Revisiting data weighting in fisheries stock assessment models

R.I.C. Chris Francis 2017

SEDAR55-RD10

12 January 2018

Fisheries Research 192 (2017) 5–15

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Revisiting data weighting in fisheries stock assessment models

R.I.C.Chris Francis

123 Overtoun Terrace, Wellington 6021, New Zealand

ARTICLE INFO

Article history:
Received 17 March 2016
Received in revised form 6 June 2016
Accepted 7 June 2016
Available online 28 June 2016

Keywords:
Fisheries stock assessment
Data weighting
Model misspecification
Data conflict

ABSTRACT

This paper revisits topics addressed in two previous papers on data weighting in fisheries stock assessment models: the first was general (Francis, 2011; Can. J. Fish. Aquat. Sci. 68, 1124–1138); the second considered the related problem of finding the best likelihood for composition data (Francis, 2014; Fish. Res. 151, 70–84). In the light of subsequent literature and experience, four topics seemed in need of increased emphasis or elaboration. (1) For composition data, it is better to think in terms of "right-weighting" (i.e., weighting that is statistically appropriate) than "down-weighting". (2) The sensitivity of some assessments to changes in weighting can sometimes be reduced by restructuring to reduce model misspecification; this is a good idea, but should be seen as complementary to data weighting, rather than an alternative to it. (3) It seems typical that more than half of the variance of composition residuals arises from process error (arising from model misspecification) rather than observation error. (4) Changing the likelihood for composition data from the multinomial to the Dirichlet-multinomial has some advantages but is not without problems. Some new topics are discussed: most iterative reweighting of composition data is multiplicative, but additive methods deserve consideration; data weighting is more complicated in state-space models; catch data should not be subject to data weighting; there are significant disadvantages in structuring age-related observations of fishery and survey catches as frequencies, rather than compositions (i.e., as numbers, rather than proportions); methods of weighting three additional data types (conditional age at length, tagging abundance; and tagging length-increment) are described.

1. Introduction

A key feature of modern fisheries stock assessment models is that they simultaneously analyse two or more types of data in a likelihood framework. This feature has been present in stock assessment models since at least the early 1980s (Fournier and Archibald, 1982). In this century it is often signalled by the adjective "integrated", both within the fisheries literature (e.g., both Maunder (2003); Maunder and Punt (2013) refer to "integrated analysis"), and in that relating to population modelling more broadly (e.g., both Newman et al. (2014); Schaub and Abadi (2011) use "integrated population modelling"). When there are multiple data types the term *data weighting* refers to decisions made by the modeller that affect the relative influence of each data type (and of each individual datum within a data type) on model outputs. In two recent papers I first showed how these decisions can have a profound effect on the outcome of stock assessments and presented some principles aimed at guiding stock assessment scientists in data weighting (Francis, 2011), and then suggested that some data-weighting difficulties could be reduced by moving away from the multinomial likelihood for composition data (Francis, 2014). In the present paper my aim is to revisit the matters covered by these papers in response both to subsequent publications and to presentations and discussions at the October 2015 workshop "Data conflict and weighting, likelihood functions, and process error" organised by CAPAM (Center for the Advancement of Population Assessment Methodology). To allow the present paper to stand alone I will combine brief summaries of the main points of the two earlier papers with new material that either elaborates on topics that I now think need more explanation, fills gaps that have become apparent, responds to subsequent literature and discussions, or identifies gaps in our knowledge.

2. Background and definitions

Much of the discussion of data weighting below will be restricted to two types of data: *abundance* (e.g, absolute or relative biomass estimates from trawl or acoustic surveys, from fishery catch per unit effort [CPUE], or from tag-recapture experiments) and *composition* (estimated proportions, by age or length, in catches from fisheries or surveys). These are by far the domi-

E-mail address: chris.francis@clear.net.nz

http://dx.doi.org/10.1016/j.fishres.2016.06.006
0165-7836/

nant (and sometimes only) data types in age-structured statistical stock assessment models. Some reference will be made to other data types.

The weight applied to each datum in a stock assessment model is defined by a parameter of the likelihood associated with that datum. This likelihood describes the assumed distribution of the error associated with that datum, which is to say the difference between O, the observed value of the datum, and E, the value expected by the model. For abundance data the most commonly used likelihoods are the lognormal or normal, and the weighting parameter is typically a c.v. (coefficient of variation) or, less frequently, an s.d. (standard deviation). For compositions, the dominant likelihood is the multinomial, for which the weighting parameter is an *effective sample size* though lognormal likelihoods (weighted either by a c.v. or by a log-space s.d.) are sometimes used (e.g., Punt and Kennedy, 1997). Here the adjective "effective" (Pennington and Vølstad 1994) is intended to emphasise that this is not an actual sample size (i.e., the number of fish measured or aged); it is usually much smaller than the actual sample size because of factors like intra-haul correlation (the fact that two fish from the same haul are typically more alike than those from different hauls). If we expect the error associated with a datum to be small (or large) then we should apply a large (or small) weight to it, which means using a small (or large) c.v. or a large (or small) effective sample size.

We can distinguish three ways of weighting data in stock assessments. The weighting is called *outside the model* if the values of the weighting parameters are calculated, and fixed, before the model is run. It is *inside the model* if the weighting parameters are estimated, along with other parameters (such as selectivities, and the population unfished biomass, B_0), each time the model is run. This includes situations in which the weighting is partially fixed before the model run, and partially estimated during the model run (e.g., we may define the c.v. of the ith observation in a set as $(c_i^2 + c^2)^{0.5}$, where the c_i are fixed beforehand and c is estimated), because it is not until the model is run that the weighting of the data set is determined. The third way of weighting data is called *iterative*, because it involves the following iterative procedure.

1. Set initial weights for the data set
2. Run the model
3. Use information from the model output to adjust the data weights
4. Repeat steps 2 and 3 as often as desired

(This terminology is slightly different from that of Francis (2011), who used the term "two-stage" to include both iterative weighting and the type of inside-the-model weighting in which the weighting is partially fixed before the model run.) Francis (2014) described likelihoods that can be weighted inside the model as *self-weighting*, and pointed out that most likelihoods commonly used for compositions (including the multinomial) are not self-weighting because they are improper. This is a major reason for the use of iterative weighting in stock assessments.

I will call the weighting of a data set in a stock assessment model *statistically appropriate* if the sizes of the errors, $(O - E)$, are consistent (in some sense) with the associated likelihood and weighting parameter(s). This is not intended to be a formal definition (thus the phrase "in some sense"), but two simple examples should demonstrate its intent. Suppose we have a set of observations, O_i, indexed by i, and the associated weighting parameters are c.v.s, c_i, so the s.d. of the ith error is given by $s_i = c_i E_i$. Then, assuming the errors are uncorrelated, our weighting will be statistically appropriate if we ensure that $\mathrm{Var}_i[(O_i - E_i)/s_i] \approx 1$. How close this variance should be to 1 for the weighting to be deemed statistically appropriate will, of course, depend on the number of observations and the values of the c_is. This method of ensuring statistical appropriateness is not valid when the O_i are the individual composition proportions because it ignores the large correlations (both positive and negative) that are common amongst the errors in these proportions and very different from the small negative correlations associated with multinomial errors (Fig. 1). For such data, Francis (2011) suggested that the weighting would be statistically more appropriate if we ensure that $\mathrm{Var}_j\left[\left(\bar{O}_j - \bar{E}_j\right)/S_j\right] \approx 1$, where, for the jth composition, $\bar{O}_j$ and $\bar{E}_j$ are the observed and expected mean age (or size), and S_j is the standard error of $\bar{O}_j$ (this is the basis of weighting methods TA1.8-10 of Francis, 2011). Francis (2014) referred to multinomial sample sizes calculated following this approach as *Pennington sample sizes*, since the approach adapts, for use with stock assessment

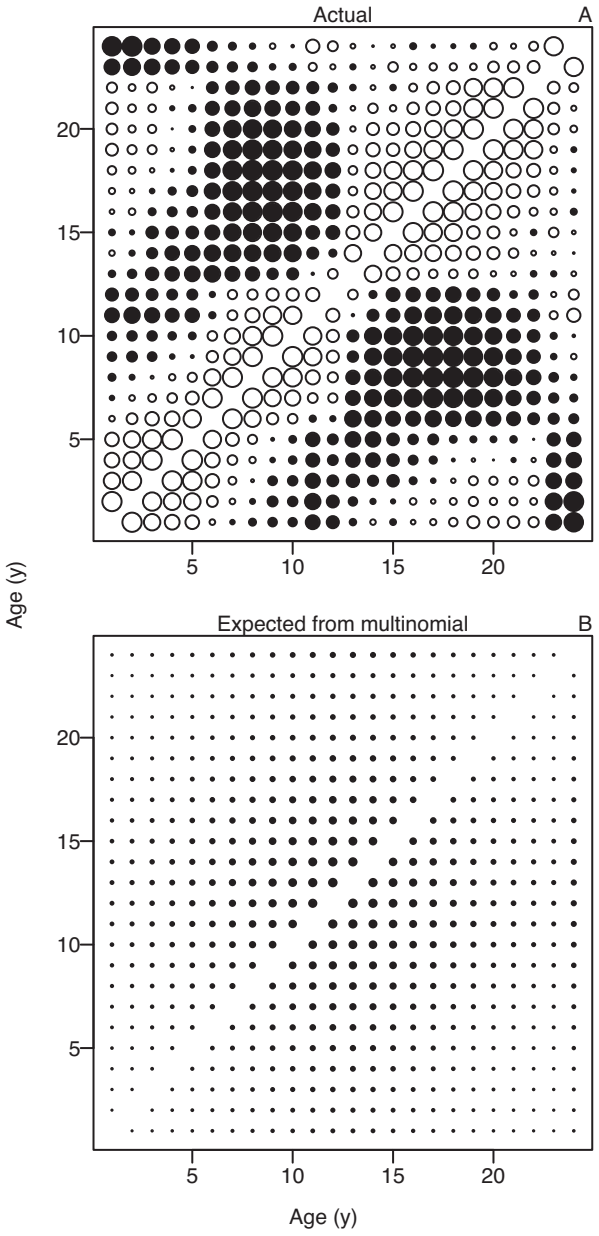

Fig. 1. Comparison, for trawl fishery data from the 2006 assessment of southern hake (*Merluccius australis*) in Chile, of A, actual between-age correlations in composition residuals with B, those expected from a multinomial distribution. The area of each circle is proportional to the corresponding correlation (with open circles for positive correlations, and filled circles for negative; the largest absolute correlation shown is 0.94). The multinomial correlations were calculated for the average (across years) of the expected compositions.

Table 1
Effect on estimated uncertainty of reweighting the (age and length) composition data in the 2014 stock assessment of the southern stock of Chilean kingclip (*Genypterus blacodes*): multinomial sample sizes for compositions (assumed to be the same for all years) and estimated standard errors of selected outputs. The reweighting used method TA1.8 of Francis (2011).

	Sample sizes		Standard errors		
Model	Age	Length	B_0 (t)	Depletion (%)	F_{final} (y^{-1})
Initial	100	50	339	2.16	0.016
Reweight	10	8	504	2.82	0.019
	Change in standard errors		+49%	+31%	+21%

residuals, a method of Pennington and Vølstad (1994) to characterise observation error in survey composition data. I acknowledge the looseness of the above definition of statistical appropriateness (particularly in the undefined precision associated with "≈"), but I can see no useful way to formalise this definition. The discussion below of the question "how many times to iterate" (Section 4.2) offers some guidance in this matter.

I will sometimes refer to the difference $(O-E)$ as the *total error* to emphasise that it can be thought of as the sum of the *observation error* $(O-T)$, where T is the true (i.e., real world) value of what we observe, and what I call the *process error*, $(T-E)$, which exists because our models are only approximations of the real world (note that the term "process error" has a different meaning in state-space models, which I discuss separately below). This decomposition of the total error is important because the two error components are very different in nature. Observation error (sometimes called sampling error) is completely independent of the model and its assumptions. Its likely size can often be estimated from the variability within the raw data from which our observations are constructed (e.g., from within-stratum variation of catch rates in a trawl survey biomass estimate), and we can often influence this size by changing our sampling intensity. By contrast, the process error is purely a consequence of our model assumptions, and its likely size may be affected by changes in these assumptions. An important example of this concerns assumptions about fishery selectivity. Real world fishery selectivities change from year to year as a consequence of changes in the spatio-temporal distribution of both the species of interest and of fishing activity (which may be affected by factors such as weather, the price of fuel, and the abundance and/or value of other species). Thus we can expect to reduce the size of process error in fishery compositions (which means that the total error will decrease, so the weighting of these data should be increased) if we allow the fishery selectivity to be time-varying, rather than the same in all years. Note that we cannot directly quantify the extent of process error, but we can estimate its variance by subtraction using $V_{process} = V_{total} - V_{observation}$ (which follows from the above definition). It's worth noting also that, although some of the correlation structure seen in composition residuals (e.g., Fig. 1) undoubtedly arises from process error, there is clear evidence that there is also some in the observation error (Hrafnkelsson and Stefánsson 2004; Miller and Skalski 2006).

The approach to data weighting presented by Francis (2011) was developed in the context of what I will call *conventional* stock assessment models (this is simply a convenient label for a common type of model; I do not wish to imply that "non-conventional" models are necessarily inferior in any way). These include the most commonly used models in north America, South Africa, Australia, and New Zealand, such as AMAK (Anonymous, 2016), ASAP (Legault and Restrepo, 1999), BAM (Craig, 2012), CASAL (Bull et al., 2012), iSCAM (Martell, 2011), MULTIFAN-CL (Fournier et al., 1998), and Stock Synthesis (Methot and Wetzel, 2013). Most of the current paper also assumes conventional models, but some different approaches to stock assessment, including state-space models, will be briefly discussed in a separate section below.

3. Why data weighting is important

Francis (2011) gave two reasons to believe that data weighting is important. The first, and more important, is that it can substantially affect the results of a stock assessment, as results from two recent stock assessment reviews show. In both reviews, models were rerun after replacing the existing weighting of composition data by that based on method TA1.8 of Francis (2011). For rougheye rockfish (*Sebastes aurora*) this reweighting reduced the estimate of *depletion* (the final year spawning biomass as a percentage of the unfished biomass) from 0.63 to 0.45 and decreased yield by "around 30%" (NMFS 2013); for Pacific sardine the reweighting was done for three alternative models and the ratios of new to old estimates of final-year 1+ biomass were 3.2, 0.38, and 0.71 (NMFS 2011). In both assessments the reweighting substantially reduced the multinomial sample sizes used to weight the composition data – by factors of about 5 for rougheye rockfish and 15 for Pacific sardine. For another example of how strongly the weighting of compositions can affect assessment outputs see Sharma et al. (2014fig. 9 [for reference points] and fig. 10 [for biomass trajectories]). It should be noted that the effects of reweighting an assessment are very variable, even when the changes in weights are as substantial as in these examples. Reweighting of a similar scale sometimes has only minor effects on assessment outputs. The important factor seems to be the extent of conflict between the different data types in the assessment: the greater this conflict, the greater the effects of reweighting. I will discuss data conflict further below.

The second reason to believe that data weighting is important is that it affects any statistical inference we may make from our assessments. The most common such inference is in the form of measures of uncertainty (e.g., a standard error or confidence interval for depletion, or the probability that the final-year spawning biomass exceeds some biological reference point). Another type of statistical inference is the use of AIC (Akaike 1974) to choose between competing models. A change in data weighting will almost always change our measures of uncertainty and may change the conclusion of an AIC-based inference. Usually, a decrease (or increase) in the weight assigned to a data set will increase (or decrease) estimated uncertainty, as illustrated in Table 1, where a 7.5- to 10-fold reduction in the weighting of composition data increased estimated standard errors by 21–49%. If our data weighting is not statistically appropriate our statistical inferences risk being invalid. It is tempting to conclude that data weighting does not matter in those assessments for which key model outputs are relatively insensitive to changes to data weighting. This conclusion will be mistaken if we attach any importance to our estimates of the uncertainty of the assessment, because these estimates are likely to be sensitive to weightings.

4. An approach to data weighting

4.1. Abundance data

Francis (2011) argued that we should prioritise abundance data in stock assessment models because (a) they contain direct information about the matters most important to stock assessment; (b)

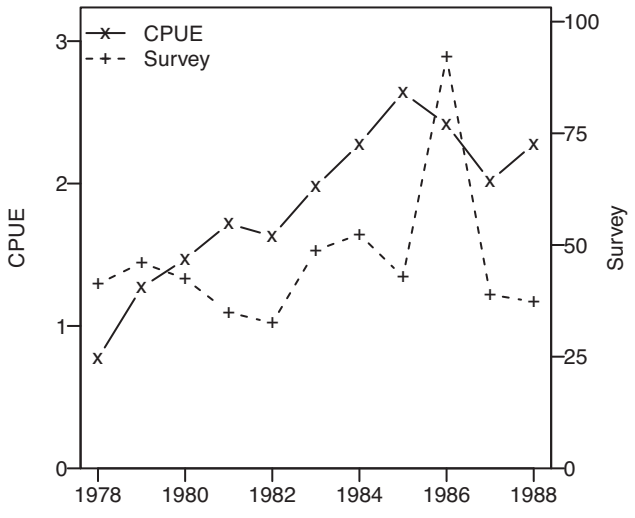

Fig. 2. A famous example of conflict between two abundance data sets: survey and CPUE data from the Canadian 2J3KL stock of cod (*Gadus morhua*) which collapsed in the early 1990s. Replotted with permission from Fig. 1 of Schnute and Hilborn (1993) [© Canadian Science Publishing or its licensors].

composition data contain comparatively little direct information about these matters (because of uncertainty about natural mortality and selectivity), although model misspecification sometimes makes it appear otherwise (see below); and (c) without such prioritisation there is a danger that the abundance data will not be well fitted in our models because their influence will be swamped by the much more numerous composition data. By "prioritise" I simply mean that we should ensure that our abundance data are fitted as well as possible. There are two techniques we can use to this end.

The first technique to encourage acceptable fits to abundance data is to weight these data outside the model. The weights must allow for both observation and process error. Francis (2011) described three approaches for setting them (adding process error to trawl survey observation-error c.v.s; a simulation technique for acoustic surveys; and, for CPUE, the method of Clark and Hare (2006) using a data smoother). Some scientists weight their abundance data inside the model (e.g., by use of concentrated likelihoods (Brodziak, 2005) or by estimating a process-error c.v. inside the model). This approach is statistically sound, and will often produce acceptable results, but I caution against it on the grounds that it can result in poor fits to the abundance data, particularly when these data are in apparent conflict with other data (a situation discussed below). When we fix the abundance weights we encourage the model to find parameter values that produce acceptable fits to these data; when we allow the model to estimate these weights we are effectively giving it permission to down-weight these data to justify a poor fit (i.e., we are not prioritising the abundance data). If the abundance data are weighted inside the model it is important to check that this does not lead to a poor fit to these data.

The second technique to encourage acceptable fits to abundance data is the use of alternative models fitted to different subsets of the available data. For example, when two abundance data sets show such different trends that it is impossible to fit both well (e.g., Fig. 2) we must acknowledge that at least one of them must be wrong. I think that the best way to express our uncertainty about which is wrong, and to evaluate the consequences of this uncertainty, is to create two alternative models, each of which uses just one of the abundance data sets (and fits it well). This uncertainty is hidden when we include both abundance indices in our model but fit neither well. When there are many abundance indices (an extreme example is the 2009 assessment of bocaccio (*Sebastes paucispinis*), where there were eight indices for adult fish and two for juveniles (Field et al., 2009)) it is more challenging to devise a small set of alternative models, each of which includes, and fits well to, a subset of the abundance indices. Another situation where this technique is useful is when there is uncertainty about whether an abundance index is representative of the stock being assessed. Francis (2011) pointed out that, rather than down-weighting this data set (a common response), we should evaluate the effect of the uncertainty by comparing outputs from a model which includes, and fits well to, the data, and an alternative model that excludes the data.

The preceding comments apply to the most common type of abundance data, arising from surveys or fishery CPUE, with a single datum for each time step. Abundance data from tag-recapture experiments are more complex, with each datum typically representing just those recaptures within a single length (or age) bin over a given time period. This data type was not considered by Francis (2011). Though it is not possible to weight these data outside the model (because the weighting depends on the extent of correlations induced by process error), they can be iteratively reweighted using an approach which is analogous to method TA1.8 (Francis, 2011) for compositions (see Appendix A).

4.2. Composition data

Composition data should not be weighted outside the model because it is not possible to quantify the process error, which often makes up a substantial proportion of the total error for these data (Table 2). Because the composition likelihoods used in most stock assessment programs are not self-weighting (Francis, 2014) [the two exceptions I am aware of are the logistic-normal likelihood in iSCAM (Martell 2011 [where is it labelled multivariate logistic]) and the recent introduction of the Dirichlet-multinomial in Stock Synthesis (Thorson et al., 2017), which I discuss below] it is usually necessary to weight composition data iteratively, rather than inside the model.

For iterative reweighting we must decide: (a) how to set the initial weights, (b) how to adjust the weights iteratively, and (c) how many times to iterate the reweighting. One method of setting initial weights is to make them correspond to the observation error, as estimated by bootstrap resampling of the raw data from which the compositions are constructed (e.g., Stewart and Hamel (2014)); Thorson (2014) proposed an alternative way. A simpler approach is to set initial multinomial sample sizes equal to the numbers of sets (or trips) sampled for each composition. In both methods the intention is to give greater weight to compositions in years when sampling was more intensive. A common approach to adjusting the initial weights after running the model is to use one of several algorithms based on two equations of McAllister and Ianelli (1997) (see Appendix B). As noted above, this approach is not statistically appropriate because it assumes composition residuals are uncorrelated; a better approach (described above) uses Pennington sample sizes, which typically produce much lower composition weights (Table 3) [lower weights make sense because the correlations reduce the amount of information in composition data]. It should be stressed that there is no 'correct' method of weighting compositions with a multinomial likelihood, because the likelihood itself is incorrect (since it does not allow the substantial correlations that are typical of these data); however, some weightings will be more statistically appropriate than others. There is no simple answer to the question as to how many times to iterate the reweighting. My experience is that it usually requires rather large changes (more than a factor of 2) in composition sample sizes to have an appreciable effect on key model outputs (e.g., the estimated spawning biomass trajectory). Thus I do not think it necessary to iterate until there is no appreciable change in the sample sizes; a better stopping criterion is no appreciable change in key outputs. Another reason to limit the

Table 2
Demonstration, using ten data sets from six New Zealand stock assessments, that a substantial percentage of the total error in composition data can be due to process error. For each composition data set (which was from either a survey or fishery) this percentage (in the final column) was calculated as $100(1-N_{total}/N_{obs})$, where N_{obs} and N_{total} are median Pennington sample sizes for the observation and total error, respectively (this calculation uses the relationship $V_{process} = V_{total} - V_{observation}$ and the fact that multinomial variances are inversely proportional to sample sizes; for details of the calculation of N_{obs} and N_{total} see the text associated with Table 1 of Francis, 2014).

Species	Assessment reference	Type	Source	N_{obs}	N_{total}	Percentage process error
Hoki (*Macruronus novaezelandiae*)	McKenzie (2013)	age	survey	116	83	28
		age	fishery	937	20	98
		age	fishery	261	69	74
Hake (*Merluccius australis*)	Horn (2013)	age	survey	89	24	73
		age	fishery	150	14	91
Ling (*Genypterus blacodes*)	McGregor (2015)	age	survey	323	152	53
		age	fishery	193	38	80
Smooth oreo (*Pseudocyttus maculatus*)	Doonan et al. (2009)	length	fishery	146	64	56
Paua (*Haliotis iris*)	Fu (2016)	length	fishery	327	65	80
	Fu (2014)	length	fishery	245	67	73

Table 3
Comparison of effective sample sizes calculated by method TA1.8 of Francis (2011), which produces Pennington sample sizes that allow for correlations, and two algorithms based on McAllister and Ianelli (1997), which do not (numbers in parentheses are approximate 95% confidence intervals for the TA1.8 sample sizes, as calculated by SSMethod.TA1.8 [Taylor et al., 2014]). The sample sizes were calculated for four age composition data sets from the 2006 assessment of southern hake (*Merluccius australis*) in Chile using a single iteration from a model run in which the input sample sizes were all 150.

		Effective sample sizes		
		McAllister and Ianelli (1997)		Francis (2011)
Composition data set	Number of compositions	Arithmetic[a]	Harmonic[a]	TA1.8
Trawl fishery	24	258	154	15 (10,33)
Commercial longline fishery	15	240	122	10 (6,32)
Artisanal longline fishery	12	557	210	25 (16,63)
Survey	6	338	294	69 (45,360)

[a] See Appendix B for details.

number of iterations is that Pennington sample sizes are often not very precisely estimated (e.g., see the confidence intervals in Table 3, which depend on the number of compositions in the data set).

The idea underlying Pennington sample sizes has also been used to develop iterative reweighting methods for two other data types: conditional age at length (Punt, 2017), and length-increment data from tag-recapture experiments (Punt et al., 2017). An important feature of the former method is that it allows for correlations both within and between length classes. [Both this method and TA1.8 are now available for users of Stock Synthesis in the R package r4ss (Taylor et al., 2014), in which they are called SSMethod.Cond.TA1.8 and SSMethod.TA1.8, respectively.]

The weighting adjustment in most commonly used iterative reweighting methods is multiplicative but I think a good case can be made for considering additive methods in some circumstances, though this will require a slightly different approach to that given in the additive methods of Francis (2011). All the reweighting methods in Table 3 use multiplicative adjustment, setting $N_{2y} = wN_{1y}$, where N_{1y} and N_{2y} are the input and adjusted samples sizes for year y, and w is an adjustment factor calculated from the model residuals. Francis (2011) described two methods with additive adjustment: in method TA1.9 (for a multinomial likelihood), $1/N_{2y} = 1/N_{1y} + 1/N_{adj}$; and in TA1.10 (for a lognormal likelihood), $c_{2y}^2 = c_{1y}^2 + c_{adj}^2$, where the cs are c.v.s (and the adjustment terms, N_{adj} and c_{adj}, are calculated from the model residuals). The difference between using additive and multiplicative adjustments can be substantial in terms of the resultant weights (Fig. 3). In the example in Fig. 3 the two reweighting methods assign about the same overall weight to the data set (i.e., the median samples sizes from the two methods are similar) because both use the idea of Pennington sample sizes. However, the strong between-year variation in the input sample sizes is preserved by the multiplicative

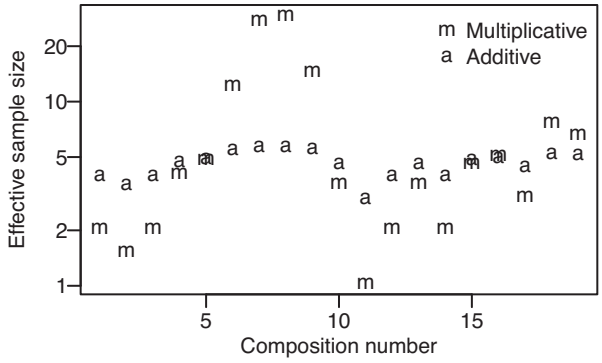

Fig. 3. Comparison of effective sample sizes for composition data reweighted using a multiplicative (TA1.8) or additive (TA1.9) method. The sample sizes were calculated for a set of 19 compositions for fishery discards from the initial base case in the 2013 assessment of rougheye rockfish (*Sebastes aurora*) (NMFS 2013). The median input sample size was 24; the TA1.8 adjustment factor, w, was 0.17; and the TA1.9 adjustment sample size, N_{adj}, was 5.9.

method, but much reduced with the additive method. The theoretical idea behind these additive methods is that to get the variance of the combination of two independent errors we simply add the variances of the individual errors (note that variances are proportional to $1/N$ for multinomial errors, and to squared c.v.s for lognormal errors). My intention in presenting methods TA1.9 and TA1.10 in Francis (2011) was that we would be adding observation and process errors, but I now realise that to properly implement this intention we need to make some changes these methods. For TA1.9 we should rewrite the above equation as $1/N_{y,total} = 1/N_{y,observation} + 1/N_{process}$, in which $N_{y,total}$, the sample size used for year y, is calculated from its two components: $N_{y,observation}$, which must be inferred in some way from the vari-

ability in the raw composition data (perhaps using methods like those of Stewart and Hamel, 2014 or Thorson, 2014); and N_{process}, which we must initially guess, and then iteratively update using the equation for TA1.9 in Francis (2011, table A1) [the $\tilde{N}_{jy}$ and N_j of that equation correspond to $N_{y,\text{observation}}$ and N_{process}, respectively, in the present notation]. A simple starting point for this iteration is to assume no process error, which means that the initial value of N_{process} is infinite, so that $N_{y,\text{total}} = N_{y,\text{observation}}$ in the initial model run of the iterative procedure (this is what I did, with a single iteration, to calculate the additive sample sizes for Fig. 3). Perhaps, given the results in Table 2, median$_y(N_{y,\text{observation}})$ would be a better initial value for N_{process}. [Ideally, the $N_{y,\text{total}}$ would be calculated inside the assessment program from user-provided values of $N_{y,\text{observation}}$ and N_{process} (as it is in CASAL) but it is not difficult for users to do this calculation outside the program and input just the $N_{y,\text{total}}$]. We need to make analogous changes to method TA1.10, for which the modified equation should be $c^2_{y,\text{total}} = c^2_{y,\text{observation}} + c^2_{\text{process}}$, and it is c_{process} that is iteratively updated. I think the additive methods are worth considering because they have some theoretical support (i.e., that variances of independent component errors are additive), whereas the above equation for the multiplicative methods is *ad hoc*. An important point to understand is that we have theoretical support for the additive methods only if we can argue that our values of $N_{y,\text{observation}}$ accurately represent the magnitude of observation error for our composition data. It does not make sense to use TA1.9 if we set $N_{y,\text{observation}}$ to some arbitrary value, like the number of trips or landings sampled in year y (unless, of course, previous analyses have shown that these values are representative of observation error). I would also say that the choice between additive and multiplicative methods is much less important than the decision to use composition weighting methods that allow for correlations.

Punt et al. (2014) suggested that the composition reweighting method TA1.8 "may fail when selectivity is modelled as a random walk because the model predictions of the mean age (length) will match the observed value very closely and the effective sample sizes could become very large". I agree that this could possibly happen but suggest that the problem here would not be with the data weighting; rather it would be in the parameterisation of a random walk structure that has allowed over-fitting of the composition data. The difficulty is in finding the right balance between providing sufficient flexibility in the selectivity parameterisation to mimic temporal changes in selectivity, but not so much flexibility that the model is fitting to noise in the composition data. As Nielsen and Berg (2014) note, approaches to addressing this difficulty are often rather *ad hoc*. One obvious indication of overfitting would be if the samples sizes after reweighting were larger than was consistent with observation error alone. A reasonable response would be to constrain the parameters of the random walk. Punt et al. (2014) also expressed concern that when the composition data are in conflict with other data method TA1.8 "could lead to the model converging to the fit which mimics the compositional data better." I think this would be a problem only if the conflict remained *after* the selectivity was made time-varying. Adding flexibility to the parameterisation of selectivity (e.g., using a random walk to allow variation with time) will tend to reduce, if not remove, data conflicts involving composition data, as Lee et al. (2014) showed. Both the potential problems discussed by Punt et al. (2014) are associated with the possibility that method TA1.8 could assign too much weight to composition data, which is surprising, given that this method typically assigns smaller weights to these data than other iterative-weighting schemes.

4.3. Catch data

Francis (2011) did not discuss the weighting of catch data because these data are usually treated as known, either exactly (i.e., without error), or almost exactly. In the former case the catch equation is solved to calculate, for each fishery f and time step t, the fishing pressure (expressed either as an exploitation rate or an instantaneous fishing mortality rate) so that model's expected catch, C^{exp}_{ft}, is equal to the observed catch, C^{obs}_{ft}. For the latter case (assuming catches are *almost* exactly true), the catches are treated as observations with very small errors (e.g., by assigning them a small c.v. – e.g., 0.05, or even 0.01), so the fishing pressures are estimated, like other parameters, rather than calculated. This latter approach should not be thought of as a form of data weighting. It is simply a computationally convenient way of ensuring that $C^{\text{exp}}_{ft} \approx C^{\text{obs}}_{ft}$, used because it is cumbersome to solve the Baranov catch equation with multiple fisheries at the same time step. Stock Synthesis offers a third, intermediate, approach referred to as "hybrid" fishing mortality (Methot and Wetzel 2013). The three approaches typically produce very similar assessments.

I believe these approaches will usually be sensible even when there is some, possibly substantial, doubt about the catch data. This is because there is rarely sufficient information to estimate the catches within the assessment model. A useful way to deal with uncertainty about catches is to construct alternative models which differ only in their catch data. For example, if the uncertainty relates to catches in the early years of the fishery, we may start with a base model which uses our best estimates of catches for these years, and bracket this with two alternative models in which the early catches are either increased or decreased by 30%, say. The difference in the outputs from the three models is then an expression of the effect of our uncertainty about the early catches.

5. Right-weighting vs down-weighting

Some recent papers have written about *down-weighting* composition data in a way that may be misleading. For example, Maunder and Punt (2013) interpreted Francis (2011) as recommending that age and length compositions be down-weighted to ensure an adequate fit to abundance data; and Lee et al. (2014) said that "Resolving model issues through the use of model process to reduce the misfit to the problematic data components rather than statistical down-weighting is preferable". The problem is that "down-weight" is a relative term: it simply means to reduce the weight currently applied to a data set. Thus a general recommendation about (or discussion of the merits of) down-weighting composition data does not make sense; down-weighting relative to what? The usage is understandable, because the recommendation by Francis (2011) to allow for correlations when weighting compositions will result in a down-weighting relative to many commonly-used weighting schemes. However, it would be better to speak of *right-weighting*, rather than down-weighting, where by the former term I simply mean applying a weighting procedure that is statistically appropriate (as described above). With regard to the preceding quote from Lee et al. (2014) I would agree that it is certainly desirable, where possible, to resolve issues (e.g., data conflict [discussed below]) by modifying assumptions concerning model processes. However, I think that data weighting should be seen *complementary* to such modifications, rather than an *alternative* to them.

6. Data conflict & misspecification

As noted above, data conflict is of importance here because changes in data weighting are likely to have greater effect on

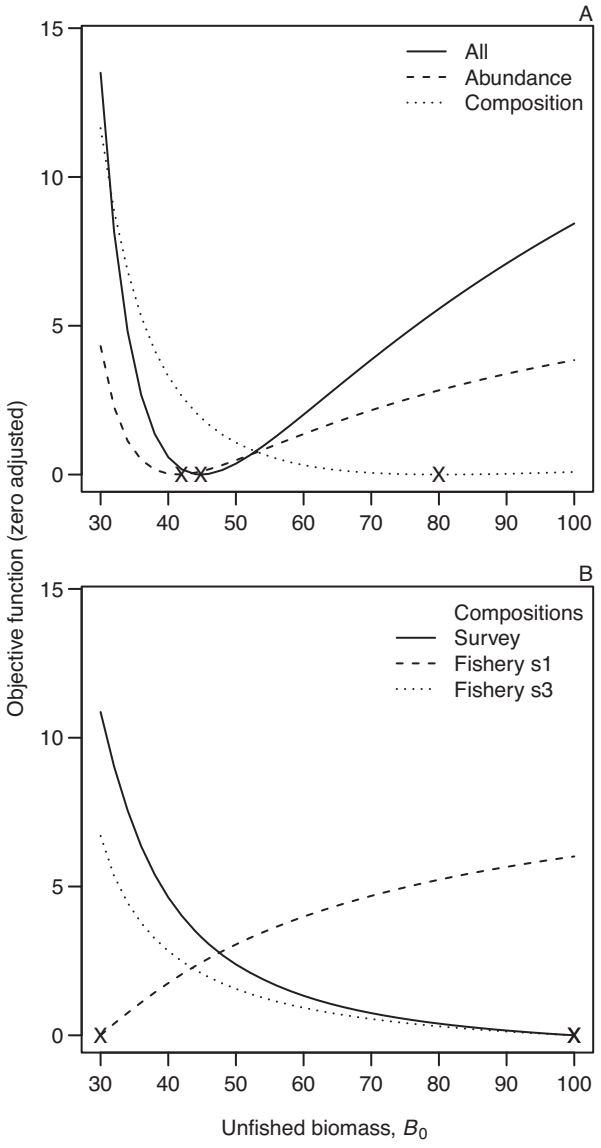

Fig. 4. A likelihood profile on unfished biomass, B_0, illustrating two types of conflict: A, between abundance and composition data; and B, within the composition data. Each plotted line shows how the objective function (i.e., the negative log-likelihood) associated with some subset of the data varies across the profile, with each line being zero adjusted (i.e., shifted vertically so that its minimum (marked "X") occurs at 0). The profile is from the 2010 assessment of New Zealand hake (Horn and Francis, 2010).

assessment outputs when there is conflict within the data. It is not uncommon in stock assessments to find evidence of conflict, both within and between data types. Amongst abundance data sets, such conflicts are usually immediately apparent from a plot of the data on a comparable scale (as in Fig. 2). Conflicts involving composition data are not usually so evident in data plots but can sometimes be seen in likelihood profiles. For example, Fig. 4A shows an apparent conflict between the abundance data, which are best fitted when $B_0 \approx 42\,000$ t, and the composition data, for which the best value of B_0 is about 80 000 t; Fig. 4B indicates a conflict *within* the composition data, with that from fishery s1 being most consistent with low values of B_0, while the other compositions favour high values.

The phrase "data conflict" is potentially misleading because it suggests that the source of the conflict is in the data, but in fact it more often arises from *model misspecification*, i.e., errors in model assumptions. For example, the conflict in Fig. 2 probably arises not because of any problem with the abundance data themselves, but because of a wrong assumption about those data: that both time series were proportional to biomass. Much of the apparent data conflict in Fig. 4 could well arise from misspecification in one or both of two processes – natural mortality and selectivity – which affect model inferences from composition data. In fitting to these data the model calculates an expected age composition to compare with each observed composition, but the expected compositions will be biased by errors in the model's representations of natural mortality and selectivity. Misspecification in these processes is inevitable. For example we should expected natural mortality to vary with time, space, and fish size, but there is rarely sufficient information to estimate this variation, so it is commonly ignored in model assumptions.

How should we deal with data conflict? First of all we should endeavour to remove (or at least reduce) it by *restructuring* our model (i.e, changing the model assumptions to reduce model misspecification), as advocated by Lee et al. (2014). A difficulty is that demonstrating the existence of conflict is much simpler than identifying its cause. Note that any restructuring of the model should be followed by a check to see whether the compositions are still right-weighted (because the statistically appropriate weighting may change when the model assumptions are modified). When restructuring fails to remove conflict, I think that our response should depend on the data type(s) involved. For conflicts amongst abundance data, the response I describe above (see first reference to Fig. 2) is consistent with the assertion of Punt and Hilborn (1997) that "The most fruitful approach to handling situations in which there are conflicting sources of information . . . is to conduct analyses for each source separately and present the results to the decision makers". This response is often not practicable when the conflict involves composition data (though it is sometimes a useful exercise to see what happens when a composition data set is dropped and the corresponding selectivity is fixed at a plausible value). Nor is it advisable when the conflict is seen in a profile on an absolute abundance parameter (as in Fig. 4) because the conflict may be illusory. Maunder and Piner (2015) noted that "relatively minor model misspecification (e.g. a too inflexible selectivity curve) can have a large impact on the information about absolute abundance [apparently] contained in the composition data". Here we should prioritise the abundance data, as described above, so the first thing to check is whether the abundance data are well fitted. Since these data were well fitted in the 2010 hake assessment, the conflict with the composition data in Fig. 4A was of less concern. Note also that an examination of the vertical scale in this plot shows that the conflict is not great. Although the minimum value of the composition objective function occurs at a value of B_0 (80 000 t) very far away from the assessment estimate of 45 000 t, the goodness of fit to the composition data is only very slightly different between these two value of B_0 (a difference of only about 2 in negative log-likelihood). Had the fit to the abundance data been poor, the next step would have been to check whether the composition data were right-weighted. If they were, then we should see whether we can achieve a satisfactory fit to the abundance data by down-weighting the compositions (or, possibly, up-weighting the abundance data). I would emphasise that this is a last resort; it is always preferable that data be right-weighted.

The interpretation of likelihood profiles is more complex than I thought when discussing the profile of Fig. 4 in Francis (2011) (in which fig. 1A shows the same profile). My statement "the estimate of B_0 could have taken any value between 42 000 t and 80 000 t, depending on the relative weights assigned to the biomass and composition data" was misguided. It was based on the simplistic assumption that the positions of the minima of the abundance and composition likelihoods (42 000 t and 80 000 t, respectively) would be unchanged if either data type were reweighted. A simple experiment showed that both these minima shifted (in the same

Table 4
Effect on the B_0 profile of Fig. 4A of reweighting either the composition or abundance data. The tabulated values are the changes (relative to the profile of Fig. 4A) in the values of B_0 (t) at which each objective function component has its minimum. The data sets were up- (or down-) weighted by halving (or doubling) the associated c.v.s (note that a lognormal likelihood was used for both data types in this assessment).

	Reweight compositions		Reweight abundance	
Objective function component	Up-weight	Down-weight	Up-weight	Down-weight
Abundance	+8 000	−2 000	−4 000	+2 000
Composition	+20 000	−26 000	−14 000	+20 000
All	+10 000	−1 000	−2 000	+3 000

direction) when the weighting on either data type was halved or doubled (Table 4). However, the change in the estimate of B_0 caused by each reweighting (shown in the bottom row of Table 4) is exactly as expected: i.e., the estimate decreased when either the abundance was up-weighted or the compositions were down-weighted, and vice versa.

7. Replacing the multinomial with the Dirichlet-multinomial

Because the multinomial likelihood does not allow substantial correlations, its use in stock assessment models is a clear misspecification, and we could improve our stock assessments if we replaced this likelihood by one that better reflected the error distributions of real composition data (Francis, 2014). Another advantage of finding a better likelihood is that it would allow us to simulate more realistic composition data. Almost all stock assessment-related simulation studies that involve generating composition data do so using the multinomial, and this must undermine any claim that conclusions from these studies would be applicable with real composition data.

The Dirichlet-multinomial, which has recently been introduced into Stock Synthesis as an alternative to the multinomial (Thorson et al., 2017), has some clear advantages. It is self-weighting [but note that, of its two weighting parameters – one Dirichlet, the other multinomial – only the former is estimable] so there is no need for iterative reweighting, which greatly simplifies the assessment process and also ensures that the uncertainty associated with the composition weighting is included in any measures of stock assessment uncertainty (e.g., standard errors or confidence intervals for assessment outputs). It also has the advantage of allowing zero proportions, which the logistic-normal likelihood, proposed by Francis (2014) as a better replacement for the multinomial, does not. Thorson et al., (2017) acknowledge that, unlike the logistic-normal, the Dirichlet-multinomial cannot hope to mimic the strong between-bin correlations that are characteristic of compositions. However, they hypothesise that compositions are also likely to be correlated among years and fleets, and suggest that the best strategy for dealing with such a complicated correlation structure is via "mixed-effects models" (presumably something akin to the state-space models discussed below). Both the hypothesis and suggested strategy certainly deserve consideration. However, it remains to be seen whether it is satisfactory to deal with an observation-error correlation in compositions by using random effects to model a time-varying process that might induce similar correlations.

The Dirichlet-multinomial suffers from two other weaknesses. Thorson et al. (2017) showed that, when used in an assessment, this likelihood gives results similar to those obtained with the multinomial likelihood subject to iterative reweighting following McAllister and Ianelli (1997). This does not seem a strong recommendation for the new likelihood in the context of conventional models (without random effects), because the evidence is that McAllister and Ianelli methods overweight composition data in such models (see, e.g., Table 3). It remains to be seen whether such weighting will be satisfactory in models with random effects (discussed in the next section). Another weakness is that the degree of overdispersion is the same for every (age or length) bin with the Dirichlet-multinomial, but has been shown to be strongly bin-dependent in real composition data (Fig. 5 of Hrafnkelsson and Stefánsson, 2004; Tables 2 and 3 of Miller and Skalski, 2006).

8. Other approaches in stock assessment

As noted above, the "conventional" stock assessment models assumed by Francis (2011) in his approach to data weighting are widespread but not universal. There are two departures from this type of model that deserve comment here.

8.1. State-space models

The first departure is the use of what have come to be called *state-space models* (e.g., Millar and Meyer, 2000; Linton and Bence, 2008; Nielsen and Berg, 2014; also see de Valpine, 2002 for an excellent introduction to the associated theory, but note that this field is rapidly evolving because of substantial technological advances since 2002). In discussing these models I will use the term *process variation* to designate what is usually called "process error" in the state-space literature, because I have used the latter term in a different sense (see definition above). Process variation refers to year-to-year changes in processes (or quantities) such as recruitment, natural mortality, fish growth, fishery selectivity, proportion mature, etc. Conventional models typically ignore all process variation except for recruitment: yearly values of recruitment (or, equivalently, log-space deviates from an expected value) are treated as fixed effects, to be estimated, like other parameters, and the parameter quantifying their variability (usually denoted σ_R) is fixed, rather than estimated. In state-space models, variation may be modelled in one or more processes, and the associated deviates are usually treated as random effects (to be integrated over, rather than estimated) whose variance parameter is usually estimated [for an explanation of why the annual values should be integrated over see the discussion by de Valpine (2002) of the difference between what he calls "true" and "errors-in-variables" likelihoods].

Including process variation in a model affects data weighting because it reduces process error. Recall that the weight assigned to each observation needs to allow for both observation and process errors. Some of the process error in a conventional model will be caused by the fact that much process variation has been ignored (i.e., processes that vary from year to year have been treated as time-invariant). Thus, when we add process variation to a model we are likely to reduce the process error associated with an observation (and thus allow greater weight to be applied to the observation). However, it would be wrong to assume that we can remove *all* process error in a state-space stock assessment model because not all process error is associated with process variation: some is due to other factors, such as errors in either fixed parameters (e.g., using the wrong parameter values for natural mortality or growth) or mathematical forms (e.g., using the wrong equation for mean length at age or a fishery selectivity, or the multinomial likelihood

for compositions that include substantial correlations). Moreover, it is not feasible to model all of the many model processes that vary from year to year. Thus, what is commonly labelled "observation error" in the state-space literature will always be a mixture of what I have called observation and process errors.

I am not aware of any need to modify the data-weighting approach of Francis (2011) for use with state-space models. However, the weighting problem becomes more complex with these models. In a conventional model we may think of the data-weighting problem as being one of partitioning the total error amongst the various data sets. Decreasing the weight given to a data set is equivalent to increasing its share of the total error. In a state-space model, we partition the total error amongst both data sets *and* time-varying processes. Increasing the variance of a time-varying process is the same as increasing its share of the total error. So, as well as asking (i) "do we have the right balance of weights amongst the different data sets?", with state-space models we need also ask (ii) "do we have the right balance between data and time-varying processes?", and, when there are multiple time-varying processes, (iii) "do we have the right balance of weights (i.e., process variances) amongst these processes?". Linton and Bence (2008) used a simulation experiment to show that (ii) was usually difficult to answer in their state-space statistical catch-at-age model (in their terminology we would say that they could not usually produce good estimates of both process and observation process error variances [NB they did obtain good estimates in the case when an informative prior was used for the process error variance, but noted that "it is unlikely that a stock assessment analyst would have the necessary data to set such an informative prior"]). It is well understood that getting (i) wrong can strongly affect the outcome of a stock assessment. What is less clear is how influential errors in (ii) and (iii) are.

8.2. Use of frequencies rather than compositions

Conventional models differ from some earlier assessment models (particularly Virtual Population Analysis [VPA] and its descendants [e.g., Pope, 1972; Shepherd, 1999]) in structuring age-related observations of fishery and survey catches as compositions, rather than frequencies (i.e., as proportions, rather than numbers). Much of the above discussion of data weighting assumes the use of compositions, and so is not relevant to some recent state-space models (e.g., Millar and Meyer, 2000; Nielsen and Berg, 2014) that have reverted to the use of frequencies, rather than compositions.

The use of age frequency observations seems to me to make much more difficult both the broader task of stock assessment, and the more specific task of data weighting. Survey data contain two fundamentally different types of information: abundance (the total quantity of fish observed – by number or weight), and age structure (how that total quantity breaks down by age). I have argued above that we should prioritise the former type, particularly when, as is not uncommon, there is conflict between the two types. This is straightforward when the two types of information are presented in separate data sets (as in conventional models), but not when they are combined in age frequencies. For fishery data, separating the two types of information allows us to easily distinguish between (and thus weight differently) the total catch for each year (which is often known relatively precisely) and the age structure of the catch (which is much less precisely known).

When fishery age frequencies (rather than compositions) are used in stock assessment models they are often aggregated across fisheries. This practice is problematic because it withholds from the model two potentially useful types of information about between-fishery heterogeneity. First, there may be, for logistical or other reasons, quite substantial between-fishery differences in sampling intensity, which means that different weights should be applied to data from different fisheries. This is not possible with aggregated data. Second, a model with aggregated age frequencies lacks information about any year-to-year changes in the contribution that each fishery makes to the total catch. These changes in contribution could be useful in tracking year-to-year changes in the aggregate fishery selectivity. For example, suppose a fishery which typically catches old fish, gradually contributes a bigger and bigger proportion of the total catch. Then we would expect the aggregate fishery selectivity to gradually shift to the right, but a model with aggregated age frequencies would lack an important piece of information (the change in contribution from the fishery) to inform the estimation of such a shift.

9. Concluding comments

In recent years there seems to have been an increasing consciousness of data weighting amongst the stock assessment community: data-weighting decisions are now more often discussed and explicit in assessment documents (where once they were tacit and implicit), and new methods of data weighting have been developed. This is a positive change, which must be improving the quality of stock assessments. Another worthwhile development is a greater attention towards removing some model misspecification, with the aim of lessening the sensitivity of assessments to data weighting by reducing (apparent) data conflicts (particularly those involving composition data). Nothing in these developments suggests a need to modify the three guiding principles of data weighting given by Francis (2011): (i) do not let other data stop the model from fitting abundance data well; (ii) when weighting age or length composition data, allow for correlations; and (iii) do not down-weight abundance data because they may be unrepresentative. Here, I have tried to extend the message of the earlier paper by (a) providing increased emphasis or elaboration of some points, and (b) to discuss a few topics that have arisen more recently (see Abstract for a summary).

There are two particular areas of research whose development I follow with interest because I think they might have a strong effect on data weighting. The more general, and active, area is the use of random effects to model process variation in stock assessment models and thus reduce process error, model misspecification, and (potentially) data conflict. Heavy computational requirements slowed progress in this area until the recent development of TMB (Kristensen et al., 2016). [To those who have, as I once did, doubts about whether random effects are really significantly different from fixed effects I offer the growth modelling example of Francis et al. (2016), where the former worked but the latter did not; for a theoretical justification see the text associated with the last reference above to de Valpine (2002)]. The other, more specific, research area of interest is the difficult search for a convincing replacement for the multinomial as a likelihood for composition data. As I say above, I suspect we can do better than the Dirichlet-multinomial proposed by Thorson et al. (2017), but the logistic-normal that I advocated (Francis, 2014) is not without problems. I am aware of promising research in this area and believe that weighting composition data inside the stock assessment model, with a likelihood which more accurately represents the error distribution of these data, will represent a great advance over methods like TA1.8, which, though the best we can currently do using the multinomial likelihood, are simply *ad hoc* solutions to the problem of using the wrong likelihood.

Acknowledgements

I am grateful to the Center for the Advancement of Population Assessment Methodology for the invitation to present an earlier version of this paper at their workshop (www.CAPAMresearch.org/

data-weighting/workshop); to Cristian Canales of the Instituto de Fomento Pesquero, Valparaíso, for permission to use data associated with Fig. 1 and Tables 1 and 3; to Ian Doonan, Peter Horn, Dan Fu and Andy McKenzie of NIWA, New Zealand, for providing the data I analysed for Table 2; and to André Punt and Jim Thorson for useful reviews of a draft of this paper.

Appendix A. Data weighting for tag-recapture abundance data

In this Appendix I describe the rationale behind a recently-proposed method for weighting tag-recapture data for stock assessments using CASAL. This method applies to situations in which the data include information about the *tag rate* (the proportion of captured fish that are tagged), and so are informative about population abundance. Without tag rate information the data are still informative about growth, but the present weighting method is not applicable (but see Punt et al., 2017).

The tag-recapture data are presented to CASAL in a series of subsets, each of which is associated with catches over a given time period in a given area. Within each subset the fish are binned (usually by length, but conceivably by age) and the data presented for the ith bin in the jth subset are

n_{ij}, the number of captured fish examined for tags, and

m_{ij}, the number of examined fish that are found to have tags

The default assumption is that the m_{ij} are independent (between bins and between subsets) and $m_{ij} \sim \text{RobustBinomial}(n_{ij}, p_{ij})$, where p_{ij} is the model's expected tag rate (see p. 77 of Bull et al., 2012 for the form of the robustification). Then $e_{ij} = n_{ij}p_{ij}$ is the expected number of tagged fish in the ith bin of the jth subset, and we define $m_j = \Sigma_i m_{ij}$ and $e_j = \Sigma_i e_{ij}$.

There is evidence that some (probably most) tag-recapture data sets are over-dispersed. That is, they are more variable than would be expected from the above assumptions, and thus should be down-weighted. To investigate this possibility we can construct residuals, $r_j = (m_j - e_j)e_j^{-0.5}$. Since the n_{ij} are typically large, the m_{ij} should be approximately Poisson(e_{ij}), and the additive property of independent Poisson distributions means that m_j should be approximately Poisson(e_j), and so have mean and variance equal to e_j. Thus we should expect that $\text{Var}_j(r_j)$, which we denote by w, should be approximately 1. With some data sets w is substantially greater than 1, i.e., the data are over-dispersed. This is presumably because, rather than being independent between bins, the m_{ij} are positively correlated.

CASAL provides a dispersion parameter, d, as an informal means of allowing for over-dispersion, with the (robust binomial) log-likelihood of each subset of tag-recapture data being divided by d. The default is $d = 1$; setting $d > 1$ implies over-dispersion and down-weights the data. The proposed tag-recapture weighting method, in which $d \approx w$, is motivated by analogy with a normal likelihood. If our observations, m_{ij}, were normally distributed with variance σ^2, then dividing our log-likelihood by a dispersion parameter, d, would have exactly the same effect as multiplying σ^2 by d. Thus setting $d = w$ for our tag-recapture data is like multiplying the assumed binomial variance of each m_{ij} [i.e., $n_{ij}p_{ij}(1 - p_{ij})$] by w, which changes the expected variance of the r_j from 1 to w, as observed, and makes the weighting of this data statistically appropriate. This weighting method was first applied in the 2013 assessment of the New Zealand SNA 1, where the change in weighting from $d = 1$ to $d = 2.7$ had a moderate effect on estimated biomass trajectories (see Fig. 10 of Francis and McKenzie, 2015).

A1.1 similarities with TA1.8

It is instructive to note several similarities between this weighting method and method TA1.8, which Francis (2011) devised to weight (length or age) composition data. Both methods involve constructing a residual, r_j, for each of a series of subsets of a data set, and then adjusting the data weighting so that the expected variance of the r_j is close to the observed variance. For the composition data, each subset is a single composition vector (i.e., the proportions at age or length in the catch from a given fishery (or survey) in a given time period) and the r_j are standardised residuals of mean length or mean age. Both weighting methods use w (=$\text{Var}_j(r_j)$): for the tag-recapture data we set $d \approx w$; for method TA1.8 we divide the initial effective sample sizes by w. Note that both methods rely on the associated data set consisting of sufficiently many subsets to obtain a reasonably reliable estimate of w. Further, both weighting methods allow for correlations within the data subsets that are not allowed for in the data likelihood.

Another similarity is that neither of the likelihoods for these weighting methods is self-weighting. That is, we can not make our weighting parameter, w, an estimable model parameter, rather than calculating it (outside the model) as $\text{Var}_j(r_j)$. Francis (2014) pointed out that the multinomial distribution is not self-weighting for composition data because, as it is used in stock assessment models it is *improper* (i.e., its integral, over all permissible values of the observations, rather than being 1, is a function of the parameters of the distribution). The same is true of the binomial distribution in the tag-recapture likelihood. In CASAL, this is made improper by two factors: the robustification and the use of $d \neq 1$. The former factor is probably minor, but the latter is not, as I shall explain. With the usual (non-robustified) binomial, the probability that we would observe m tagged fish in the ith bin of the jth data subset is given by $P_{ij}(m) = p_{ij}^m \left(1 - p_{ij}\right)^{n_{ij}-m} n!/\left[m!\left(n_{ij} - m\right)!\right]$. Standard theory shows that if we sum this probability over all permissible values of m we get 1 (i.e., $\sum_{m=0}^{n_{ij}} P_{ij}(m) = 1$). However, introducing a dispersion parameter, d, is equivalent to replacing $P_{ij}(m)$ by $P_{ij}(m)^{1/d}$, and this makes our distribution improper because $\sum_{m=0}^{n_{ij}} P_{ij}(m)^{1/d}$ is no longer equal to 1, and in fact depends on the distribution parameters, n_{ij}, p_{ij}, and d. If we try to estimate d we find that our estimate tends to infinity, and this is simply because $P_{ij}\left(m_{ij}\right)^{1/d}$, the likelihood of our observation m_{ij}, is monotone increasing as a function of d.

Appendix B Some iterative reweighting algorithms based on McAllister and Ianelli (1997)

McAllister and Ianelli (1997, equations (2.5), (2.6) in Appendix B) described an iterative reweighting algorithm for composition data with a multinomial likelihood. Their first equation calculates, from the assessment model output, an effective sample size, $N_{\text{eff},ij}$, for the jth composition in the ith data set [NB the sample sizes labelled "effN" in Stock Synthesis output are the $N_{\text{eff},ij}$]; and the second equation calculates the new sample size for all compositions in the ith data set as $N_{\text{new},i} = \text{mean}_j(N_{\text{eff},ij})$.

This algorithm seemed reasonable at a time when (as McAllister and Ianelli (1997) say) it was "common practice" to use the same input sample size for all compositions in a data set (i.e., $N_{\text{input},ij} = N_{\text{input},i}$ for all j). However, it is not reasonable now that it is common to recognise year-to-year variations in sampling intensity by assigning a different input sample size to each composition. To allow for this change of practice a common variant of the original algorithm retains the first equation, but replaces the second by $N_{\text{new},ij} = N_{\text{input},ij}\ \text{mean}_j(N_{\text{eff},ij}/N_{\text{input},ij})$. Both the original and variant algorithms are strongly affected by outliers in the form of the extremely high values of $N_{\text{eff},ij}$ that occur when, by chance, one or more observed compositions happens to lie very close to its expected value. In recent years a second vari-

ant algorithm has become more common in which the arithmetic mean in this last equation is replaced by the harmonic mean, which is less sensitive to outliers [the harmonic mean of x_1, x_2, ..., x_n is $n/\sum_i (1/x_i)$]. These two variants of the McAllister and Ianelli (1997) reweighting algorithms are denoted "Arithmetic" and "Harmonic" in Table 3 above. A third variant uses $N_{\text{new},ij} = N_{\text{input},ij}$ harmonic-mean$_j(N_{\text{eff},ij})$/mean$(N_{\text{input},ij})$.

References

Akaike, A., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Anonymous, 2016. Assessment Model for Alaska Description of GUI and Instructions. https://github.com/NMFS-toolbox/AMAK, (accessed 21.04.16.).

Brodziak, J. 2005. Technical Description of STATCAM Version 1.2. National Marine Fisheries Service, Woods Hole, Massachusetts. http://nft.nefsc.noaa.gov/Download.html, (accessed 10.12.15.).

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H., Bian, R., Fu, D., 2012. CASAL (C + + algorithmic stock assessment laboratory): CASAL User Manual v2.30-2012/03/21. NIWA Technical Report 135, 280p, Available from http://www.niwa.co.nz/fisheries/tools-resources/casal.

Clark, W.G., Hare, S.R., 2006. Assessment and management of Pacific halibut: data, methods, and policy. Scientific Report 83. International Pacific Halibut Commission, Seattle, Wash.

Craig, K., 2012. The Beaufort Assessment Model (BAM) with Application to Cobia: Mathematical Description, Implementation Details, and Computer Code. SEDAR28-RW01. SEDAR, North Charleston, SC, pp. 37, http://www.sefsc.noaa.gov/sedar/Sedar Documents.jsp?WorkshopNum=28&Folder Type=Review 21 October 2013.

Doonan, I.J., Coburn, R.P., McMillan, P.J. 2009. Assessment of OEO 3A smooth oreo for 2008–09. New Zealand Fisheries Assessment Report 2009/47, 40 pp.

Field, J.C., Dick, E.J., Pearson, D., MacCall, A.D., 2009. Status of bocaccio, *Sebastes paucispinis*, in 2009. http://www.pcouncil.org/groundfish/stock-assessments/by-species/bocaccio-rockfish/, (accessed 10.12.15.).

Fournier, D.A., Archibald, C.P., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39 (8), 1195–1207.

Fournier, D.A., Hampton, J., Sibert, J.R., 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat Sci. 55 (9), 2105–2116.

Francis, R.I.C.C., Aires-da-Silva, A.M., Maunder, M.N., Schaefer, K.M., Fuller, D.W., 2016. Estimating fish growth for stock assessments using both age-length and tagging-increment data. Fish. Res. 180, 113–118.

Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138.

Francis, R.I.C.C., 2014. Replacing the multinomial in stock assessment models: a first step. Fish. Res. 151, 70–84.

Francis, R.I.C.C., McKenzie, J.R., 2015. Assessment of the SNA 1 stocks in 2013. New Zealand Fisheries Assessment Report 2015/76. 94 p.

Fu, D., 2014. The 2013 stock assessment of paua (*Haliotis iris*) for PAU 5B. New Zealand Fisheries Assessment Report, 2014/45. 51 p.

Fu, D., draft. The 2015 stock assessment of paua (*Haliotis iris*) for PAU 7. New Zealand Fisheries Assessment Report. 20xx/xx.

Horn, P.L., Francis, R.I.C.C. 2010. Stock assessment of hake (Merluccius australis) on the Chatham Rise for the 2009–10 fishing year. New Zealand Fisheries Assessment Report 2010/14.

Horn, P.L., 2013. Stock assessment of hake (*Merluccius australis*) on the Chatham Rise (HAK 4) and off the west coast of South Island (HAK 7) for the 2012–13 fishing year, New Zealand Fisheries Assessment Report 2013/31, 58 pp.

Hrafnkelsson, B., Stefánsson, G., 2004. A model for categorical length data from groundfish surveys. Can. J. Fish. Aquat. Sci. 61 (7), 1135–1142.

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H.J., Bell, B., 2016. TMB: automatic differentiation and laplace approximation. J. Stat. Software 70 (5).

Lee, H.-H., Piner, K.R., Methot, R.D., Maunder, M.N., 2014. Use of likelihood profiling over a global scaling parameter to structure the population dynamics model: an example using blue marlin in the Pacific Ocean. Fish. Res. 158, 138–146.

Legault, C.M., Restrepo, V.R., 1999. A flexible forward age-structured assessment program. ICCAT Coll. Vol. Sci. Pap. 49 (2), 246–253.

Linton, B.C., Bence, J.R., 2008. Evaluating methods for estimating process and observation error variances in statistical catch-at-age analysis. Fish. Res. 94 (1), 26–35.

Martell, S., 2011. iSCAM Users Guide Version 1.0, Available from https://sites.google.com/site/iscamproject/, (accessed 21.10.13.).

Maunder, M.N., Piner, K.R., 2015. Contemporary fisheries stock assessment: many issues still remain. ICES J. Mar. Sci. 72 (1), 7–18.

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74.

Maunder, M.N., 2003. Paradigm shifts in fisheries stock assessment: from integrated analysis to Bayesian analysis and back again. Nat. Resour. Model. 16 (4), 465–475.

McAllister, M.K., Ianelli, J.N., 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54 (2), 284–300.

McGregor, V., 2015. Stock assessment of ling (Genypterus blacodes) on the Chatham Rise (LIN 3&4) for the 2014–15 fishing year. New Zealand Fisheries Assessment Report 2015/82, 50 pp.

McKenzie, A., 2013. Assessment of hoki (*Macruronus novaezelandiae*) in 2012. New Zealand Fisheries Assessment Report 2013/27, pp. 65.

Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.

Millar, R.B., Meyer, R., 2000. Bayesian state-space modeling of age-structured data: fitting a model is just the beginning. Can. J. Fish. Aquat. Sci. 57 (1), 43–50.

Miller, T.J., Skalski, J.R., 2006. Integrating design- and model-based inference to estimate length and age composition in North Pacific longline catches. Can. J. Fish. Aquat. Sci. 63 (5), 1092–1114.

NMFS, 2011. Pacific Sardine STAR Panel Meeting Report. National Marine Fisheries Service, Silver Springs, Maryland. http://www.pcouncil.org/wp-content/uploads/F2b_ATT5_SARDINE_STAR_NOV2011BB.pdf, (accessed 01.12.15.).

NMFS, 2013. Stock Assessment Review (STAR) Panel Report for Rougheye (and Blackspotted) Rockfish. National Marine Fisheries Service, Silver Springs, Maryland. http://www.pcouncil.org/wp-content/uploads/Rougheye_and_Blackspotted_2013_STAR.pdf, (accessed 27.11.15.).

Newman, K.B., Buckland, S.T., Morgan, B.J.T., King, R., Borchers, D.L., Cole, D.J., Besbeas, P., Gimenez, O., Thomas, L., 2014. Modelling Population Dynamics. Springer, New York.

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. 158, 96–101.

Pennington, M., Vølstad, J.H., 1994. Assessing the effect of intrahaul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics 50 (3), 725–732.

Pope, J.G., 1972. An investigation of the accuracy of virtual population analysis using cohort analysis. Res. Bull. Int. Comm. N.W. Atlantic Fish. 9, 65–74.

Punt, A.E., Hilborn, R., 1997. Fisheries stock assessment and decision analysis: the Bayesian approach. Rev. Fish Biol. Fish. 7, 35–63.

Punt, A.E., Kennedy, R.B., 1997. Population modelling of Tasmanian rock lobster, Jasus edwardsii, resources. Mar. Freshwater Res. 48, 967–980.

Punt, A.E., Hurtado-Ferro, F., Whitten, A.R., 2014. Model selection for selectivity in fisheries stock assessments. Fish. Res. 158, 124–134.

Punt, A.E., Deng, R.A., Siddeek, M.S.M., Buckworth, R.C., Vanek, V., 2017. Data weighting for tagging data in integrated size-structured models. Fish. Res. 192, 94–102.

Punt, A.E., 2017. Some insights into data weighting in integrated stock assessments. Fish. Res. 192, 52–65.

Schaub, M., Abadi, F., 2011. Integrated population models: a novel analysis framework for deeper insights into population dynamics. J. Ornithol. 152 (1), 227–237.

Schnute, J.T., Hilborn, R., 1993. Analysis of contradictory data sources in fish stock assessment. Can. J. Fish. Aquat. Sci. 50 (9), 1916–1923.

Sharma, R., Langley, A., Herrera, M., Geehan, J., Hyun, S.-Y., 2014. Investigating the influence of length-frequency data on the stock assessment of Indian ocean bigeye tuna. Fish. Res. 158, 50–62.

Shepherd, J.G., 1999. Extended survivor analysis: an improved method for the analysis of catch-at-age data and abundance indices. ICES J. Mar. Sci. 56 (5), 584–591.

Stewart, I.J., Hamel, O.S., 2014. Bootstrapping of sample sizes for length- or age-composition data used in stock assessments. Can. J. Fish. Aquat. Sci. 71, 581–588.

Taylor, I., Stewart, I., Hicks, A., Garrison, T., Punt, A., Wallace, J., Wetzel, C., Thorson, J., Takeuchi, Y., Monnahan, C. 2014. Package r4ss. https://github.com/r4ss.

Thorson, J.T., Johnson, K.F., Methot, R.D., Taylor, I.G., 2017. Model-based estimates of effective sample size in Stock Synthesis using the Dirichlet-multinomial distribution. Fish. Res. 192, 84–93.

Thorson, J.T., 2014. Standardizing compositional data for stock assessment. ICES J. Mar. Sci. 71, 1117–1128.

de Valpine, P., 2002. Review of methods for fitting time-series models with process and observation error and likelihood calculations for nonlinear, non-Gaussian state-space models. Bull. Mar. Sci. 70, 455–471.