Data weighting in statistical fisheries stock assessment models

Chris Francis

SEDAR56-RD11

9 Jan 2018

PERSPECTIVE / PERSPECTIVE

Data weighting in statistical fisheries stock assessment models

R.I.C. Chris Francis

Abstract: The conclusions drawn from fisheries stock assessment models can depend strongly on the relative weights assigned to different data sets. However, there is no consensus amongst practitioners as to the best approach to data weighting. From a discussion of some key questions concerning data weighting in stock assessment models, I draw three guiding principles: (*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. I propose an approach to data weighting based on these principles. Two factors that complicate this approach are that some decisions are inevitably subjective (which underlines the need for expert knowledge in stock assessment), and some technical problems are unresolved.

Résumé : Les conclusions tirées des modèles d'évaluation des stocks des pêches dépendent fortement des poids relatifs attribués aux différents ensembles de données. Il n'y a cependant aucun consensus entre les utilisateurs sur la meilleure manière de pondérer les données. Une discussion de quelques-unes des questions principales sur la pondération des données dans les modèles d'évaluation des stocks me permettent de tirer trois principes directeurs: (*i*) ne pas permettre à d'autres données d'empêcher le modèle de bien s'ajuster aux données d'abondance, (*ii*) en pondérant des données de composition en âge et en longueur, tenir compte des corrélations et (*iii*) ne pas assigner une pondération inférieure à des données d'abondance parce qu'elles ne seraient pas représentatives — la méthodologie de pondération des données proposée ici est basée sur ces principes. Deux facteurs compliquent cette approche: certaines décisions sont inévitablement subjectives (ce qui souligne la nécessité d'obtenir des opinions d'experts dans l'évaluation des stocks) et certains problèmes techniques restent irrésolus.

[Traduit par la Rédaction]

Introduction

Most model-based fisheries stock assessments use multiple data sets. Thus one decision that must be made, explicitly or implicitly, during an assessment is how much weight to assign to each data set. This can be important because the estimated status of the stock being assessed will depend, possibly strongly, on these weights. Although there is wide agreement that data weighting is important (National Research Council 1998; Breen et al. 2003; Hulson et al. 2008), there seems to be no consensus amongst stock assessment practitioners as to how it should be addressed. Reports on individual stock assessments (mostly unpublished) often devote considerable space to data weighting, but there is wide variation amongst these reports in the methods used to weight different data sets. In this paper I discuss the main issues involved in data weighting in a stock assessment setting, and propose an approach to this problem. The topic is not straightforward because, as I will try to show, there is no satisfactory objective method of data weighting, and there are a number of unresolved questions concerning technical details of weighting schemes. Nevertheless, I think it is possible to develop some useful guiding principles, which are based as much on pragmatism and experience as on statistical theory. Before describing these principles, and a proposed approach to the problem, I will define some terminology and discuss some key data-weighting questions. The more technical details are given in two appendices.

Notation and terminology

In understanding the notation and terminology used below (Table 1), it is useful to view the stock assessment task at a series of levels of increasing complexity. At the simplest

Received 13 July 2010. Accepted 28 February 2011. Published at www.nrcresearchpress.com/cjfas on 13 July 2011.
J21919

Paper handled by Associate Editor Ray Hilborn.

R.I.C.C. Francis. National Institute of Water & Atmospheric Research, Private Bag 14901, Wellington, New Zealand.

Corresponding author: R.I.C.C. Francis (e-mail: c.francis@niwa.co.nz).

Table 1. Notation used in discussing and defining data weighting.

Type	Symbol	Description
Vector	$\boldsymbol{O}$	Vector of all data (observations) that are fitted to in a stock assessment model
	$\boldsymbol{P}$	Vector of all parameters being estimated in a stock assessment model
Function	L	Either the objective function in a stock assessment model or, if subscripted, the component of that objective function associated with an individual data point (in the latter case, L represents a negative log-likelihood)
Variable	O	An individual data point (observation) (e.g., a survey biomass estimate, or the proportion of the catch from a given year that is of a given age)
	E	The model's expected value for an individual data point
	T	The true (real world) value for an individual data point
	m	The number of individual data points in a data set
	m_{year}	The number of years covered by an abundance or composition data set
	m_{bin}	The number of bins in a composition data set
Subscript	i	Indexes the abundance data sets used in a stock assessment model
	j	Indexes the composition data sets used in a stock assessment model
	y	Indexes the years within an abundance or composition data set
	b	Indexes the bins within a composition data set (these bins may be defined by age or length, possibly in combination with sex)
Weighting parameters		
	λ, w	Simple weights
	c	Coefficient of variation (CV) used in weighting data
	σ	Standard deviation (SD) used in weighting data
	N	Multinomial sample size used in weighting data

level, this task involves fitting our data (contained in the vector $\boldsymbol{O}$) to a model by estimating the "best" values for those model parameters (in the vector $\boldsymbol{P}$) whose values are not known, where the best values are defined as being those that minimize our objective function $\mathrm{L}(\boldsymbol{O}, \boldsymbol{P})$. There are likely to be three main types of parameters in $\boldsymbol{P}$: (*i*) those that determine the historical dynamics of the fish population (e.g., the biomass of the unfished population, the recruitment in each year, and parameters for natural mortality and growth); (*ii*) those that describe the action of the fishery (e.g., the fishing mortality in each year, and the age at which fish are 50% selected in a particular fishery); and (*iii*) those from the error distributions for the data (e.g., a standard deviation (SD), or coefficient of variation (CV), for an abundance index). Our focus will be on type *iii* parameters, which affects data weighting.

Looking at the data in more detail, we will, for the most part, consider only two types of data: (*i*) observations (or estimates) of abundance (in biomass or numbers), and (*ii*) observations of the length or age composition of the catch from a fishery or a survey. These are the dominant, and often only, types of data used in stock assessments, and thus the most important in discussions of data weighting (other data types are discussed briefly below). We will think of the data vector as being made up of a number of data sets, each of which consists of a collection of individual data points, which are written as O_{iy}, for abundance data, or O_{jby}, for composition data. For example, one abundance data set might consist of estimates of (relative or absolute) abundance from a series of trawl surveys, with each individual data point being an abundance estimate for one of the survey years. In a typical composition data set, the individual data point is an estimate of the proportion of the catch for a given year that lies in a given length or age bin (bins may be defined by combinations of sex and age or length). If there are several fisheries, there may be one composition data set for each fishery; there may also be composition data sets associated with surveys.

Now, looking more closely at the objective function, this may be written as follows:

$$\mathrm{L}(\boldsymbol{O}, \boldsymbol{P}) = \sum\nolimits_{iy} \mathrm{L}_{iy}(O_{iy}, \boldsymbol{P}) + \sum\nolimits_{jby} \mathrm{L}_{jby}(O_{jby}, \boldsymbol{P}) + \text{other terms} \quad (1)$$

I will consider only those stock assessments that use "statistical" models, like those typically developed using computer packages such as Stock Synthesis (NOAA 2011), CASAL (Bull et al. 2008), A-SCALA (Maunder and Watters 2003), Gadget (Begley 2005), or ADMB (ADMB Foundation 2011). I explicitly mean to exclude methods such as Virtual Population Analysis, and its descendents (e.g., Pope 1972; Shepherd 1999), in which composition data are treated as observed without error, and so not subject to data weighting. In statistical models, we can ignore the "other terms" (which might include penalty functions and, if the assessment is Bayesian, prior distributions) because they are not relevant to the question of data weighting. Our focus is on the functions L_{iy} and L_{jby}, which for these models are negative log-likelihoods, i.e., they describe the assumed error distribution for each data point. Many different error distributions have been used in stock assessment models but, all the common examples are expressed in terms of the expected value for the observation (E_{iy} or E_{jby}, which are functions of the parameters in $\boldsymbol{P}$) and a weighting parameter (see Table 2). The assumed (or estimated) values of weighting parameters determine how much weight is given to each data point. Thus, for example, if we assume a "normal by CV" error distribution for an abundance data point (T2.2 in Table 2) we can assign high

Table 2. Examples of negative log-likelihoods (ignoring constant terms) used in stock assessment models for abundance (O_{iy}) or composition (O_{jby}) data, showing the parameter that is used to weight each data point and the sense in which this parameter works (i.e., does a high weighting parameter value imply a high weight (+) or a low weight (–)).

		Weighting		
	Negative log-likelihood (L_{iy} or L_{jby})	Description	Parameter	Sense
T2.1A	$\lambda_{iy}(O_{iy} - E_{iy})^2$	Weighted sum of squares	λ_{iy}	+
T2.1B	$\lambda_{jby}(O_{jby} - E_{jby})^2$	Weighted sum of squares	λ_{jby}	+
T2.2	$\log(c_{iy}E_{iy}) + 0.5\left(\frac{O_{iy}-E_{iy}}{c_{iy}E_{iy}}\right)^2$	Normal by CV	c_{iy}	–
T2.3	$\log(\sigma_{iy}) + 0.5\left(\frac{O_{iy}-E_{iy}}{\sigma_{iy}}\right)^2$	Normal by SD	σ_{iy}	–
T2.4A	$\log(\sigma_{iy}) + 0.5\left[\frac{\log(O_{iy}/E_{iy})}{\sigma_{iy}} + 0.5\sigma_{iy}\right]^2$	Lognormal by CV[a]	c_{iy}[b]	–
T2.4B	$\log(\sigma_{jby}) - \log\left\{\exp\left[-0.5\left(\frac{\log(O_{jby}/E_{jby})}{\sigma_{jby}} + 0.5\sigma_{jby}\right)^2\right] + 0.01\right\}$	Robust lognormal	c_{jby}[c]	–
T2.5	$\log(\sigma_{iy}) + 0.5\left[\frac{\log(O_{iy}/E_{iy})}{\sigma_{iy}}\right]^2$	Lognormal by CV[a]	c_{iy}[b]	–
T2.6	$-N_{jy}O_{jby}\log E_{jby}$	Multinomial	N_{jy}	+
T2.7	$0.5\log(E'_{jby}) - \log\left\{\exp\left[\frac{-(O_{jby}-E_{jby})^2}{2E'_{jby}/N_{jy}}\right] + 0.01\right\}$	Robust multinomial[d]	N_{jy}	+

[a]The difference between the two lognormals is that in T2.4A, it is assumed that the expected value of O_{iy} is E_{iy}, whereas in T2.5 the expected value of $\log O_{iy}$ is assumed to be $\log E_{iy}$;

[b]$c_{iy} = \sqrt{\exp(\sigma_{iy}^2) - 1}$

[c]$c_{jby} = \sqrt{\exp(\sigma_{jby}^2) - 1}$

[d]$E'_{jby} = (1 - E_{jby})E_{jby} + 0.1/m_{\text{bin},j}$

(or low) weight to that observation by setting c_{iy} to be low (or high). Another way of thinking of this is that the weight we assign to an observation O_{iy}, is determined by how close we expect it to be to the model's expected value, E_{iy}; a high weight means O_{iy} is expected to be close to E_{iy} (so its CV will be small). Any adjustment of weighting parameters that gives more (or less) weight to a data set is said to up-weight (or down-weight) that data set.

It is important to understand that some sorts of information used in stock assessments are not "data", in the sense used in this paper, and thus not relevant to discussions of data weighting. For example, a set of age and length measurements on individual fish may be used outside the assessment model to estimate growth parameters (such as von Bertalanffy coefficients), which are then treated as fixed parameters in the model. In this case, neither the age–length measurements nor the growth parameters should be thought of as data, and thus subject to weighting. The age–length measurements would be considered as data only if they appeared in the objective function (in which case the growth parameters would be estimated within the model). Another type of information that I will not treat as data, and thus subject to weighting, is that used in a Bayesian model to construct a prior distribution for a model parameter. I will also ignore the parameter, common to many assessment models, that describes the degree of year-to-year variation in recruitment (e.g., the parameters σr of McAllister and Ianelli 1997; and λF of Savereide and Quinn 2004) because although it looks, mathematically, like a data-weighting parameter, it does not actually serve to weight any data set.

For readers interested in technical details, there are three points to be made about the examples in Table 2. First, some of the terms in the negative log-likelihoods may be considered optional. For example, many authors omit the $\log(\sigma_{iy})$ term in examples T2.3–T2.5. This is appropriate if, as is often true, the parameter σ_{iy} is assumed to be known; but it is inappropriate if σ_{iy} is to be estimated. Second, the weighted sum of squares (examples T2.1A, T2.1B) is not a negative log-likelihood, but it is included here because its use is not uncommon (e.g., Taylor et al. 2007; Hulson et al. 2008), and it may be interpreted as a "normal by SD" negative log-likelihood simply by setting $\sigma_{iy} = \sqrt{0.5/\lambda_{iy}}$ (for T2.1A), or $\sigma_{jby} = \sqrt{0.5/\lambda_{jby}}$ (for T2.1B). Third, for examples T2.4A, T2.4B, and T2.5 we have a choice in weighting parameters between the SDs in log space (σ_{iy} or σ_{jby}) and CVs in natural space (c_{iy} or c_{jby}). I have chosen the latter because they are more easily interpretable.

Two-stage weighting

It is common practice to weight some or all data sets in two stages. Stage 1 weights are devised before the model is run, and generally use information about the way in which the data were collected (i.e., sample sizes and structures). It is not uncommon for each data point to have a different weight. Stage 2 weighting occurs after the model has been run (or sometimes during a model run), and is intended to make the data weights more consistent with the model output. The weighting adjustments at stage 2 usually apply to whole data sets, rather than individual data points, using formulations like those in Table 3. For example, if equation T3.4 (Table 3) is used for composition data, then before the model is run, the stage 1 weights, $\widetilde{N}_{jy}$, will be fixed and provisional values will be assigned to the stage 2 weights, w_j. The assessment model is then run, information from that run

Table 3. Examples of equations that allow, for the example negative log-likelihoods of Table 2, two-stage data weighting.

			Weighting parameters	
	Equation	Corresponding example(s) in Table 2	Stage 1	Stage 2
T3.1	$\lambda_{iy} = \widetilde{\lambda}_{iy} w_i$	T2.1A	$\widetilde{\lambda}_{iy}$	w_i
T3.2	$c_{iy} = \widetilde{c}_{iy}/w_i$	T2.2, T2.4A, T2.5	$\widetilde{c}_{iy}$	w_i
T3.3	$\sigma_{iy} = \widetilde{\sigma}_{iy}/w_i$	T2.3	$\widetilde{\sigma}_{iy}$	w_i
T3.4	$N_{jy} = \widetilde{N}_{jy} w_j$	T2.6, T2.7	$\widetilde{N}_{jy}$	w_j
T3.5	$c_{iy} = \sqrt{\widetilde{c}_{iy}^2 + c_i^2}$	T2.2, T2.4A, T2.5	$\widetilde{c}_{iy}$	c_i
T3.6	$c_{jby} = \sqrt{\widetilde{c}_{jby}^2 + c_j^2}$	T2.4B	$\widetilde{c}_{jby}$	c_j
T3.7	$\sigma_{iy} = \sqrt{\widetilde{\sigma}_{iy}^2 + \sigma_i^2}$	T2.3	$\widetilde{\sigma}_{iy}$	σ_i
T3.8	$N_{jy} = 1/[(1/\widetilde{N}_{jy}) + (1/N_j)]$	T2.6, T2.7	$\widetilde{N}_{jy}$	N_j

Note: The tilde (~) above the symbol for a weighting parameter indicates that it applies to stage 1.

is used to adjust the stage 2 weights, and the model is run again with the adjusted weights. The process of adjusting the weights may be applied several times until the weights reach stable values (this is called iterative reweighting).

Some key data-weighting questions

Answers to the questions addressed in this section will help to set the scene for the rest of the paper.

Why is data weighting important?

The main reason that data weighting is important in stock assessments is that it can substantially change the results of the assessment. This can sometimes be demonstrated by constructing a profile on a key parameter (called a likelihood profile if estimation is by maximum likelihood, or a posterior profile if it is Bayesian; as illustrated in Fig. 1) (a profile is constructed by refitting the assessment model many times, each time with a different fixed value of the chosen parameter – unfished biomass, B_0, in this case). In this example, the profile shows that that the abundance data (a single time series of trawl survey biomass estimates) were best fitted when B_0 was 42 000 t; the composition data (proportions at age in catches from the survey and two fisheries) were best fitted at B_0 = 80 000 t; and in this assessment the data have been weighted in such a way that the model estimate of B_0 (44 840 t) is quite close to the best estimate from the abundance data. Data weighting was important in this assessment because 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. Also, the estimated current biomass (a key management indicator in New Zealand stock assessments) could have taken any value between 44% B_0 and 57% B_0, depending on the data weighting (Fig. 1*b*).

Another reason that data weighting is important in stock assessments is that it affects all the usual tools of statistical inference that are used in these assessments (Deriso et al. 2007). For example, we may want to use the Akaike Information Criterion (AIC; Akaike 1974) to decide which of two alternative equations for a selectivity curve is most consistent with our data, or to test the hypothesis that a fishery selectivity has changed in recent years. The decisions we make in these cases may be affected by a change in data weighting. Also, any estimated confidence (or credibility) intervals for parameters (or for derived quantities, like current biomass) will change when data weightings are changed.

Why is stage 2 weighting necessary?

There is an argument that says that we ought to be able to do without stage 2 weighting. After all, correct weighting requires knowledge of the error distribution of our observations, and many types of data allow us to estimate that error distribution before we start modelling. This appears to be the argument behind the claim by Maunder (2003, p. 470) that weighting factors are not needed in what he calls “the new integrated analysis."

The argument is wrong because it doesn’t recognise that there are three types of error to consider (Fig. 2). These arise because for any quantity we observe for our stock assessment (e.g., a biomass, or a proportion at age in a given year) there are three different values: (*i*) the value we observe, *O*; (*ii*) the value expected by our model, *E*; and (*iii*) the true (real world) value, *T*. *O* differs from *T* because of observation error, which is the error whose distribution, and likely size, we may be able to infer from our data (some ways of doing this are discussed below). *E* differs from *T* because of process error, by which I mean all the ways in which our model is only an approximation to the real world. The error we are interested in (i.e., that described in our negative log-likelihoods) is that between *O* and *E*. I call this total error because it is the sum of observation error and process error.

This characterization of errors gives an obvious interpretation to the two stages of data weighting. At stage 1, we assign the weights (or CVs) appropriate for observation error; at stage 2, we adjust those weights to allow for process error. This adjustment is usually done either multiplicatively (as in eqs. T3.1–T3.4 of Table 3) or additively (as in T3.5–T3.8). Note that when we add a process error with CV c_{proc} to an observation error with CV c_{obs}, the total error has CV $\sqrt{c_{\text{proc}}^2 + c_{\text{obs}}^2}$ (assuming the errors are independent, which

Fig. 1. Results from a profile on unfished biomass, B_0, in the New Zealand hake assessment of Horn and Francis (2010): (*a*), the (total) objective function (solid line), and the components of it associated with abundance (dashed line) and composition data (dotted line), (all zero-adjusted), with plotted points showing the minimum for each curve; and (*b*), the relationship between B_0 and current biomass in the profile.

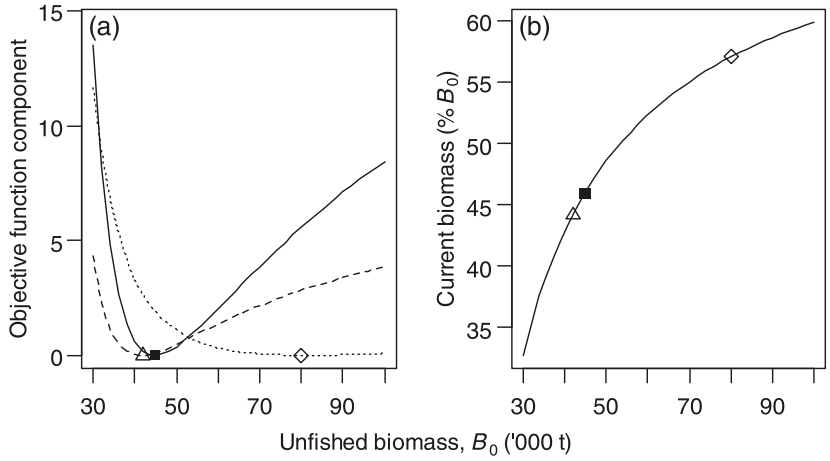
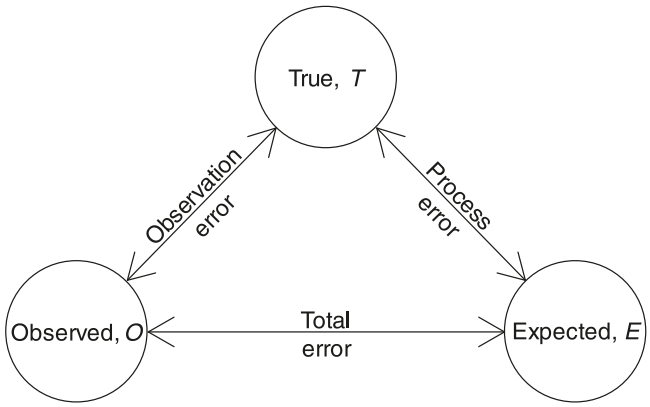


Fig. 2. Schematic illustration of the three types of error that exist between the three values of any quantity of interest (such as a biomass or proportion at age): the observed value, *O*; the true (real world) value, *T*; and the value *E* expected by our stock assessment model.



seems reasonable, given their sources). A similar logic applies to equation T3.8: if both *O* and *E* are multinomially distributed, with parameters (N_{obs},*T*) and (N_{proc},*T*), respectively, then the variance of ($O - E$) is the same as that of a multinomial distribution with parameters (N_{total},*T*), where $N_{total} = 1/[(1/N_{obs}) + (1/N_{proc})]$.

To avoid stage 2 weighting we must be able to estimate the likely size of both observation and process error for all our data before we run our model. I will show that it is possible to do this for some types of abundance data, but never for composition data. Thus, stage 2 weighting will always be necessary, at least for composition data.

Can we estimate process error outside the stock assessment model?

The answer to this question is yes for some types of abundance data, but no for composition data.

Several studies have attempted to estimate the process error associated with trawl survey biomass data by comparing the estimated observation errors with the sizes of the residuals when these biomass estimates are used in a stock assessment model. Pennington and Godø (1995) analysed four survey time series and concluded that the variance of the total error was approximately twice as large as that of the observation error (i.e., process and observation error were approximately equal in variance). Francis et al. (2003) took a different approach, assuming that the CV of the process error was the same for each of 18 trawl-survey time series, and estimated that a process error CV of about 0.2 would be needed to best explain the size of the stock assessment residuals. These two results are broadly similar because the mean observation-error CV in the latter study was 0.22. Both studies assumed that the main source of this process error was year-to-year variation in survey catchability, although this assumption was not necessary for their analyses. A third study, by Millar and Methot (2002), focussed primarily on between-species variation in mean catchability for six species of rockfish caught in a triennial series of eight trawl surveys off the west coast of the US. However, they also estimated year-to-year changes in catchability (assumed to be the same for all species), which suggest a process error CV of about 0.35 for this survey (this is the approximate CV, in natural space, of the midpoints of the distributions in Millar and Methot 2002's Fig. 5). This CV may be atypically high for research surveys because this particular survey series uses chartered commercial vessels and, as noted by Millar and Methot (2002), it is not possible to use the same vessels and skippers each time.

For acoustic survey biomass estimates there are several factors whose year-to-year variation has been identified as contributing to process error. These include mean target strength, the abundance of nontarget species, the proportion of the target population that is not detectable (because it is too close to the sea floor), and the timing of the survey relative to that of spawning activity. As long as the likely scale of amongst-survey variation is known for each factor, it is

possible to estimate the extent of process error using a simulation procedure (Rose et al. 2000; O'Driscoll 2004).

Process error in composition data is much more complex because it occurs in two dimensions (time and age or length), or sometimes three (if the data are sex-specific). The model assumptions that contribute most to this error are those associated with natural mortality and selectivities. For example, in many assessments (including that associated with Fig. 1) it is assumed that natural mortality is independent of both age and time, and fishery selectivities do not vary with time. In principle, we could use a simulation approach, similar to that used for acoustic surveys, to estimate the distribution of process error arising from the failure of these (or other similar) assumptions. This does not work in practice because we do not know how wrong the assumptions are (e.g., we do not know the scale of year-to-year variations in natural mortality, or how this scale varies with age). Thus it is not realistic to try to estimate the process error in composition data outside the stock assessment model.

Why abundance data should have primacy?

Some approaches to data weighting treat all data types in the same way. I suggest that this is a mistake, and that primacy should be given to abundance data, by which I mean that special attention should be devoted to ensuring that the abundance data are well fitted by the model. My reasons for this have to do with the purposes of stock assessment modelling.

Most of the key questions we address in stock assessments are to do with abundance. We want to know what is the current stock abundance (usually relative to historical levels, or to the unfished abundance), and whether it is increasing or decreasing. We might also want to know what the effect of possible future levels of fishing is likely to be, and we will express this effect in terms of biomass trends (rates of increase/decrease). Another quantity of interest is some measure of fishing pressure (usually either an instantaneous rate of fishing mortality, or a catch/biomass ratio) which, of course, is directly related to abundance (for a given catch, the greater the abundance is in any given year, the lower the fishing pressure must have been). We should give primacy to abundance data because they provide direct information about the stock assessment quantities that are of most interest, whereas composition data provide only very indirect information about these quantities. If we do not grant this primacy, then there is a danger that any signal from abundance data will be swamped by that from composition data, simply because the latter data type is typically much more numerous (in terms of individual data points).

It is easy to be misled by plots like Fig. 1, which suggest that composition data contain much useful information about abundance. The point to remember is that the likelihood calculations that produced this plot depend on very strong assumptions, which we know to be false, but which we need to make to have a useful stock assessment model. In this model it was assumed that natural mortality was independent of age and year, and that the selectivities of the two fisheries did not vary by year. Such assumptions are useful in allowing us to infer average selectivity curves for the survey and fisheries, and which year classes are particularly strong or weak. However, it seems to be a mistake to rely on them in making abundance inferences from composition data. In this particular assessment we would see the weakness of such inferences if we were to add a separate line in for each of the three composition data sets to Fig. 1. These additional lines (not plotted) would show that these data sets are inconsistent in their (apparent) abundance signal: the line for the catch at age data from one of the fisheries would have its minimum at the lowest value of B_0 (30 000 t), whereas for both the survey and the other fishery the minimum would be at the highest value (100 000 t).

The situation to avoid, is one in which the relative weighting given to composition data causes a poor fit to the abundance data. We should accept such a fit only if we are confident that the composition data provide clear evidence of abundance trends that differ from those suggested by our abundance data. This, I suggest, will be a rare occurrence. We rely on composition data to tell us about strong and weak year classes and the shape of selectivity curves, not about abundance trends.

How should we deal with abundance data that may be unrepresentative?

It is not uncommon in stock assessment documents and meetings for doubt to be expressed as to whether an abundance data set is representative, i.e., whether the trend in that data set is actually the same (allowing for observation error) as that in the population being assessed. For example, a catch per unit effort (CPUE) data set will be unrepresentative if it is from a fishery in which CPUE is not related to abundance. Survey estimates that cover only a part of the population will be unrepresentative if the population fraction covered by the survey is very different in different years. If two abundance data sets are clearly contradictory (i.e., they show very different trends over the same years) then at least one of them is likely to be unrepresentative.

It may seem reasonable to down-weight data sets that may be unrepresentative, but this is a bad idea. A better response is to consider alternative assessments in which possibly unrepresentative data sets are omitted. Thus, if there is only one suspect data set we should produce two assessments: one including this data set, and the other excluding it. These two assessments correspond to the two logical possibilities: A, the data set is representative; or B, it is not. An important uncertainty that must be communicated to fishery managers is that although only one of the two assessments is likely to be correct, we do not know which one. If we simply down-weight the suspect data set we will produce a result that lies somewhere between these two assessments. This result will be wrong in case A, and it will be wrong in case B, and we will not have alerted fishery managers to the possibility that one data set is unrepresentative. Both Richards (1991) and Schnute and Hilborn (1993) made a similar point when discussing the problem of contradictory data sets in stock assessments. In this case it is important to acknowledge the uncertainty about which data set is unrepresentative; a single assessment using all data sets is likely to be wrong, no matter which of the data sets turns out to be unrepresentative (Schnute and Hilborn (1993) describe a likelihood that allows both contradictory data sets to be used in the same assessment and also acknowledges the uncertainty about which one is unrepresentative, but this seems to me to be an elegant

way of demonstrating the problem, rather than a practical way of dealing with it in actual assessments).

Sometimes there will be clear evidence that an abundance data set is unrepresentative. For example, consider an abundance index that is initially stable and then suddenly increases (or decreases) sharply. This would clearly be unrepresentative in a fishery in which catches were stable and there was no evidence of extreme (very strong or weak) year classes. Data sets believed to be unrepresentative should be discarded; they certainly should not be retained and downweighted.

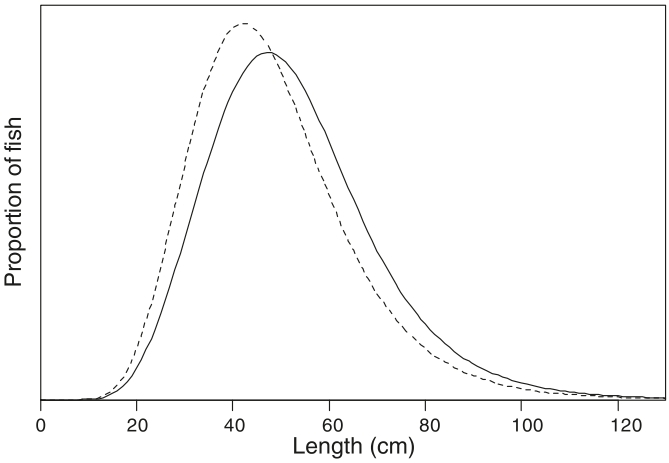
How should we deal with correlations in composition data?

It is commonly observed that fish in the same catch (e.g., from the same tow of a trawl net, or the same set of a longline) tend to be more similar to each other in length or age than are fish from different catches. Pennington and Vølstad (1994) called this intrahaul correlation. Sometimes this correlation can occur at a higher level, with fish caught by the same vessel, in a multiple-catch fishing trip, being more alike than fish caught by different vessels (e.g., see Fig. 10 of Francis 2006). These phenomena can induce substantial correlations in age or length composition data sets that are constructed from samples from multiple catches (whether from surveys or fisheries). The proportions O_{jby} and $O_{jb'y}$ will tend to be positively correlated if the bins b and b', are close together, and negatively correlated if they are distant (e.g., see Fig. 12 of Francis 2006). A very similar pattern was shown in Fig. 1 of Hrafnkelsson and Stefánsson (2004), although they were plotting different, but related, correlations (those between the length frequencies at different survey stations, rather than within the overall length frequency from a survey). These correlations are much bigger than those that always occur amongst multinomial proportions (the latter, which are caused by the fact that the proportions for each year must sum to 1, are always negative and usually small, having the following value: $-[E_{jby}E_{jb'y}/(1 - E_{jby})(1 - E_{jb'y})]^{0.5}$).

The effect of these correlations is to reduce the amount of information in composition data sets. This was demonstrated by Pennington and Vølstad (1994), who analysed length frequency data for haddock in 26 trawl surveys on Georges Bank. For each survey they calculated what they called the effective sample size: the number of fish that would be required, in a simple random (i.e., uncorrelated) sample from the population, to estimate the population mean length with the same precision as was achieved in the survey. Over the 26 surveys, the median effective sample size was 21, which was half the median number of tows that caught haddock, and only 2.5% of the median number of fish measured (these numbers come from Table 1 of Pennington and Vølstad 1994). In a similar analysis, based on trawl surveys conducted in the Barents Sea, off Namibia, and off South Africa, Pennington et al. (2002) found that the effective sample size for each survey was, on average, about the same as the number of tows.

The correlations discussed so far are part of the observation error for composition data, but there is also likely to be correlation in the process error. Consider, for example, a stock assessment in which a fishery selectivity is assumed to be constant, but actually varies from year to year. In a year in which the fishery happens to target larger (or older) fish than usual, the selectivity curve will be shifted to the right, compared with that expected in the model, and so will the observed length (or age) frequency from the catch (Fig. 3). Notice that the residuals ($O_{jby} - E_{jby}$) are all negative for the smaller lengths (up to about 48 cm), and all positive for the greater lengths (Fig. 3). This pattern would be reversed in a year in which the selectivity curve was shifted to the left. Thus, this year-to-year shifting of the selectivity would induce the same type of correlations in the process error for the proportions at length as was described above for the observation error (i.e., positive correlations between bins that are close together, and negative correlations for distant bins). Other types of process error would produce a similar effect (e.g., year-to-year variation in the natural mortality on small fish – perhaps caused by fluctuations in their food supply).

Fig. 3. Illustration of the relationship between the observed (solid line) and expected (broken line) catch length frequencies in a year in which the true fishery selectivity is shifted to the right of that assumed in the stock assessment model.



How should we take account of the correlations associated with composition data sets? There are three possible approaches, none of which is completely satisfactory. The first, and most common approach, is to ignore the correlations (i.e., to act as if they were zero). A simple error model is used for the composition data (e.g., T2.4B or T2.6 in Table 2), and some method that assumes zero correlations is used to set weights at stages 1 and 2. Some of the many variants of this approach are described in Appendix A. The danger with this approach is that the composition data sets are likely to be over-weighted (remember that correlations tend to reduce the amount of information in composition data, and less information should mean less weight), and this could cause poor fits to abundance data. The second approach is to use a complicated error model with an appropriate correlation structure. One problem with this approach is that although there have been several attempts to model observation error in composition data (e.g., Kvist et al. 2001; Hrafnkelsson and Stefánsson 2004; Miller and Skalski 2006) we have no way of knowing whether these will be appropriate for total (i.e., observation + process) error. Another problem is that these error models typically require many more

Table 4. Final (stage 2) multinomial sample sizes, estimated using four alternative methods of calculation (TA1.1, TA1.2, TA1.3, TA1.8; see Appendix A for details), for four age composition data sets with multinomial errors (and assumed sample size 150) in one model run from the 2006 assessment of southern hake (*Merluccius australis*) in Chile.

	Estimated sample size			
Composition data set	TA1.1*	TA1.2*	TA1.3*	TA1.8†
Trawl fishery	258	115	116	15
Commercial longline fishery	240	154	152	10
Artisanal longline fishery	557	251	248	25
Survey	338	232	233	69

*These methods do not allow for correlations.
†This method allows for correlations.

parameters than the simple models. I am not aware of any assessments using this second approach. I favour the third approach, which is to use simple error models, but to allow for correlations by using a data-weighting approach analogous to the method of Pennington and Vølstad (1994) for calculating effective sample sizes. That is to say, the weights (in the form of CVs or multinomial sample sizes) are calculated so as to be consistent with the size of the errors in mean length (for length composition data) or mean age (for age compositions) (see Appendix A for details of three methods using this approach).

To see the effect of allowing for correlations, consider a model run from the 2006 assessment of southern hake (*Merluccius australis*) in Chile, in which there were four age composition data sets, all of which were assumed to have a multinomial error distribution with sample size 150 (the same for all years). If we use stage-2 weighting methods that ignore correlations, the resulting adjusted sample sizes are mostly greater than 150 (see columns 2–4, Table 4), suggesting that insufficient weight was given to the composition data sets in this model. However, when we allow for correlations, the adjusted sample sizes are all much smaller than 150 (last column, Table 4), implying that the composition data were over-weighted. This pattern is not uncommon; multinomial sample sizes calculated with allowance for correlations are usually much smaller than those in which correlations are ignored. An effect of the correlations is illustrated (Fig. 4). Although there is broad agreement between the observed mean ages and those expected from the model (in that both show an overall decline), the expected values are mostly outside the 95% confidence intervals for the observed values. In other words, year-to-year variation in the age composition data was much greater than is consistent with a simple random sample size of 150 (as was assumed in calculating the confidence intervals). Use of this sample size implies that these data sets contain more information than they really do.

I should emphasise that my preference for stage-2 weighting methods based on errors in mean length or age is not based on any statistical theory supporting these methods. It is simply a pragmatic response to (*i*) the fact that the commonly used methods are based on an assumption (of uncorrelated residuals) that is demonstrably wrong; and (*ii*) the empirical observation (in plots like Fig. 4) that these methods give too much weight to composition data.

How can robust likelihoods help?

Robust likelihoods are particularly important for composition data because they help to avoid the situation in which the model fit is driven by a small number of composition data points that are extreme outliers. These likelihoods allow a better fit to the majority of composition data points and reduce the probability that the composition data will prevent good fits to abundance data.

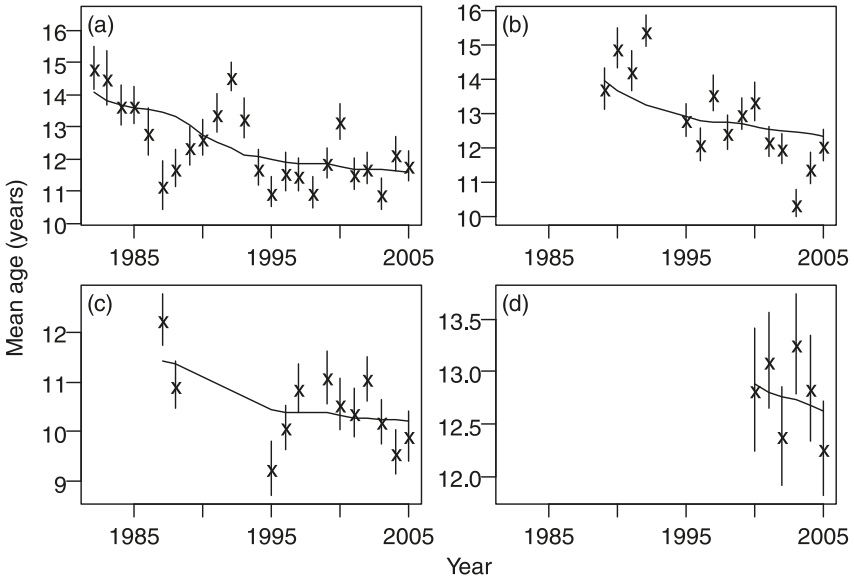
Fournier et al. (1990) pointed out that standard maximum-likelihood methods can perform poorly when applied to length composition data because they are too sensitive to the deviations from the model's hypotheses (i.e., process error) that are common with such data. They devised a robust alternative to the multinomial likelihood (T2.7 in Table 2: I have called this a robust multinomial because of the way it is parameterised, but technically it is a robust multivariate normal), which is used in MULTIFAN-CL (Fournier et al. 1998), a stock assessment program used for many tuna stocks. [In the original version of the robust multinomial, the term N_{jy} was replaced by $\min(N_{jy}, 1000)$; I omitted this here because I think it will rarely be used in stock assessments with effective stage 2 weighting]. A modification of T2.7, in which $E'_{jby} = (1 - O_{jby})O_{jby} + 0.1/m_{\mathrm{bin},j}$, is available in some programs (e.g., Maunder and Watters 2003; Bull et al. 2008) following the finding by Starr et al. (1999) that the original likelihood can produce biased estimates. An approach similar to that of Fournier et al. (1990) was used to make a robust lognormal (T2.4B in Table 2) that is available in CASAL (Bull et al. 2008).

Proposed approach to data weighting

From the questions discussed above I derive three principles and an end point to aim for in addressing the problem of data weighting in stock assessments. Principle 1: Do not let other data stop the model from fitting abundance data well. Principle 2: When weighting composition data, allow for correlations. Principle 3: Do not down-weight abundance data because they may be unrepresentative.

From a data-weighting point of view, the ideal end point of a stock assessment is a single assessment model in which

Fig. 4. Observed (×) and expected (curved line) mean age for the four age composition data sets in the stock assessment of Table 4: (*a*) commercial trawl; (*b*) commercial longline; (*c*) artisanal longline; and (*d*) survey. The vertical lines are 95% confidence intervals for mean age, calculated assuming simple random sampling with sample size 150 in each year.



all abundance data sets are fitted well. This may not be possible when there are conflicts amongst the data sets. In that case, we should aim for a set of alternative assessment models, in each of which one or more data sets has been omitted but all remaining abundance data sets are well fitted. The following approach to data weighting is intended to achieve this desired end point while respecting the above three principles.

Weighting abundance data

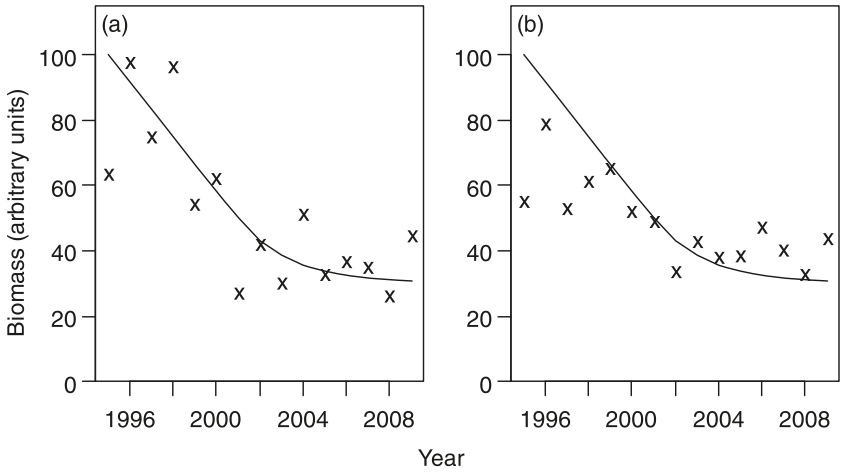
To apply Principle 1, we must have some idea, before we run the stock assessment model, of how we are going to decide whether the model has fitted the abundance data well. That is to say, we need to know how large the CVs (or SDs) of the total errors for these data sets should be. I have described above some methods for estimating these total CVs for trawl and acoustic indices. For other abundance data sets, I suggest using the approach adopted by Clark and Hare (2006, p. 9) for a CPUE data set: use the CV of the residuals of the fit of a data smoother to the abundance data (this is equivalent to saying that we expect the stock assessment model to fit these data as well as the smoother). These total CVs should be applied at stage 1 (i.e., before the model is run). They should not be adjusted at stage 2, because they have already been set to reflect our expectations as to how well the model should fit these data.

Weighting other types of data

For composition data, I suggest using robust likelihoods (e.g., T2.4B or T2.7 from Table 2), to reduce the influence of extreme outliers, and setting the initial (stage 1) weights to represent the approximate size of observation errors. Because these initial weights will be adjusted, often substantially, with stage-2 weighting, their relative sizes (within data sets) are more important than their absolute sizes. For example, it is important that years with poorer sampling of the catch be given less weight than those with better sampling. The weighted sum of squares (T2.1B in Table 2) should be avoided for composition data, because it implies that for a proportion p, $\mathrm{CV}(p) \propto 1/p$, whereas the more typical pattern is that for a multinomial proportion, i.e., $\mathrm{CV}(p) \propto \sqrt{(1-p)/p}$ (e.g., see Fig. 3, Crone and Sampson 1998). Because composition data typically derive from complex multistage samples, bootstrap resampling is an excellent (although computationally intensive) way of estimating observation-error CVs. Some other approaches have been suggested by Morton and Bravington (2008) and Robotham et al. (2008). If the (robust) multinomial distribution is preferred to the lognormal, a nonlinear regression (of CV on proportion) can be used to calculate the equivalent multinomial sample size from the bootstrap-estimated CVs (Crone and Sampson 1998). If these more complicated approaches are not possible, then I see little harm in the use of the (preferably robust) multinomial with sample sizes set by some ad hoc rule (e.g., Crone and Sampson 1998 estimated a relationship with number of trips sampled; Gilbert and Phillips 2003 set sample sizes to be 5× the number of landings sampled; Maunder and Watters 2003 used number of wells sampled; see also the discussion of length composition sample sizes in Thompson et al. 2009). I don't think it is important to allow for correlations (Principle 2) at stage 1.

Stage-2 weighting is important for composition data because it allows us to include process error, which cannot be estimated before the model is run, and thus avoid overweighting these data. Many different methods of stage-2 weighting have been proposed (see Appendix A). Those methods that allow for correlations are to be preferred, because they almost always produce smaller weights for the composition data (as in Table 4) and thus reduce the probability that these data will cause a poor fit to the abundance data (Principle 1).

Fig. 5. Demonstration that the standard deviation of the normalized residuals (SDNR) by itself is not a good measure of how well a model biomass trajectory (line) fits a set of biomass observations (×). The SDNR is exactly the same in both panels but the residual patterns indicate a good fit in panel (*a*), and a poor fit in panel (*b*).



One weakness of stage-2 weighting methods that allow for correlations is that they don't work well with data sets that cover few years. For example, if we have a 5-year composition data set with 20 age bins, the stage-2 weights from such methods are not well estimated, because they derive from just 5 mean-age errors, whereas those from most methods ignoring correlations (e.g., methods TA1.1–TA1.3) are based on the errors at 100 data points (or 200, if sex is included). It will sometimes be sensible to make inferences from other larger data sets. For example, if we find for our large composition data sets that the adjusted sample sizes with method TA1.8 are approximately 20%, say, of those calculated with TA1.1, then we might assume that this ratio is true for a small data set. Don't be concerned if the final weights for composition data are rather small. Recall the results quoted above from Pennington and Vølstad (1994) and Pennington et al. (2002), which show that effective sample sizes for length compositions can be very small (those for age compositions based on age-length keys tend, in my experience, to be larger). These small weights will reduce the chance that composition data will cause a poor fit to abundance data (Principle 1) and often do not greatly affect estimates of those parameters about which these data are most informative (year-class strengths and selectivity curves).

It is difficult to provide specific advice for data types other than abundance and composition, e.g., proportions mature by length (Taylor et al. 2007), and length increments from mark–recapture experiments (Breen et al. 2003). The only general point I would make is that Principle 1 should apply to these data.

Judging whether abundance data are well fitted

When we have completed stage 2 weighting and rerun the model with the adjusted weights, we should check whether the abundance data are well fitted. A common way to do this is to calculate, for each abundance data set, the standard deviation of the normalized (or standardized) residuals (SDNR) (Breen et al. 2003) (see Appendix B for methods of calculation). For an abundance data set to be well fitted, the SDNR should not be much greater than 1 (a value much less than 1, which means that the data set is fitted better than was expected, is not a cause for concern). What is meant by "much greater than 1" depends on m (the number of years in the data set). Since the normalized residuals are (approximately) normally distributed, a reasonable guideline is that the SDNR should be less than $[\chi^2_{0.95,m-1}/(m-1)]^{0.5}$, where $\chi^2_{0.95,m-1}$ is the 95th percentile of a χ^2 distribution with $m-1$ degrees of freedom (e.g., this means upper limits of 1.54, 1.37, and 1.26 for $m = 5$, 10, and 20, respectively).

Although an SDNR not much greater than 1 is a necessary condition for a good fit, it is not sufficient. It is important to plot the observed and expected abundances to be sure that the fit is good (Fig. 5).

What to do when abundance data are not well fitted

When an abundance data set is not well fitted by an assessment model, the first thing to check is whether Principle 1 is being violated. Can we get a good fit by either up-weighting the abundance data or down-weighting the composition (or any other) data? If so, this should be done. A point to note is that if, to achieve a good fit, an abundance data set is up-weighted, its SDNR after up-weighting should be calculated using the original weights (because these weights represent our best information about how well this data set should be fitted).

What should be done if the problem does not lie with the composition data? If there is only one abundance data set, then this should be considered unrepresentative (unless there is some acceptable way to modify the model structure to achieve a good fit), and thus discarded. If there are multiple abundance data sets, and one or more is not well fitted, then there is good reason to believe that at least one data set is unrepresentative (although not necessarily one of those that is not well fitted). The approach to take here is as outlined above in the section on unrepresentative data sets. Create a set of alternative models, in each of which one or more abundance data sets is omitted, and all the remaining abundance data sets are well fitted. The set of alternative models should

be as small as possible. The point to note here is that we should not simply down-weight potentially unrepresentative data sets (Principle 3); instead we should investigate the effect of omitting them, or other data sets which may be in conflict with them. The aim is to present the set of alternative models to fishery managers and say that each one of these models could be true, but we don't know which one. It may be possible to use the expertise of people with knowledge of the data, and the fishery, to provide an indication as to which of the alternative models are more, or less, likely to be true.

Dealing with doubt about abundance data

Principles 1 and 3 should not be interpreted as suggesting that stock assessment scientists should ignore any doubts they may have about the representativeness of their abundance data. Nor do I want to suggest that it is easy to determine whether such data are representative. It is not. What I am advocating is that we deal with any doubt about the representativeness of a particular abundance data set by constructing an alternative model (or models) that omits that data set. Any model that includes this data set is intended to address the possibility that the data set *is* representative, so for this model we apply Principles 1 and 3. The model (or models) without this data set addresses the possibility that the data set is not representative.

Goodness of fit to composition data

The above data-weighting procedure does not involve checking whether the composition data are well fitted. This is because, in my experience, a poor fit to composition data is much more likely to be caused by poor model assumptions than by inappropriate data weights. For example, fits to fisheries composition data may be poor if the wrong selectivity curve is assumed (e.g., if a logistic selectivity is assumed when the actual selectivity is strongly domed). Note that SDNRs are not an appropriate measure of goodness of fit for composition data because the theory underlying them assumes that the errors are uncorrelated.

Discussion

This paper offers three aids to those dealing with the difficult problem of data weighting in fisheries stock assessments: (*i*) some guiding principles; (*ii*) an end to aim for; and (*iii*) a proposed approach to achieve that end. Of these, the last is least important. For many assessments, even a simple trial and error approach, as suggested by Fournier and Archibald (1982), may be adequate if it is aimed towards the end I have described, and supported by the guiding principles.

One source of difficulty in data weighting is that some of the decisions involved are inevitably subjective. For example, there is no objective way of deciding whether abundance data are adequately well fitted, or of setting the degree of smoothing to be used in setting CVs for abundance data sets. This necessity for subjective decisions underlines the importance of expert knowledge (concerning the fishery, the data, and the models) in stock assessment. I don't think the problem of subjectivity is sufficiently severe that we should abandon statistical population models, as advocated by Cotter et al. (2004).

Another difficulty is that there are still many unresolved technical problems. For example, should observation and process errors be combined multiplicatively (as in T3.1–T3.4) or additively (as in T3.5–T3.7)? It is unclear why method TA1.1 performs so differently from TA1.2 and TA1.3 in the example of Table 4, although the derivations of these methods show they are intended to achieve the same thing (see Appendix A). I know of only one approach to stage-2 weighting of composition data that allows for correlations (method TA1.8, with variants TA1.9 and TA1.10). If other researchers addressed the problem of devising such a method their approaches might be equally plausible, while producing quite different results. There is a need for a simple and plausible error distribution for composition data that allows for substantial correlations.

Even when the data-weighting approach described above appears to work well, and produces a single assessment model in which all data sets are well fitted, it is a useful exercise to experiment with different data weightings (or exclusion of some data sets). This provides important information about the robustness of our assessment. Not all stock assessments are sensitive to changes in data weighting (Breen et al. 2003; Savereide and Quinn 2004) but we can't know about any such sensitivity unless we investigate alternative weightings.

I offer brief comments on some earlier studies that consider data weighting. McAllister et al. (2001) considered just the weighting of abundance sets, in a setting different from that addressed here in that composition data were either not used (surplus production models) or were treated as known without error (Virtual Population Analysis). Gavaris and Ianelli (2002) identified data weighting as one of four generic statistical issues important in stock assessments and described two versions of two-stage weighting (one uses a special case of TA1.1 [see their Table 1]; in the other, the terms "extrinsic" and "intrinsic" [see their p. 257] apparently correspond to my "stage 1" and "stage 2"). Breen et al. (2003) estimated a parameter (their $\tilde{\sigma}$) that is the standard deviation of a component of error common to all data sets. This is an interesting approach that has little effect on point estimates of parameters, but has the merit of reducing the effect of changes in data weights on measures of the uncertainty of parameter estimates. Taylor et al. (2007), applying an approach to data weighting proposed by Stefánsson (2003), set weights for each data set according to how well it was fitted by the model when it was strongly up-weighted. This idea seems worth pursuing, but I think it would be much better if the objective function used proper likelihoods, rather than simple sums of squares, and of course Principle 1 must apply. Irwin et al. (2008) used concentrated likelihoods (i.e., weighting parameters were replaced by their maximum likelihood estimates), so that their data sets were essentially self-weighting, and then, as a sensitivity analysis, investigated the effect of sequentially up-weighting each data set by a factor of 10. This approach does not address the problem of correlations in composition data, nor does it give primacy to abundance data. Candy (2008) described a complicated stage-2 weighting method for composition data which involved treating the error in the model fit to these data as being the sum of two components: random error and systematic lack of fit. Though this characterization of the error seems sensible, I do not agree that the stage-2 weighting should include only the first component.

Acknowledgments

I am grateful to Cristian Canales and Juan Carlos Quiroz of the Instituto de Fomento Pesquero in Valparaiso for permission to use data from the Chilean southern hake assessment; to Alistair Dunn, Mark Maunder, Dave Fournier, and Paul Starr for technical information about some forms of data weighting; and to three anonymous referees who alerted me to some important studies that I missed, and whose comments have, I hope, enabled me to clarify some of my arguments.

References

ADMB Foundation. 2011. AD Model Builder version 10.1. ADMB Foundation, University of Hawaii, Manoa, Hawaii. Available from admb-project.org/ [accessed 16 May 2011].

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.

Begley, J. 2005. Gadget User Manual. Technical Report 120. Marine Research Institute, Reykjavik, Iceland.

Breen, P.A., Kim, S.W., and Andrew, N.L. 2003. A length-based Bayesian stock assessment model for the New Zealand abalone *Haliotis iris.* Mar. Freshw. Res. **54**(5): 619–634. doi:10.1071/MF02174.

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H., and Bian, R. 2008. CASAL (C++ algorithmic stock assessment laboratory): CASAL User Manual v2.20–2008/02/14. NIWA Technical Report 130. National Institute of Water & Atmospheric Research, Wellington, New Zealand.

Candy, S.G. 2008. Estimation of effective samplesizefor catch-at-age and catch-at-length data using simulated data from the Dirichlet-multinomial distribution. CCAMLR Science, **15**: 115–138. Available from www.ccamlr.org/ccamlr_science/Vol-15-2008/07candy.pdf.

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

Cotter, A.J.R., Burt, L., Paxton, C.G.M., Fernandez, C., Buckland, S. T., and Pan, J.-X. 2004. Are stock assessment methods too complicated? Fish Fish. **5**: 235–254.

Crone, P.R., and Sampson, D.B. 1998. Evaluation of assumed error structure in stock assessment models that use sample estimates of age composition. *In* Fishery Stock Assessment Models. *Edited by* F. Funk, T.J. Quinn, J. Heifetz, J.N. Ianelli, J.E. Powers, J. F. Schweigert, P.J. Sullivan, and C.-I. Zhang. Alaska Sea Grant College Program Report No. AK-SG-98–01. University of Alaska, Fairbanks, Alaska. pp. 355–370.

Deriso, R.B., Maunder, M.N., and Skalski, J.R. 2007. Variance estimation in integrated assessment models and its importance for hypothesis testing. Can. J. Fish. Aquat. Sci. **64**(2): 187–197. doi:10.1139/F06-178.

Fournier, D.A., and Archibald, C.P. 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. **39**(8): 1195–1207. doi:10.1139/f82-157.

Fournier, D.A., Sibert, J.R., Majkowski, J., and Hampton, J. 1990. MULTIFAN a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*). Can. J. Fish. Aquat. Sci. **47**(2): 301–317. doi:10.1139/f90-032.

Fournier, D.A., Hampton, J., and 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. doi:10.1139/cjfas-55-9-2105.

Francis, R.I.C.C. 2006. Some recent problems in New Zealand orange roughy assessments. New Zealand Fisheries Assessment Report 2006/43, Ministry of Fisheries, Wellington, New Zealand.

Francis, R.I.C.C., Hurst, R.J., and Renwick, J.A. 2003. Quantifying annual variation in catchability for commercial and research fishing. Fish Bull. **101**: 293–304.

Gavaris, S., and Ianelli, J.N. 2002. Statistical issues in fisheries' stock assessments. Scand. J. Stat. **29**(2): 245–267. doi:10.1111/1467-9469.00282.

Gilbert, D.J., and Phillips, N.L. 2003. Assessment of the SNA 2 and Tasman Bay/Golden Bay (SNA 7) snapper fisheries for the 2001–02 fishing year. New Zealand Fisheries Assessment Report 2003/45, Ministry of Fisheries, Wellington, New Zealand.

Horn, P.L., and 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, Ministry of Fisheries, Wellington, New Zealand.

Hrafnkelsson, B., and Stefánsson, G. 2004. A model for categorical length data from groundfish surveys. Can. J. Fish. Aquat. Sci. **61**(7): 1135–1142. doi:10.1139/f04-049.

Hulson, P.-J.F., Miller, S.E., Quinn, T.J., Marty, G.D., Moffitt, S.D., and Funk, F. 2008. Data conflicts in fishery models: incorporating hydroacoustic data into the Prince William Sound Pacific herring assessment model. ICES J. Mar. Sci. **65**(1): 25–43. doi:10.1093/icesjms/fsm162.

Irwin, B.J., Treska, T.J., Rudstam, L.G., Sullivan, P.J., Jackson, J.R., VanDe Valk, A.J., and Forney, J.L. 2008. Estimating walleye (*Sander vitreus*) density, gear catchability, and mortality using three fishery independent data sets for Oneida Lake, New York. Can. J. Fish. Aquat. Sci. **65**(7): 1366–1378. doi:10.1139/F08-062.

Kvist, T., Gislason, H., and Thyregod, P. 2001. Sources of variation in the age composition of sandeel landings. ICES J. Mar. Sci. **58**(4): 842–851. doi:10.1006/jmsc.2001.1075.

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. doi:10.1111/j.1939-7445.2003.tb00123.x.

Maunder, M.N., and Watters, G.M. 2003. A-SCALA: an age-structured statistical catch-at-length analysis for assessing tuna stocks in the eastern pacific ocean. Bull. I-ATCC, **22**: 435–437.

McAllister, M.K., and 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. doi:10.1139/cjfas-54-2-284.

McAllister, M., Babcock, E.A., and Pikitch, E.K. 2001. Evaluating the relative merits of alternative methods to weight different time series of abundance indices in stock assessment. Col. Vol. Sci. Pap. ICCAT, **52**: 1094–1151.

Millar, R.B., and Methot, R.D. 2002. Age-structured meta-analysis of U.S. West Coast rockfish (Scorpaenidae) populations and hierarchical modeling of trawl survey catchabilities. Can. J. Fish. Aquat. Sci. **59**(2): 383–392. doi:10.1139/f02-009.

Miller, T.J., and 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. doi:10.1139/F06-022.

Morton, R., and Bravington, M. 2008. Comparison of methods for estimating age composition with application to Southern Bluefin Tuna (*Thunnus maccoyii*). Fish. Res. **93**(1-2): 22–28. doi:10.1016/j.fishres.2008.02.009.

National Research Council. 1998. Improving fish stock assessments. National Academy Press, Washington, D.C.

NOAA. 2011. Stock synthesis. Version 3. NOAA Fisheries Toolbox. National Oceanographic and Atmospheric Administration, Wodds Hole, Mass. Available from nft.nefsc.noaa.gov/SS3.html [accessed 16 May 2011].

O'Driscoll, R.L. 2004. Estimating uncertainty associated with acoustic surveys of spawning hoki (*Macruronus novaezelandiae*) in Cook Strait, New Zealand. ICES J. Mar. Sci. **61**(1): 84–97. doi:10.1016/j.icesjms.2003.09.003.

Pennington, M., and Godø, R.O. 1995. Measuring the effect of changes in catchability on the variance of marine survey abundance indices. Fish. Res. **23**(3-4): 301–310. doi:10.1016/0165-7836(94)00345-W.

Pennington, M., and Vølstad, J.H. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics, **50**(3): 725–732. doi:10.2307/2532786.

Pennington, M., Burmeister, L.-M., and Hjellvik, V. 2002. Assessing the precision of frequency distributions estimated from trawl-survey samples. Fish Bull. **100**: 74–80.

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

Richards, L.J. 1991. Use of contradictory data sources in stock assessments. Fish. Res. **11**(3-4): 225–238. doi:10.1016/0165-7836(91)90003-X.

Robotham, H., Young, Z.I., and Saavedra-Nievas, J.C. 2008. Jacknife method for estimating the variance of the age composition using two-phase sampling with an application to catches of swordfish (*Xiphias gladius*). Fish. Res. **93**(1-2): 135–139. doi:10.1016/j.fishres.2008.03.007.

Rose, G., Gauthier, S., and Lawson, G. 2000. Acoustic surveys in the full monte: simulating uncertainty. Aquat. Living Resour. **13**(5): 367–372. doi:10.1016/S0990-7440(00)01074-3.

Savereide, J.W., and Quinn, T.J., II. 2004. An age-structured assessment model for chinook salmon (*Oncorhynchus tshawytscha*). Can. J. Fish. Aquat. Sci. **61**(6): 974–985. doi:10.1139/f04-039.

Schnute, J.T., and Hilborn, R. 1993. Analysis of contradictory data sources in fish stock assessment. Can. J. Fish. Aquat. Sci. **50**(9): 1916–1923. doi:10.1139/f93-214.

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. doi:10.1006/jmsc.1999.0498.

Starr, P.J., Bentley, N., and Maunder, M.N. 1999. Assessment of the NSN and NSS stocks of red rock lobster (*Jasus edwardsii*) for 1998. New Zealand Fisheries Assessment Research Document 99/34. Ministry of Fisheries, Wellington, New Zealand.

Stefánsson, G. 2003. Issues in multispecies models. Nat. Resour. Model. **16**(4): 415–437. doi:10.1111/j.1939-7445.2003.tb00121.x.

Taylor, L., Begley, J., Kupcal, V., and Stefansson, G. 2007. A simple implementation of the statistical modelling framework Gadget for cod in Icelandic waters. Afr. J. Mar. Sci. **29**(2): 223–245. doi:10.2989/AJMS.2007.29.2.7.190.

Thompson, G.G., Ianelli, J.N., and Wilkins, M.E. 2009: Assessment of Pacific cod in the Gulf of Alaska. *In* Stock Assessment and Fishery Evaluation Report for the Groundfish Resources of the Gulf of Alaska. *Edited by* North Pacific Fishery Management Council, Anchorage, Alaska. pp. 165–352.

Appendix A: Some methods of stage-2 weighting for composition data

In this appendix I provide equations for 10 methods of stage-2 weighting for composition data (Table A1) and describe their derivations.

McAllister and Ianelli (1997) presented a method of stage-2 weighting assuming a multinomial error structure. They assumed that the stage-2 sample size is the same for all years in a data set, and they took this sample size as the average of the N_{jy} calculated using TA1.1 (see eqs. (2.5) and (2.6) in their Appendix 2). Their approach is based on the fact that, with a multinomial distribution, $\mathrm{Var}(O_{jby} - E_{jby}) = \mathrm{Var}(O_{jby}) = E_{jby}(1 - E_{jby})/(w_j\widetilde{N}_{jy})$, and also that $\mathrm{Var}(O_{jby} - E_{jby}) \approx (O_{jby} - E_{jby})^2$. Therefore, $E_{jby}(1 - E_{jby})/(w_j\widetilde{N}_{jy}) \approx (O_{jby} - E_{jby})^2$. Equation TA1.1, which can be derived by summing this last equation over b and rearranging the terms, is a generalization of the method of McAllister and Ianelli (1997) in which the stage-2 sample size for each year is assumed to be a multiple of that for stage 1 (i.e., $N_{jy} = \widetilde{N}_{jy}w_j$).

Method TA1.3 applies to the same situation as TA1.1 (i.e., weighting assumption T3.4) but its derivation is based on the assumption, that for each j, $\sum_{by}\frac{N_{jy}(O_{jby}-E_{jby})^2}{E_{jby}}$ has a χ^2 distribution with $m_{\mathrm{year},j}(m_{\mathrm{bin},j} - 1)$ degrees of freedom (this is the assumption underlying the usual χ^2 test used in the analysis of contingency tables) and so has expected value $m_{\mathrm{year},j}(m_{\mathrm{bin},j} - 1)$. Dunn and Hanchet (2009) used this same assumption to create method TA1.6 for the situation where weighting assumption T3.8 (i.e., $1/N_{jy} = 1/\widetilde{N}_{jy} + 1/N_j$) is to be used. Note that eq. TA1.6 has no explicit solution for N_j (i.e., it must be solved numerically). However, in the special case where the initial sample sizes are the same for all years (i.e., $\widetilde{N}_{jy} = \widetilde{N}_j$ for all y) there is an explicit solution, which is method TA1.7.

The methods discussed so far use a multinomial error for the composition data. If, instead, we assume a lognormal error, with additive weighting assumption T3.6, it is possible to estimate the stage-2 weighting parameter, c_j, directly as a model parameter (i.e., c_j is chosen to minimize the objective function). This is method TA1.4, which was devised for CASAL (Bull et al. 2008).

In developing methods TA1.2 and TA1.5 I applied the same general approach to two specific situations. I will first describe the general approach, and then how it was applied in these two situations. The aim in this approach was to standardize the errors ($O_{jby} - E_{jby}$) so that they all have the same variance. That is, I wanted to find m quantities, X_{jby}, (which will be functions of the weighting parameter, w_j or c_j), so that the standardized error, $S_{jby} = (O_{jby} - E_{jby})/X_{jby}$, had constant variance, i.e., $\mathrm{Var}(S_{jby}) = k_j$, for all years y, and bins b. Here, I am thinking of the S_{jby} as being a set of m random variables whose distributions (determined by the assumptions of the multinomial or lognormal error model) all have mean 0 and variance k_j. When we do our stage-2 weighting, after running our stock assessment model, we can calculate the actual value of each of these standardized errors, and how that value changes as we change our weighting parameter. Our aim is to find the value of the weighting parameter that makes the variance of this set of m standardized errors, which I write as $\mathrm{Var}(S_{jby})$, equal to k_j. [Note the important distinction between $\mathrm{Var}(S_{jby})$, which is the variance of a random variable, and $\mathrm{Var}_{by}(S_{jby})$, which is the variance of a set of m numbers calculated from the model output].

Now I show how I applied this general approach to two specific situations. If we assume multinomial errors and $N_{jy} = w_j\widetilde{N}_{jy}$ (i.e., weighting assumption T3.8) we can use

Table A1. Equations for some methods of stage-2 weighting of composition data, grouped by the weighting assumption (from Table 3) on which they are based, and whether they allow for substantial correlations (see text for explanations and sources).

Method	Weighting assumption	Equation	Allows for correlations?
TA1.1	T3.4	$w_j = (1/\widetilde{N}_{jy}) \times \{[\sum_b E_{jby}(1-E_{jby})]/[\sum_b (O_{jby}-E_{jby})^2]\}$	No
TA1.2	T3.4	$w_j = 1/\mathrm{Var}_{by}\{(O_{jby}-E_{jby})/[E_{jby}(1-E_{jby})/\widetilde{N}_{jby}]^{0.5}\}$	No
TA1.3	T3.4	$w_j = [m_{\mathrm{year},j}(m_{\mathrm{bin},j}-1)]/[\sum_{by} \widetilde{N}_{jy}(O_{jby}-E_{jby})^2/E_{jby}]$	No
TA1.4	T3.6	c_j estimated as model parameter	No
TA1.5	T3.6	$\mathrm{Var}_{by}\{(O_{jby}-E_{jby})/[E_{jby}(\widetilde{c}^2_{jby}+c_j^2)^{0.5}]\} = 1$	No
TA1.6	T3.8	$\sum_{by}\{(O_{jby}-E_{jby})^2/[E_{jby}(1/\widetilde{N}_{jy}+1/N_j)]\} = m_{\mathrm{year},j}(m_{\mathrm{bin},j}-1)$	No
TA1.7	T3.8[a]	$\frac{1}{N_j} = \{[\sum_{by}(O_{jby}-E_{jby})^2/E_{jby}]/[m_{\mathrm{year},j}(m_{\mathrm{bin},j}-1)]\} - \frac{1}{\widetilde{N}_j}$	No
TA1.8	T3.4	$w_j = 1/\mathrm{Var}_y[(\overline{O}_{jy}-\overline{E}_{jy})/(v_{jy}/\widetilde{N}_{jy})^{0.5}]$	Yes
TA1.9	T3.8	$\mathrm{Var}_y\{(\overline{O}_{jy}-\overline{E}_{jy})/[v_{jy}(1/\widetilde{N}_{jy}+1/N_j)]^{0.5}\} = 1$	Yes
TA1.10	T3.6	$\mathrm{Var}_y[(\overline{O}_{jy}-\overline{E}_{jy})/(S_2^2+S_1^2S_3-2S_1S_4)^{0.5}] = 1$[b]	Yes

Note: Var is the usual finite-sample variance function [for a sample $x_1,\ldots,x_n$, $\mathrm{Var}_k(x_k) = \sum_k (x_k-\bar{x})^2/(n-1)$, where $\bar{x}$ is the sample mean].

[a]With the additional assumption that $\widetilde{N}_{jy} = \widetilde{N}_j$ for all y.

[b]See text for definitions of S_1, S_2, S_3, and S_4.

the same equation as was used by McAllister and Ianelli (1997), viz. $\mathrm{Var}(O_{jby}-E_{jby}) = E_{jby}(1-E_{jby})/(w_j\widetilde{N}_{jy})$. This means that the errors can be standardized by setting $X_{jby} = [E_{jby}(1-E_{jby})/\widetilde{N}_{jby}]^{0.5}$, which makes $k_j = 1/w_j$, and leads to the equation for TA1.2. With lognormal errors, $\mathrm{Var}(O_{jby} - E_{jby}) = (E_{jby}c_{jby})^2$, so to standardize the errors we simply set $X_{jby} = E_{jby}c_{jby}$ and $k_j = 1$, which, together with the assumption $c^2_{jby} = \widetilde{c}^2_{jby} + c_j^2$ (weighting assumption T3.6), produces method TA1.5. As with method TA1.6, this method requires numerical solution.

Methods allowing for correlations

None of the methods discussed so far has allowed for the possibility of substantial correlations within a data set. Methods TA1.1, TA1.2, TA1.4, and TA1.5 assume there are no correlations; the remaining methods allow only for the small negative correlations that are associated with the fact that all proportions must sum to 1 in each year.

To develop methods that allow for substantial correlations I used an approach similar to that described above for TA1.2 and TA1.5, except that, following Pennington and Vølstad (1994), the error to be standardized was $(\overline{O}_{jy} - \overline{E}_{jy})$, where $\overline{O}_{jy} = \sum_b (x_b O_{jby})$ and $\overline{E}_{jy}$ (defined similarly) are the observed and expected mean ages (or lengths), and x_b is the age (or length) associated with the bth bin. If our composition data set has a multinomial distribution then $\mathrm{Var}(\overline{O}_{jy} - \overline{E}_{jy}) = \mathrm{Var}(\overline{O}_{jy}) = v_{jy}/N_{jy}$, where $v_{jy} = \sum_b (x_b^2 E_{bjy}) - \overline{E}^2_{jy}$ is the variance of the expected age (or length) distribution. Therefore, with $N_{jy} = w_j\widetilde{N}_{jy}$ (i.e., weighting assumption T3.4), $\mathrm{Var}(\overline{O}_{jy} - \overline{E}_{jy}) = v_{jy}/(w_j\widetilde{N}_{jy})$ and we can standardize our errors by setting $X_{jy} = (v_{jy}/\widetilde{N}_{jy})^{0.5}$, and $k_j = 1/w_j$. This produces stage-2 weighting method TA1.8, which can be solved explicitly for w_j. However, if $1/N_{jy} = 1/\widetilde{N}_{jy} + 1/N_j$ (i.e., weighting assumption T3.8), we set $X_{jy} = [v_{jy}(1/\widetilde{N}_{jy} + 1/N_j)]^{0.5}$ and $k_j = 1$, which leads to the equation for method TA1.9, which can be solved numerically for N_j.

When the composition data are assumed to have a lognormal error distribution, the calculations are a bit more complicated. Here, we need to impose the constraint (which is implicit with multinomial, but not lognormal, errors) that proportions sum to 1. Thus we treat $\overline{O}_{jy}$ as being equal to $\sum_b x_b O_{jby} / \sum_b x_b$, and calculate its variance using the standard approximation for the variance of a ratio of two random variables (eq. 10.17 of Stuart and Ord 1987), which produces $\mathrm{Var}(\overline{O}_{jy}) = S_2^2 + S_1^2 S_3 - 2S_1S_4$, where $S_1 = \sum_b x_b E^2_{jby}$, $S_2 = \sum_b (x_b c_{jby} E_{jby})^2$, $S_3 = \sum_b (c_{jby}E_{jby})^2$, and $S_4 = \sum_b x_b (c_{jby}E_{jby})^2$. Then we simply set $X_{jy} = \mathrm{Var}(\overline{O}_{jy})^{0.5}$ and $k_j = 1$, which leads to method TA1.10. Again, the equation for this method must be solved numerically (in this case we search for the value of c_j which makes the equation true, remembering that $c^2_{jby} = \widetilde{c}^2_{jby} + c_j^2$).

Methods requiring numerical solutions

Finally, a word about what I mean above by "numerical solution" (and a reassurance that this solution is not difficult to achieve). Consider method TA1.5, where we are solving for the weighting parameter c_j. To solve this equation we first evaluate the left-hand side of the equation for $c_j = 0$, 0.1, 0.2, …, continuing until a value greater than 1 is found. We might find, for example that for $c_j = 0$, 0.1, 0.2, and 0.3 the left-hand side of the equation has values 0.37, 0.56, 0.83, and 1.25, respectively, so that our solution lies between 0.2 and 0.3. Simple linear interpolation between the last two values produces the solution $c_j = 0.24$ $(= [0.2(1.25-1) + 0.3(1-0.83)]/(1.25-0.83))$.

References

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H., and Bian, R. 2008. CASAL (C++ algorithmic stock

Table B1. Formulae for calculating normalized residuals for the example negative log-likelihoods of Table 2.

Table 2 likelihood	Formula
T2.1A	$(2\lambda_{iy})^{0.5}(O_{iy} - E_{iy})$
T2.2	$(O_{iy} - E_{iy})/ (c_{iy}E_{iy})$
T2.3	$(O_{iy} - E_{iy})/\sigma_{iy}$
T2.4A	$\{[\log(O_{iy}/E_{iy})]/\sigma_{iy}\} + 0.5\sigma_{iy}$
T2.5	$[\log(O_{iy}/E_{iy})]/\sigma_{iy}$

assessment laboratory): CASAL User Manual v2.20–2008/02/14. NIWA Technical Report 130, National Institute of Water & Atmospheric Research, Wellington, New Zealand.

Dunn, A., and Hanchet, S.M. 2009. Assessment models for Antarctic toothfish (*Dissostichus mawsoni*) in the Ross Sea for the years 1997–98 to 2008–09. Report WG FSA 09/40, CCAMLR, Hobart, Australia.

McAllister, M.K., and 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. doi:10.1139/cjfas-54-2-284.

Pennington, M., and Vølstad, J.H. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics, **50**(3): 725–732. doi:10.2307/2532786.

Stuart, A., and Ord, J.K. 1987. Kendall's advanced theory of statistics. Vol. 1. Distribution theory. Charles Griffin and Company Limited, London, UK.

Appendix B: Calculation of SDNRs

For each data point, O_{iy}, a normalized residual is a number that indicates how different that observation is from the model's expected value (E_{iy}), and that has been transformed so that its distribution is approximately normal, with mean 0 and SD 1 (assuming the likelihood for the data point is correct). To calculate the SDNR (standard deviation of the normalized residuals) for a data set, we first calculate the normalized residual for each data point (using formulae like those in Table B1) and then use the usual formula to calculate the standard deviation of these residuals. No formulae are given for the composition-data likelihoods because SDNRs are inappropriate for this type of data (because of correlations).