original range was 1955–1972), (*b*) different bias step sizes for Table 3 (the original step size was 5%), and (*c*) different random number seeds.

		(*a*) Range of sample years			(*b*) Bias step size (%)		(*c*) Different random number seeds			
	Original	1950–1980	1950–1972	1955–1980	1	2				
Lower (%)	–3.2	–3.4	–3.2	–3.4	–2.9	–3.2	–3.1	–3.2	–3.3	–3.2
Upper (%)	+18.6	+18.9	+18.5	+18.2	+17.7	+18.2	+18.7	+18.2	+18.5	+18.9

lated data points). A different random number seed leads to a different distribution for *h* in Fig. 6 and thus different confidence intervals for *h* and bias. When the simulations were repeated with four different seeds, the effect on the bias confidence interval was small (Table 4*c*). It is because of the uncertainties illustrated in Table 4 that we chose to round the bounds of our bias confidence intervals to the nearest whole percentage point.

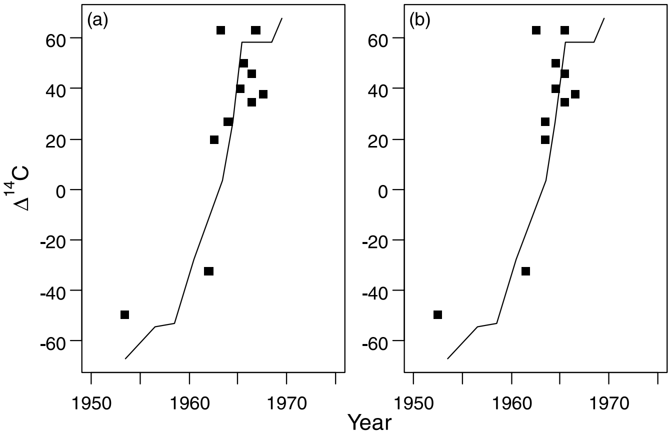
Another arbitrary decision was the choice of method used to fit a line to a set of observed or simulated reference points (see algorithm A, Appendix C). This choice will affect the estimated bias confidence interval, particularly when sample sizes are small. The method we chose, R function “isoreg” (http://www.r-project.org/), has two advantages. First, the fitted line is always increasing, which avoids ambiguity in the calculation of *h*. Second, unlike locally weighted regression procedures such as LOWESS (Cleveland 1981), it does not require the user to make a subjective decision about the degree of smoothing to be applied. Initial exploratory analyses using LOWESS showed that this procedure was unsuitable for present purposes because the estimated confidence interval for bias was found to be strongly affected by the degree of smoothing chosen by the user.

Error in age estimation ought to apply to both the fish age and the sample age. However, in simulating data, we ignored error in the sample ages (see step 5.1, Appendix C). This was done because each simulated data set was intended to be an example of what could have been observed, given the simulation assumptions. In our data sets, the calculated sample years were always integers for the bluenose–snapper data (because ages and sample ages were integers) and always half years (e.g., 1968.5, 1969.5) for the haddock–NWA data (because ages were integers, and sample ages were all 0.5 years). It seemed important that the sample years for the simulated data sets should have the same properties. We achieved this by (*i*) adding ageing errors to the “true” ages, (*ii*) rounding the resulting ages to the nearest year, and (*iii*) calculating the sample years using these simulated ages and the observed catch years and sample ages.

Other models for ageing bias

The proportional model for ageing bias, which we have assumed above, has only one parameter: the percentage bias. The only other one-parameter bias model assumes constant bias. For example, we may assume that the average difference between estimated age and true age is +1 year, independent of the true age. Of these two models, the former seems most widely applicable for fish ageing and is most consistent with the widespread use of the coefficient of variation and average percentage error in quantifying ageing precision (Campana 2001). Certainly, when the ages in a data set cover a wide range (as they do for bluenose — see Table 1), it would be unreasonable to assume that bias is independent of age. An argument supporting the proportional model considers the process of ageing as a series of decisions as to which marks in an otolith are annual rings and which are false. The older a fish is, the more decisions the otolith reader must make. A reader who is overly conservative about these decisions (i.e., too reluctant to count a ring as annual) will typically underestimate age. If, for example, they label 10% of annual rings as false, their estimates will have a bias of +10%. The constant bias model is more plausible for a data set with a narrow range of ages. However, note that the narrower the range of ages is, the less there is to choose between the two models (e.g., if all fish in a sample are aged close to 20 years, there is little difference between a proportional bias of –10% and a constant bias of –2 years).

Fig. 8. The haddock (*Melanogrammus aeglefinus*) – Northwest Atlantic (NWA) data set replotted assuming (*a*) a proportional bias of –17% (as estimated in this study) and (*b*) a constant bias of –3 years.



It would not be difficult to modify our approach to use a constant bias model. However, it should be noted that, given the small sample sizes that are typical in bomb validation data sets (particularly when these data sets are restricted to the period of rapidly rising Δ^{14}C), it will often not be possible to decide which bias model is superior. For example, a plot of the haddock–NWA data suggests that there is little to choose between our estimated proportional bias of –17% for haddock and a constant bias of –3 years (Fig. 8).

These one-parameter models are probably unrealistically simple for most species, because the clarity of annual rings in fish otoliths often varies with ring count. Thus, it would often be more realistic to describe bias as varying with age. For example, in the original analysis of the haddock data, Campana (1997) concluded that there was no significant bias up to age 10 years but that there was a negative bias of 2–3 years between ages 14 and 22 years. Such a model would require at least three parameters (two for the levels of bias for "young" and "old" fish and one for the age at which fish became "old"). However, when we are fitting a model to data (as is done in our approach to bomb radiocarbon validation), we should apply Occam's razor and use the simplest model consistent with the data. We believe that it would be very rare to find a bomb radiocarbon data set that allowed the estimation of more than one parameter for ageing bias.

Discussion

We have argued that the usual approaches to validating fish ageing methods are poor because (*i*) they rely on subjective decisions (usually about whether or not some plotted data fit an expected pattern) and (*ii*) they provide no measure of the strength of the validation. These approaches foster an implausible black-and-white world view in which ageing methods may be classified only as good (validated), bad (not validated), or unknown (unvalidated). The approach that we propose, for validations based on bomb radiocarbon data, avoids both these criticisms by estimating a 95% confidence interval for bias. This allows a formal test of the null hypothesis that the ageing method is unbiased (the hypothesis is rejected if the confidence interval excludes 0). Also, the width of the confidence interval provides a measure of the strength of the validation and thus allows a more realistic world view in which shades of grey are perceptible in the form of strong and weak validations. For example, although our analysis of the bluenose–snapper data can be said to validate the bluenose ageing method, this validation is weak because the confidence interval (–3%, +19%) is broad. It is important that people using this validated ageing method know that their estimated ages are probably positively biased and that this bias could well be over 10%. For example, in a stock assessment setting, they may want to conduct model runs with bias-adjusted ages to see how sensitive their assessment is to this bias. Even when the validation fails (i.e., the null hypothesis of no bias is rejected), as happened with the haddock–NWA data, it is useful to know how wide the bias confidence interval is. Although our best estimate for bias in haddock ages is –17%, this estimate is uncertain and the actual bias could be considerably smaller than this.

What validation can, or should, be done when our "same environment" assumption (see above) does not hold? This is a contentious question about which we have not reached consensus. Suppose, for example, that the test and reference data sets both show a rapid rise in $\Delta^{14}C$ but that this rise starts and (or) ends at different $\Delta^{14}C$ levels in the two data sets. This difference of starting and (or) ending levels of $\Delta^{14}C$ indicates that the test and reference species inhabit different environments, so our "same environment" assumption is violated. In this situation, some researchers have rescaled the data so that the test and reference levels of $\Delta^{14}C$ agree in the periods both before and after the rise. They then consider the test species ageing method to be validated if the rescaled data sets are in agreement about the timing of the rapid rise in $\Delta^{14}C$. The question is, is such a validation legitimate? We are in agreement that it is not appropriate to use rescaled data in our approach to formal validation; however, we are not in agreement about whether it is legitimate to rescale in a less formal validation.

Two related studies

We offer comments on two studies that are similar to ours in that they address the problem of formalising age validations. Okamura and Semba (2009) proposed a formal method of inference for validations based on marginal increments. Their inference was set in a different framework from that of most validation studies: model selection rather than hypothesis testing. They provided an objective way (the Akaike information criterion (AIC)) of choosing between three alternative models of ring formation: annual, biannual, and acyclic. The ageing method would be considered validated if the first model was selected. This approach is a step in the right direction. It provides an objective way of dealing with the situation in which the main uncertainty is about the periodicity of ring formation. However, it does not deal well with the perhaps more common situation in which ring formation is believed to be annual, but the annual rings are not always clear. Here, the main uncertainty concerns bias. The direction and extent of any bias depends on the relative frequency of the two types of miscount: annual rings that are not counted because they are unclear, and false rings that are wrongly interpreted as annual. The method of Okamura and Semba (2009) will not quantify this bias and will not detect it unless it is extreme.

Kastelle et al. (2008) suggested three new statistical methods for use in bomb radiocarbon validations. We comment only on the second of these methods. This is quite similar to the approach proposed here in that it involves calculating a statistic, SSR (analogous to our *h*), that measures the distance between the test and reference data sets as a function of assumed ageing bias. The best estimate of bias is that associated with the smallest value of SSR (cf. our best estimate at $h = 0$). There are three differences between this approach and ours that are of little consequence, but worth mentioning for clarity. First, it uses constant bias rather than proportional bias. This is irrelevant because both methods could be modified to use a different bias model. Second, the labelling of bias by Kastelle et al. (2008) is inverted (e.g., what is labelled as an age bias of +5 years in their table 2 actually means a negative bias, or underaging, as is clear from the black drum results on their p. 1106). Third, the *x* variable in their plots (and thus in their calculation of SSR) is birth year rather than sample year (as defined above). This difference does not matter as long as the sample ages for the test and reference data sets are the same (as they are for their main data set in their fig. 1).

There are several ways in which we believe our approach is better than that of Kastelle et al. (2008). First, and most importantly, our approach provides a confidence interval for bias and thus a formal method of testing the null hypothesis of zero ageing bias. Second, because the effect of ageing bias is to displace the plotted test data from the reference line in a horizontal direction, it seems more logical to devise a statistic that measures horizontal displacement (as h does) rather than vertical displacement (as is measured by SSR). Third, our approach makes use of information about imprecision in the test and reference data sets (i.e., the SEs for age and Δ^{14}C). Fourth, we avoided the subjective decision involved in specifying parameters controlling the degree of smoothing to be applied in fitting a line to the reference data. The LOESS function used by Kastelle et al. (2008) requires two such parameters (span and degree); it is unclear to what extent their results are sensitive to these parameters.

Final comments and recommendations

We can understand that some researchers may be cautious about using our proposed method because of its rather complex calculations, which may be based on relatively few data. We recommend exploratory plots as a means of confirming that the best estimate of bias, as well as the associated confidence interval, are plausible.

We offer some recommendations to researchers wishing to apply our method. First, if possible, use formal estimates of ageing error SEs based on replicate ages (see Appendix B) rather than informal estimates, such as were used in three of our four data sets. The more realistic these SEs are, the more reliable the validation will be. Second, if there is any doubt about the range of years to use when restricting the data to the period of rapidly increasing Δ^{14}C, then the validation should be repeated with different ranges of years to see how sensitive the results are to this decision. Campana et al. (2008, p. 736) provided an objective way of defining the first year in the period of rapid Δ^{14}C increase: this is their Y_T, which is the year in which Δ^{14}C first exceeds the value C_T, which lies at 10% of the way between C_L and C_P, the lowest and peak values of Δ^{14}C. It is simple to extend this concept to define the last year in the period of rapid Δ^{14}C increase as being that at which Δ^{14}C first exceeds the level 90% of the way between C_L and C_P.

We are not able to specify recommended, or minimum acceptable, sample sizes for bomb radiocarbon validation data sets. The location of the data points and the size of measurement errors are probably just as important as their number. Of course, with our method, data points outside the period in which Δ^{14}C is rapidly increasing make no contribution to the validation. Even within this period, location is likely to be important: the bias confidence interval for haddock would probably have been narrower had the test data points been spread across the whole period of rising ^{14}C, rather than being mostly at the end of that period. Another point to note is that what is acceptable from a validation depends on the current state of knowledge. For example, although it is rather wide, the bias confidence interval for bluenose was useful because it substantially reduced uncertainty about the longevity of this species, whose otoliths are very difficult to read.

We caution against the practice of limiting the test data in a validation set to the "best" otoliths, as was done by Kastelle et al. (2008). They limited their data in two ways: (*i*) by using only otoliths with "clearer" annuli (i.e., those with good agreement between two independent readings), and (*ii*) by ignoring 11% (4/35) of the remaining test data points, which they deemed to be "noteworthy outliers" (all of which were well to the right of the reference line, by amounts suggesting that they were underaged by about 20%). The purpose of most age validations is to provide some confidence that age data used in stock assessments will not be substantially biased. For this purpose, a validation using only the "best" otoliths will be misleading unless the otoliths used in stock assessments can be similarly limited.

It is hoped that this paper will stimulate other researchers to propose methods to formalise age validations based on other types of data (e.g., radiometric or marginal increment). Though our approach for bomb radiocarbon validations is not without weaknesses (as discussed above), it seems greatly preferable to the informal and subjective methods that are the norm in this area.

Acknowledgements

We are grateful to the New Zealand Ministry of Fisheries for partial funding of this study and to Peter Horn, Peter Marriott, Larry Paul, and Rodger Sparks for help in assembling the bluenose data.

References

Andrews, A.H., Tracey, D.M., and Dunn, M.R. 2009. Lead–radium dating of orange roughy (*Hoplostethus atlanticus*): validation of a centenarian life span. Can. J. Fish. Aquat. Sci. **66**(7): 1130–1140. doi:10.1139/F09-059.

Beamish, R.J., and McFarlane, G.A. 1983. The forgotten requirement for age validation in fisheries biology. Trans. Am. Fish. Soc. **114**: 847–850.

Campana, S.E. 1997. Use of radiocarbon from nuclear fallout as a dated marker in the otoliths of haddock *Melanogrammus aeglefinus*. Mar. Ecol. Prog. Ser. **150**: 49–56. doi:10.3354/meps150049.

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**(2): 197–242. doi:10.1111/j.1095-8649.2001.tb00127.x.

Campana, S.E., Casselman, J.M., and Jones, C.M. 2008. Bomb radiocarbon chronologies in the Arctic, with implications for the age validation of lake trout (*Salvelinus namaycush*) and other Arctic species. Can. J. Fish. Aquat. Sci. **65**(4): 733–743. doi:10.1139/F08-012.

Cappo, M., Eden, P., Newman, S.J., and Robertson, S.G. 2000. A new approach to validation of periodicity and timing of opaque zone formation in the otoliths of eleven species of Lutjanus from the central Great Barrier Reef. Fish. Bull. (Washington, D.C.), **98**: 474–488.

Cleveland, W.S. 1981. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. Am. Stat. **35**(1): 54. doi:10.2307/2683591.

Fowler, A.J. 1990. Validation of annual growth increments in the otoliths of a small, tropical coral reef fish. Mar. Ecol. Prog. Ser. **64**: 25–38. doi:10.3354/meps064025.

Franks, J.S., Warren, J.R., and Buchanan, M.V. 1999. Age and growth of cobia, *Rachycentron canadum*, from the northeastern Gulf of Mexico. Fish. Bull. (Washington, D.C.), **97**: 459–471.

Kalish, J.M. 1993. Pre- and post-bomb radiocarbon in fish otoliths. Earth Planet. Sci. Lett. **114**(4): 549–554. doi:10.1016/0012-821X(93)90082-K.

Kastelle, C.R., Kimura, D.K., and Goetz, B.J. 2008. Bomb radiocarbon age validation of Pacific ocean perch (*Sebastes alutus*) using new statistical methods. Can. J. Fish. Aquat. Sci. **65**(6): 1101–1112. doi:10.1139/F08-038.

Morison, A.K., Robertson, S.G., and Smith, D.G. 1998. An integrated system for production fish aging: image analysis and quality assurance. N. Am. J. Fish. Manage. **18**(3): 587–598. doi:10.1577/1548-8675(1998)018<0587:AISFPF>2.0.CO;2.

Okamura, H., and Semba, Y. 2009. A novel statistical method for validating the periodicity of vertebral growth band formation in elasmobranch fishes. Can. J. Fish. Aquat. Sci. **66**(5): 771–780. doi:10.1139/F09-039.

Stevens, M.M., Andrews, A.H., Caillet, G.M., and Coale, K.H. 2004. Radiometric validation of age, growth, and longevity for the blackgill rockfish (*Sebastes melanostomus*). Fish. Bull. (Washington, D.C.), **102**: 711–722.

Appendix A. Generation of the bluenose data

This appendix describes the procedures followed in generating the bluenose data of Table 1. Archived otolith pairs of 12 bluenose obtained from commercial trawl or line-fishing catches off the central eastern coast of the North Island, New Zealand, were selected for examination. The fish had dates of capture from 1980 to 1999, and preliminary age estimations from transverse sections indicated that their likely birth dates covered the period before, during, and after the increase of surface oceanic radiocarbon.

One of each pair of otoliths was sectioned transversely through the nucleus to produce a section ~0.35 mm thick that was mounted on a glass slide. Using a binocular microscope at ×80 magnification with illumination by transmitted light, all dark zones visible on either side of the sulcus were counted. This is likely to be very similar to the Australian procedure for ageing the same species (Morison and Robertson 1995).

To obtain an estimated age for the otolith core sample, electronic images were made for each otolith of the thin section on which the growth zones had been counted, and of the thicker sections (from the paired otolith) from which the carbonate samples were to be taken for stable isotope and radiocarbon analyses. On a print of the thin section, the growth zone positions were marked and labelled in association with careful re-examination of the otolith section under the microscope. On the print of the thick section, the growth zones were similarly marked, and these marked zones were used to position the drill tracks for carbonate samples. These images were used to determine the single growth zone estimated to be in the centre of the sample.

Reference

Morison, A.K., and Robertson, S.G. 1995. Growth, age composition and mortality of blue-eye trevalla (*Hyperoglyphe antarctica*). Victoria Department of Conservation and Natural Resources, Victoria Fisheries Research Institute, Internal Report 218, Queenscliff, Australia.

Appendix B. The precision of snapper age estimates

The bomb radiocarbon reference set developed by Kalish (1993) and based on New Zealand snapper (*Pagrus auratus*) otoliths does not include any estimate of the precision of the snapper ages. For the present study, this precision was estimated from an analysis of between-reader differences in a separate sample of snapper from an area as close as possible to that from which the sample of Kalish (1993) came (“the east coast of the North Island, New Zealand, between East Cape and Hawke Bay”). This comprised 1828 otoliths that were collected off the eastern coast of the North Island between the southern Bay of Plenty and Wellington (observer area CEE), with each otolith having been aged separately by two readers. Two alternative ageing-error models were fitted, by maximum likelihood, to this data set, with ageing error assumed to be normally distributed and either the standard error (SE), or the coefficient of variation (CV) of this error, being a linear function of age. The former model was found to fit the data better, based on AIC (Akaike 1974), and this model estimated the SE of an age estimate, a, to be $0.349 + 0.0168a$.

References

Akaike, A. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19**(6): 716–723. doi:10.1109/TAC.1974.1100705.

Kalish, J.M. 1993. Pre- and post-bomb radiocarbon in fish otoliths. Earth Planet. Sci. Lett. **114**(4): 549–554. doi:10.1016/0012-821X(93)90082-K.

Appendix C. Calculating a 95% confidence interval for ageing bias

In this appendix, we present the full details of our proposed procedure for calculating a 95% confidence interval for ageing bias (a set of R functions to do these calculations is available from the first author). Six variables are defined for each datum in the test and reference data sets, and these are the first six variables in Table C1 (which correspond to the first six columns of Table 1). We show in step 1 below how the seventh variable, sample year (Y), is calculated from three of these first six variables (y, A, a). The pair of variables (Y, C) may be thought of as defining a point, or a set of points, on a plot like those in Fig. 3. When we want to describe a single point, we will add a subscript, so that (Y_i, C_i) stands for the ith point in the set of points (Y, C). Table C1 describes the various superscripts that are used to define what type of point, or points, are being referred to. For example, (Y^{r}, C^{r}) is the set of reference points; (Y^{rl}, C^{rl}) is the reference line fitted to this set of points; (Y^{St}, C^{St}) is a set of simulated test points, and so on.

For the following calculations, both the reference and test data sets must be restricted to the period in which Δ^{14}C is rising rapidly (see main text). The calculation of the confidence interval uses the following six steps.

1. Calculate sample years for reference and test data: $Y^{\mathrm{r}} = y^{\mathrm{r}} - (A^{\mathrm{r}} - a^{\mathrm{r}})$ and $Y^{\mathrm{t}} = y^{\mathrm{t}} - (A^{\mathrm{t}} - a^{\mathrm{t}})$ (this is the same as the calculation described in footnote *c* to Table 1)
2. Make the reference line, (Y^{rl}, C^{rl}), by fitting a line to the reference data, (Y^{r}, C^{r}), using algorithm A (described below)
3. Find the relationship between assumed bias, b, and the statistic h:

   3.1 Select a trial value, b, of assumed percentage ageing bias;

   3.2 Calculate the sample year corrected for this bias $Y^{\mathrm{ct}} = y^{\mathrm{t}} - (A^{\mathrm{ct}} - a^{\mathrm{t}})$, where $A^{\mathrm{ct}} = \mathrm{round}[A^{\mathrm{t}}/(1 + b/$

Table C1. Notation used in describing the calculation of a 95% confidence interval for ageing bias.

Variables		Superscripts	
Quantity	Symbol	Type	Symbol
Catch year	y	Reference	r
Fish age	A	Test	t
SE of fish age	s_A	Line	l
Sample age	a	True	T
Δ^{14}C	C	Simulated	S
SE of Δ^{14}C	s_C	Corrected for bias	c
Sample year	Y		

100)] is the bias-corrected age, which is rounded to the nearest year (as the observed ages are whole numbers);

3.3 Calculate the statistic h, which measures the extent to which the corrected test points, (Y^{ct}, C^{t}), are horizontally displaced from the reference line, (Y^{rl}, C^{rl}), in terms of the errors, s_A^{t}, using algorithm B (described below);

3.4 Repeat steps 3.1–3.3 for a range of trial values of b, tabulating the relationship between b and h in a table like Table 3

4. Define the "true" values for use in the simulation experiment:

4.1 Define the "true" reference line to be the same as the reference line fitted at step 2, i.e., define (Y^{Trl}, C^{Trl}) = (Y^{rl}, C^{rl});

4.2 Define the "true" reference points, (Y^{Tr}, C^{Tr}), as being the points on the "true" reference line, (Y^{Trl}, C^{Trl}), that are closest to the actual reference points, (Y^{r}, C^{r}), in terms of the errors, $\left(s_A^{\mathrm{r}}, s_C^{\mathrm{r}}\right)$ using algorithm C (described below)1;

4.3 Use the same procedure to define the "true" test points, (Y^{Tt}, C^{Tt}), using the "true" reference line, (Y^{Trl}, C^{Trl}), the actual test points, (Y^{t}, C^{t}), and the standard errors, $\left(s_A^{\mathrm{t}}, s_C^{\mathrm{t}}\right)$;

4.4 Calculate the "true" ages associated with the "true" reference and test data $A^{\mathrm{Tr}} = y^{\mathrm{Tr}} - Y^{\mathrm{Tr}} + a^{\mathrm{Tr}}$ and $A^{\mathrm{Tt}} = y^{\mathrm{Tt}} - Y^{\mathrm{Tt}} + a^{\mathrm{Tt}}$

5. Do simulations to calculate (h_{lo}, h_{hi}), a 95% confidence interval for h under the assumption that there is no ageing bias:

5.1 Generate simulated reference data, (Y^{Sr}, C^{Sr}), by adding zero-mean normal errors with SDs $\left(s_A^{\mathrm{r}}, s_C^{\mathrm{r}}\right)$ to the "true" reference data, (Y^{Tr}, C^{Tr}):

$$C^{\mathrm{Sr}} = C^{\mathrm{Tr}} + s_C^{\mathrm{r}} Z_C$$

$$A^{\mathrm{Sr}} = \mathrm{round}(A^{\mathrm{Tr}} + s_A^{\mathrm{r}} Z_A)$$

$$Y^{\mathrm{Sr}} = y^{\mathrm{r}} - (A^{\mathrm{Sr}} - a^{\mathrm{r}})$$

where Z_C and Z_A are generated as a standard normal random variables;

5.2 Use the same procedure to generate simulated test data (Y^{St}, C^{St}) from $\left(s_A^{\mathrm{t}}, s_C^{\mathrm{t}}\right)$ and (Y^{Tt}, C^{Tt});

5.3 Calculate a simulated reference line (Y^{Srl}, C^{Srl}) from (Y^{Sr}, C^{Sr}) using algorithm A;

5.4 Calculate h, measuring the displacement of the test points (Y^{St}, C^{St}) from the line (Y^{Srl}, C^{Srl}) using the errors s_A^{t} and algorithm B;

5.5 Repeat steps 5.1–5.4 5000 times to generate 5000 h values;

5.6 Define (h_{lo}, h_{hi}) as the 0.025 and 0.975 quantiles of the 5000 h values

6. Calculate (b_{lo}, b_{hi}), the 95% confidence interval for ageing bias, by interpolation using (h_{lo}, h_{hi}) and the relationship tabulated at step 3.4.

Algorithm A involves fitting a line to a set of reference points, (Y^{r}, C^{r}). First, if there are any groups of points with the same sample year Y, each group is replaced by a single point by averaging the C values within the group. Then the R function "isoreg" (http://www.r-project.org/) is used to fit a monotone increasing line to the resulting set of points. This line is defined as a set of line segments joining a series of (Y, C) points (i.e., a continuous linear spline) in which both the Y and C values are in increasing order.

Algorithm B calculates the statistic h, which quantifies the extent to which a set of test points, (Y^{t}, C^{t}), with ageing error SEs, s_A^{t}, is shifted to the left (negative h) or right (positive h) of a reference line (Y^{rl}, C^{rl}). First define, for the ith test point, $\left(Y_i^{\mathrm{t}}, C_i^{\mathrm{t}}\right)$, the quantity $h_i = \left(Y_i^{\mathrm{t}} - Y_i^{\mathrm{rl}}\right)/s_{A,i}^{\mathrm{t}}$, where Y_i^{rl} is the point on the reference line that has $C = C_i^{\mathrm{t}}$ (Y_i^{rl} is calculated by interpolation on the continuous linear spline described in algorithm A). h is defined as the median of the h_i. Note that Y_i^{rl} will not be defined for any test point for which C_i^{t} happens to lie outside the range of the C^{rl}, so such a point is ignored in calculating h.

Algorithm C finds the point (Y, C) on a reference line, (Y^{rl}, C^{rl}), that is closest to a point (Y_i, C_i) in terms of the standard errors (s_A, s_C). Define the distance between (Y_i, C_i) and (Y, C) as $d = \left[\left(\frac{Y - Y_i}{s_{A,i}}\right)^2 + \left(\frac{C - C_i}{s_{C,i}}\right)^2\right]^{0.5}$. It is straightforward to find the point within each of the line segments making up the reference line that is closest to (Y_i, C_i). Then search among these points to find the one closest to (Y_i, C_i).