APPENDIX 5. THE CASA LENGTH STRUCTURED STOCK ASSESSMENT MODEL

The stock assessment model described here is based on Sullivan et al.'s (1990) CASA model.[4] CASA is entirely length-based with population dynamic calculations in terms of the number of individuals in each length group during each year. Age is largely irrelevant in model calculations. Unlike many other length-based stock assessment approaches, CASA is a dynamic, non-equilibrium model based on a forward simulation approach. CASA incorporates a very wide range of data with parameter estimation based, in the broadest sense, on maximum likelihood. CASA can incorporate prior information about parameters such as survey catchability in a quasi-Bayesian fashion. The implementation described here was programmed in AD-Model Builder (Otter Research Ltd.).[5]

Population dynamics

Time steps in the model are the same as the time periods used to tabulate catch and other data. In principle, the accuracy of calculations improves as time steps in the model become shorter, but data considerations often limit time steps to years. In this description, time steps are referred to as "years" without loss of generality. If time steps are years, then instantaneous rates have units y^{-1}. The number of years in the model n_y is flexible and can be changed easily (e.g. for retrospective analyses), usually by making a single change to the input data file.

The definition of length groups (or length "bins") is a key element in the CASA model and length-structured stock assessment modeling in general. Length bins are identified by their lower bound. With 10 mm length bins, for example, the 20 mm size bin includes individual 20-29.9 mm. Calculations requiring information about length (e.g. length-weight) use the mid-length ℓ_j of each bin.

In the current implementation, the user must specify the size of length bins (L_{bin}) in the data and model, the minimum size (L_{min}) at the lower bound of the first length bin in the data and model, and the maximum asymptotic length (L_∞). Based on these specifications, the model determines the number (n_L) of length bins to include in modeling. The last bin is a "plus-group" containing individuals L_∞ and larger. The number of length groups in catch at length and other data should be $\geq n_L$. Based on user specifications, the program takes care adjusting the original data to the length groups used in the model.

[4] Original programming in AD-Model Builder by G. Scott Boomer and Patrick J. Sullivan (Cornell University), who bear no responsibility for errors in the current implementation.

[5] AD-Model Builder can be used to calculate variances for any estimated or calculated quantity in a stock assessment model, based on the Hessian matrix with "exact" derivatives and the delta method. Experience with other models (e.g. Overholtz et al., 2004) suggests that variances estimates from AD-Model Builder, which consider the variance of all model parameters, are similar to variances calculated by the common method of bootstrapping survey abundance data.

Growth

Although age is not considered, Von Bertalanffy growth models are implicit in several of the configurations of the CASA model. The growth parameter L_∞ is not estimable because it is used in defining length bins prior to the parameter estimation phase.[6] The von Bertalanffy growth parameter t_0 is not estimable because it is irrelevant in length-based models that predict growth during a year based on the von Bertalanffy growth parameter K, L_∞ and size at the beginning of the year.

At the beginning of the year, scallops in each size group grow (or not) based on growth terms *P(b,a)* that measure the probability that a surviving individual that starts in bin *a* will grow to bin *b* by the beginning of the next year (columns index initial size and rows index subsequent size). Growth probabilities do not include any adjustments for mortality. In the CASA model, growth occurs immediately at the beginning of each year and the model assumes that no growth occurs during the year.

Growth probabilities depend on growth increments because:

$$L_2 = L_1 + \iota$$

where L_1 is the starting length, L_2 is length after one year of growth and ι is the growth increment. Following Sullivan et al. (1990), and for simplicity, growth probabilities are calculated assuming that all individuals start at the middle of their original length bin ℓ_a, and then grow to sizes that cover the whole range of each possible subsequent size bin. Thus:

$$P(b,a) = \int_{j=\ell_b - L_{bin}/2}^{\ell_b + L_{bin}/2} P(j \mid \ell_a)\partial j = \aleph(\ell_b + L_{bin}/2 \mid \ell_a) - \aleph(\ell_b - L_{bin}/2 \mid \ell_a)$$

where $P(j \mid \ell_a)$ is the probability of increment j for an individual originally in bin a (at mid-length ℓ_a). $\aleph(a \mid \ell_a)$ is the initial size-specific cumulative distribution function for growth increments. In CASA, cumulative distributions for growth increments are computed by numerical integration based on Simpson's rule (Press et al., 1990) and a user-specified number of steps per bin. The user can change the number of steps to balance the accuracy of the calculation against time required for growth calculations.

Growth probabilities *P(b,a)* are calculated in CASA by one of four options. Option 1 is similar to Sullivan et al.'s (1990) approach in that growth probabilities are calculated by numerical integration assuming that increments follow gamma distributions. The gamma distributions for growth increments are starting size-specific

[6] "Estimable" means a potentially estimable parameter that is specified as a variable that may be estimated in the CASA computer program. In practice, estimability depends on the available data and other factors. It may be necessary to fix certain parameters at assumed fix values or to use constraints of prior distributions for parameters that are difficult to estimate, particularly if data are limited.

and are specified in terms of mean increments and CV's. Mean increments $\bar{\imath}_a$ are from the von Bertalanffy growth curve:

$$\bar{\imath}_a = \left(L_\infty - \ell_a\right)\left(1 - e^{-K}\right)$$

where $K=e^{\chi}$ is the von Bertalanffy growth coefficient and χ is an estimable parameter.[7] Under Option 1, CVs are a log-linear function of length:

$$CV_L = e^{\kappa+\lambda L}$$

where κ and γ are estimable parameters. Sullivan et al. 1990 assumed constant CV's for growth. This implementation of the CASA model includes the special case of constant CV's when λ=0.

Option 2 constructs a transition matrix directly from size-specific annual growth data (i.e. data records consisting of starting length, length after one year and number of observations). Under Option 2:

$$P(b,a) = \frac{n(b \mid a)}{\sum_{j=a}^{n_L} n(j \mid a)}$$

where $n(b|a)$ is the number of individuals that started at size a and grew to size b after one year.

Under option 3, mean increments are from the von Bertalanffy growth curve as in option 1, but with length-specific CVs (and other model parameters) estimated in the model based on growth increments and other data (see below for goodness of fit calculations). Under option 3, the von Bertalanffy growth parameter K, which describes mean growth, and parameters for variance in growth (κ and γ) are estimable. Option 4 uses a constant, user-specified transition matrix provided as data to the model.

Growth calculations based on assumed gamma distributions (Sullivan et al. 1990) might be unrealistic for some species because the gamma distribution predicts growth increments of zero to infinity. Therefore, with options 1-3, the user may specify minimum and maximum growth increments for each size. Probabilities from truncated gamma distributions for growth increments between the minimum and maximum values are normalized to sum to one before use in population dynamics calculations. Size bins outside those specified are ignored in all model calculations.

[7] Most intrinsically positive or intrinsically negative parameters are estimated in log scale to ensure estimates do not change sign, and to enhance statistical properties of estimates.

Abundance, recruitment and mortality

Population abundance in each length bin during the first year of the model is:

$$N_{1,L} = N_1 \pi_{1,L}$$

where L is the size bin, and $\pi_{1,L}$ is the initial population length composition expressed as proportions so that $\sum_{L=1}^{n_L} \pi_L = 1$. $N_1 = e^{\eta}$ is total abundance at the beginning of the first modeled year and η is an estimable parameter. It is not necessary to estimate recruitment in the first year because recruitment is implicit in the product of N_1 and π_L. The current implementation of CASA takes the initial population length composition as data supplied by the user.

Abundance at length in years after the first is calculated:

$$\vec{N}_{y+1} = P\left(\vec{N}_y \otimes \vec{S}_y\right) + \vec{R}_{y+1}$$

where $\vec{N}_y$ is a vector (length n_L) of abundance in each length bin during year y, P is the matrix (n_L x n_L) of growth probabilities $P(b,a)$, $\vec{S}_y$ is a vector of length- specific survival fractions for year y, $\otimes$ is for the element-wise product , and $\vec{R}_y$ is a vector holding length-specific abundance of new recruits at the beginning of year y.

Survival fractions are:

$$S_{y,L} = e^{-Z_{y,L}} = e^{-\left(M + F_{y,L}\right)}$$

where $Z_{y,L}$ is the total instantaneous mortality rate. The natural mortality rate $M=e^{\varpi}$ (ϖ estimable) is the same for all length groups in all years. Length-specific fishing mortality rates are $F_{y,L} = F_y s_{y,L}$ where $s_{y,L}$ is the size-specific selectivity for the fishery in year y (scaled to a maximum of one at fully recruited size groups), and F_y is the fishing mortality rate on fully selected individuals.[8] Fully recruited fishing mortality rates are $F_y = e^{\phi + \delta_y}$ where ϕ is an estimable parameter for the log of the geometric mean of fishing mortality in all years, and δ_y is an estimable "dev" parameter.[9]

Given abundance in each length group, natural mortality, and fishing mortality,

[8] In this context, "selectivity" describes the combined effects of all factors that affect length composition of catch or landings. These factors include gear selectivity, spatial overlap of the fishery and population, size-specific targeting, size-specific discard, etc.

[9] Dev parameters are a special data type for estimable parameters in AD-Model Builder. Each set of dev parameters (e.g. for all recruitments in the model) is constrained to sum to zero. Because of the constraint, the sums $\phi + \delta_y$ involving n_y+1 terms amount to only n_y parameters.

predicted fishery catch-at-length in numbers is:

$$C_{y,L} = \frac{F_{y,L}\left(1 - e^{-Z_{y,L}}\right)N_{L,y}}{Z_{y,L}}$$

Total catch number during each year is $C_y = \sum_{j=1}^{n_L} C_{y,L}$. Note that, because the catches are in effect assumed to be taken at the beginning of the year, model catches (by weight) will tend to be biased low, especially during years when mostly smaller scallops were taken.

Recruitment (the sum of new recruits in all length bins) at the beginning of each year after the first is calculated based on estimable parameters that measure annual deviations γ_y from the log-scale geometric mean ρ:

$$Ry = e^{\rho+\gamma_y}$$

Proportions of recruits in each length group are calculated based on a standard beta distribution *B(w,r)* over the first n_r length bins. Proportions of new recruits in each size group are the same from year to year. Beta distribution coefficients must be larger than zero and are calculated $w=e^{\omega}$ and $r=e^{\rho}$, where ω and ρ are estimable parameters.

Population summary variables

Total abundance at the beginning of the year is the sum of abundance at length $N_{y,L}$ at the beginning of the year. Average annual abundance is:

$$\overline{N}_{y,L} = N_{y,L}\frac{1 - e^{-Z_{y,L}}}{Z_{y,L}}$$

The current implementation of the NC model assumes that weight-at-length is the same for the stock and fishery and a single set of length-weight conversion parameters is used in all calculations. For example, total stock biomass is:

$$B_y = \sum_{L=1}^{n_L} N_{y,L} w_L$$

where w_L is weight at length computed at the midpoint of each length bin using the length-weight relationship specified by the user. Total catch weight is:

$$W_y = \sum_{L=1}^{n_L} C_{y,L} w_L$$

F_y estimates for two years are comparable if fishery selectivity in the model was the same in both years. A simpler exploitation index is calculated for use when fishery selectivity changes over time:

$$U_y = \frac{C_y}{\sum_{j=x}^{n_L} N_{y,L}}$$

where x is a user-specified length bin (usually at or below the first bin that is fully selected during all fishery selectivity periods). U_y exploitation indices from different years with different selectivity patterns may be relatively comparable if w is chosen carefully.

Surplus production during each year of the model can be computed approximately from biomass and catch estimates (Jacobson et al., 2002):

$$P_t = B_{t+1} - B_t + \delta C_t$$

where δ is a correction factor that adjusts catch weight to population weight at the beginning of the next year by accounting for mortality and growth. The adjustment factor depends strongly on the rates for growth and natural mortality and only weakly on the natural mortality rate. In the absence of a direct estimate, useful calculations can be carried out assuming δ=1.

Fishery and survey selectivity

The current implementation of CASA includes six options for calculating fishery and survey selectivity patterns. Fishery selectivity may differ among "fishery periods" defined by the user. Selectivity patterns that depend on length are calculated using lengths at the mid-point of each bin (ℓ). After initial calculations (described below), selectivity curves are rescaled to a maximum value of one.

Option 1 is a flat with s_L=1 for all length bins. Option 2 is an ascending logistic curve:

$$s_{y,\ell} = \frac{1}{1 + e^{A_Y - B_Y \ell}}$$

Option 3 is an ascending logistic curve with a minimum asymptotic minimum size for small size bins on the left.

$$s_{y,\ell} = \left(\frac{1}{1 + e^{A_Y - B_Y \ell}} \right) \left(1 - D_y\right) + D_y$$

Option 4 is a descending logistic curve:

$$s_{y,\ell} = 1 - \frac{1}{1 + e^{A_Y - B_Y \ell}}$$

Option 5 is a descending logistic curve with a minimum asymptotic minimum size for large size bins on the right:

$$s_{y,\ell} = \left(1 - \frac{1}{1 + e^{A_Y - B_Y \ell}}\right)\left(1 - D_y\right) + D_y$$

Option 6 is a double logistic curve used to represent "domed-shape" selectivity patterns with highest selectivity on intermediate size groups:

$$s_{y,\ell} = \left(\frac{1}{1 + e^{A_Y - B_Y \ell}}\right)\left(1 - \frac{1}{1 + e^{D_Y - G_Y \ell}}\right)$$

The coefficients for selectivity curves A_Y, B_Y, D_Y and G_Y carry subscripts for time because they may vary between fishery selectivity periods defined by the user. All options are parameterized so that the coefficients A_Y, B_Y, D_Y and G_Y are positive. Under options 3 and 5, D_y is a proportion that must lie between 0 and 1.

Depending on the option, estimable selectivity parameters may include α, β, δ and γ. For options 2, 4 and 6, $A_Y = e^{\alpha_Y}$, $B_Y = e^{\beta_Y}$, $D_Y = e^{\delta_Y}$ and $G_Y = e^{\gamma_Y}$. Options 3 and 5 use the same conventions for A_Y and B_Y, however, the coefficient D_Y is a proportion estimated as a logit-transformed parameter (i.e. $\delta_Y = \ln[D_Y/(1-D_y)]$) so that:

$$D_Y = \frac{e^{\delta_Y}}{1 + e^{\delta_Y}}$$

The user can choose, independently of all other parameters, to either estimate each fishery selectivity parameter or to keep it at its initial value. Under Option 2, for example, the user can estimate the intercept α_Y, while keep the slope β_Y at its initial value.

Tuning and goodness of fit

There are two steps in calculating the negative log likelihood (NLL) used to measure how well the model fits each type of data. The first step is to calculate the predicted values for data. The second step is to calculate the NLL of the data given the predicted value. The overall goodness of fit measure for the model is the weighted sum of NLL values for each type of data and each constraint:

$$\Lambda = \sum \lambda_j L_j$$

where λ_j is a weighting factor for data set j (usually λ_j=1, see below), and L_j is the NLL for the data set. The NLL for a particular data is itself is usually a weighted sum:

$$L_j = \sum_{i=1}^{n_j} \psi_{j,i} L_{j,i}$$

where n_j is the number of observations, $\psi_{j,i}$ is an observation-specific weight (usually $\psi_{j,i}$ =1, see below), and $L_{j,i}$ is the NLL for a single observation.

Maximum likelihood approaches reduce the need to specify *ad-hoc* weighting factors (λ and ϕ) for data sets or single observations, because weights can often be taken from the data (e.g. using CVs routinely calculated for bottom trawl survey abundance indices) or estimated internally along with other parameters. In addition, robust maximum likelihood approaches (see below) may be preferable to simply down-weighting an observation or data set. However, despite subjectivity and theoretical arguments against use of *ad-hoc* weights, it is often useful in practical work to manipulate weighting factors, if only for sensitivity analysis or to turn an observation off entirely. Observation specific weighting factors are available for most types of data in the CASA model.

Missing data

Availability of data is an important consideration in deciding how to structure a stock assessment model. The possibility of obtaining reliable estimates will depend on the availability of sufficient data. However, NLL calculations and the general structure of the CASA model are such that missing data can usually be accommodated automatically. With the exception of catch data (which must be supplied for each year, even if catch was zero), the model calculates that NLL for each datum that is available. No NLL calculations are made for data that are not available and missing data do not generally hinder model calculations.

Likelihood kernels

Log likelihood calculations in the current implementation of the CASA model use log likelihood "kernels" or "concentrated likelihoods" that omit constants. The constants can be omitted because they do not affect slope of the NLL surface, final point estimates for parameters or asymptotic variance estimates.[10]

For data with normally distributed measurement errors, the complete NLL for one observation is:

$$L = \ln(\sigma) + \ln\left(\sqrt{2\pi}\right) + 0.5\left(\frac{x-u}{\sigma}\right)^2$$

The constant $\ln\left(\sqrt{2\pi}\right)$ can always be omitted. If the standard deviation is known or assumed known, then $\ln(\sigma)$ can be omitted as well because it is a constant that does not affect derivatives. In such cases, the concentrated NLL is:

[10] Likelihood kernels in the present implementation prevent use of AD-Model Builder's MCMC algorithms for Bayesian statistical approaches.

$$L = 0.5\left(\frac{x-\mu}{\sigma}\right)^2$$

If there are N observations with possible different variances (known or assumed known) and possibly different expected values:

$$L = 0.5\sum_{i=1}^{N}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2$$

If the standard deviation for a normally distributed quantity is not known and is estimated (implicitly or explicitly) by the model, then one of two equivalent calculations is used. Both approaches assume that all observations have the same variance and standard deviation. The first approach is used when all observations have the same weight in the NLL:

$$L = 0.5N\ln\left[\sum_{i=1}^{N}(x_i-u)^2\right]$$

The second approach is equivalent but used when the weights for each observation (w_i) may differ:

$$L = \sum_{i=1}^{N} w_i\left[\ln(\sigma)+0.5\left(\frac{x_i-u}{\sigma}\right)^2\right]$$

In the latter case, the maximum likelihood estimator:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\hat{x})^2}{N}}$$

(where $\hat{x}$ is the average or predicted value from the model) is used explicitly for σ. The maximum likelihood estimator is biased by $N/(N-d_f)$ where d_f is degrees of freedom for the model. The bias may be significant for small sample sizes, which are common in stock assessment modeling, but d_f is usually unknown.

If data x have lognormal measurement errors, then ln(x) is normal and L is calculated as above. In some cases it is necessary to correct for bias in converting arithmetic scale means to log scale means (and *vice-versa*) because $\bar{x} = e^{\bar{\chi}+\sigma^2/2}$ where χ=ln(x). It is often convenient to convert arithmetic scale CVs for lognormal variables to log scale standard deviations using $\sigma = \sqrt{\ln\left(1+CV^2\right)}$.

For data with multinomial measurement errors, the likelihood kernel is:

$$L = n\sum_{i=1}^{n} p_i \ln(\theta_i) - K$$

where n is the known or assumed number of observations (the "effective" sample size), p_i is the proportion of observations in bin i, and θ_i is the model's estimate of the probability of an observation in the bin. The constant K is used for convenience to keep L to a manageable number of digits. It measures the lowest value of L that could be achieved if the data fit matched the model's expectations exactly:

$$K = n\sum_{i=1}^{n} p_i \ln(p_i)$$

For data x that have measurement errors with expected values of zero from a gamma distribution:

$$L = (\gamma - 1)\ln\left(x/\beta\right) - x/\beta - \ln(\beta)$$

where $\beta>0$ and $\gamma>0$ are gamma distribution parameters in the model. For data that lie between zero and one with measurement errors from a beta distribution:

$$L = (p-1)\ln(x) + (q-1)\ln(1-x)$$

where $p>0$ and $q>0$ are parameters in the model.

In CASA model calculations, distributions are usually described in terms of the mean and CV. Normal, gamma and beta distribution parameters can be calculated mean and CV by the method of moments. Means, CV's and distributional parameters may, depending on the situation, be estimated in the model or specified by the user.

Robust methods

"Robust" maximum likelihood calculations are available for noisy data in the CASA model that might be assumed otherwise to have normally distributed measurement errors. Robust likelihood calculations assume that measurement errors are from a Student's t distribution with user-specified degrees of freedom d_f. Degrees of freedom are specified independently for each observation so that robust calculations can be carried out for as many (or as few) cases as required. The t distribution is similar to the normal distribution for $d_f \geq 30$. As d_f are reduced, the tails of the t distribution become fatter so that small observations seem more probable (have higher probability) and have less effect on model estimates. If $d_f=0$, then measurement errors are assumed in the model to be normally distributed.

The first step in robust NLL calculations is to standardize the measurement error residual $t = (x - \bar{x})/\sigma$ based on the mean and standard deviation. Then:

$$L = \ln\left(1 + \frac{t^2}{d_f}\right)\left(1 - \frac{1 - d_f}{2}\right) - \frac{\ln\left(d_f\right)}{2}$$

Catch weight data

In the current version of the CASA model, catch data are for a single or "composite" fishery. The terms "catches" and "landings" are used interchangeably in the current version because discard and non-landed fishery induced mortality are not distinguished. In the current version, total catch and must be specified in units of weight. Ideally, catch data should include all fishery-induced mortality and fishery length composition data (if available) should be represent the size distribution of all individuals that suffered fishery-induced mortality.

Catch data are assumed to have normally distributed measurement errors with a user specified CV. The standard deviation for catch weight in a particular year is $\sigma_Y = \kappa \hat{C}_y$ where "^" indicates that the variable is a model estimate. The standardized residual used in computing NLL for a single catch observation and in making residual plots is $r_Y = \left(C_Y - \hat{C}_Y\right) / \sigma_Y$.

Fishery length composition data

Data describing numbers or relative numbers of individuals at length in catch data (fishery catch-at-length) are modeled as multinomial proportions $c_{y,L}$:

$$c_{y,L} = \frac{C_{y,L}}{\sum_{j=1}^{n_L} C_{y,j}}$$

The NLL for the observed proportions in each year is computed based on the kernel for the multinomial distribution, the model's estimate of proportional catch-at-length $\left(\hat{c}_Y\right)$ and an estimate of effective sample size ${}^{C}N_Y$ supplied by the user. Care is required in specifying effective sample sizes, because catch-at-length data typically carry substantially less information than would be expected based on the number of individuals measured (Fournier and Archibald, 1982; Pennington et al., 2002). Typical conventions make ${}^{c}N_Y \le 200$ or set ${}^{C}N_Y$ equal to the number of trips or tows sampled. Effective sample sizes are sometimes chosen based on goodness of fits in preliminary model runs (Methot, 2000; Butler et al., 2003).

Standardized residuals are not used in computing NLL fishery length composition data. However, approximate standardized residuals $r_y = \left(c_{y,L} - \hat{c}_{y,L}\right) / \sigma_{y,L}$ with standard deviations $\sigma_{y,L} = \sqrt{\hat{c}_{y,L}\left(1 - \hat{c}_{y,L}\right) / {}^{c}N_y}$ based on the theoretical variance for proportions are computed for use in making residual plots.

Survey index data

In CASA model calculations, "survey indices" are data from any source that reflect relative proportional changes in annual abundance or biomass over time. In the current implementation of the CASA model, survey indices are assumed to be linear indices of abundance or biomass so that changes in the index (apart from measurement error) are assumed due to proportional changes in the population. Nonlinear commercial catch rate data are handled separately (see below).

In general, survey index data give one number that summarizes relative abundance for a wide range of length bins. Catch at length data from surveys are handled separately (see below). For example, a survey index might consist of stratified mean numbers per tow for all size bins in a bottom trawl survey carried out over a series of years, with one observation of the index per year of sampling.

NLL calculations for survey indices use predicted values calculated:

$$\hat{I}_{k,y} = q_k A_{k,y}$$

where q_k is a scaling factor for survey index k, and $A_{k,y}$ is abundance or biomass available to the survey. Scaling factors are calculated $q_s = e^{\varpi_s}$ where ϖ_s is estimable and survey-specific. Available abundance is:

$$A_{k,y} = \sum_{L=first_k}^{last_k} s_{k,L} N_{y,L} e^{-Z_{y,L}\tau_{k,y}}$$

where $s_{k,L}$ is size-specific selectivity of the survey, $\tau_{k,y}$=$J_{k,y}$/365 where $J_{k,y}$ is the mean Julian date of the survey, and $e^{-Z_y\tau_{k,y}}$ is a correction for mortality prior to the survey. Options and procedures for estimating survey selectivity patterns are the same as for fishery selectivity patterns, but survey selectivity patterns are not allowed to change over time. Available biomass is calculated in the same way except that body weights w_L are included in the product on the right hand side.

The range of lengths ($first_k \geq 1$ to $last_k \leq n_L$) included in the calculation of $A_{k,y}$ is specified by the user for each survey. In addition, the user specifies whether $first_k$ and $last_k$ are plus-groups meant to contain smaller or larger individuals.

NLL calculations for survey index data assume that log scale measurement errors are either normally distributed (default approach) or from a *t* distribution (robust estimation approach). In either case, log scale measurement errors are assumed to have mean zero and log scale standard errors either estimated internally by the model or calculated from the arithmetic CVs supplied with the survey data.

The standardized residual used in computing NLL for one survey index

observation is $r_{k,y} = \ln\left(I_{k,y} / \hat{I}_{k,y}\right) / \sigma_{k,y}$ where $I_{k,y}$ is the observation. The standard deviations $\sigma_{k,y}$ will vary among surveys and years if CVs are used to specify the variance of measurement errors. Otherwise a single standard deviation is estimated internally for the survey as a whole.

Survey length composition data

NLL calculations for survey length composition data are roughly analogous to calculations for fishery length composition data, except that measurement errors in length data can be modeled explicitly. Survey length composition data represent a sample from the true population length composition which is modified by survey selectivity, sampling errors (due to having a limited number of tows) and, if applicable, errors in recording length data (i.e. errors in observations to size bins). For example, with errors in length measurements, individuals belonging to length bin j, might be mistakenly assigned to adjacent length bins j-2, j-1, j+1 or j+2. Well-tested methods for dealing with errors in length data can be applied if some information about the distribution of the errors is available (e.g. Methot 2000).

Survey length composition data are treated as multinomial proportions calculated:

$$i_{k,y,L} = \frac{n_{k,y,L}}{\sum_{j=first_k}^{last_k} n_{k,y,j}}$$

The model's estimate of length composition for the population available to the survey is:

$$A_{k,y,L} = \frac{s_{k,L} N_{y,L} e^{-Z_{y,j}\tau_{k,y}}}{\sum_{first_k}^{last_k} s_{k,j} N_{y,j} e^{-Z_{y,j}\tau_{k,y}}}$$

The expected length composition $\vec{A}'_{k,y}$ for survey catches, including length measurement errors is:

$$\vec{A}'_{k,y} = \vec{A}_{k,y} E_k$$

where E_k is an error matrix that simulates errors in collecting length data by mapping true length bins in the model to observed length bins in the data.

The error matrix E_k has n_L rows (one for each true length bin) and n_L columns (one for each possible observed length bin). For example, row k and column j of the error matrix gives the conditional probability P(k|j) of being assigned to bin k, given that an individual actually belongs to bin j. More generally, column j gives the probabilities that an individual actually belonging to length bin j will be recorded as being in length

bins j-2, j-1, j, j+1, j+2 and so on. The columns of E_k add to one to account for all possible outcomes in assigning individuals to observed length bins.

In CASA, the probabilities in the error matrix are computed from a normal distribution with mean zero and $CV = e^{\pi_k}$, where π_k is an estimable parameter. The normal distribution is truncated to cover a user-specified number of observed bins.

The NLL for observed proportions at length in each survey and year is computed with the kernel for a multinomial distribution, the model's estimate of proportional survey catch-at-length $\left(\hat{i}_{k,y,L}\right)$ and an estimate of effective sample size ${}^{I}N_{Y}$ supplied by the user. Standardized residuals for residual plots are computed as for fishery length composition data.

LPUE data

Commercial landings per unit of fishing effort (LPUE) data are modeled in the current implementation of the CASA model as a linear function of average biomass available to the fishery, and as a nonlinear function of average available abundance. The nonlinear relationship with abundance is meant to reflect limitations in "shucking" capacity for sea scallops.[11] Briefly, tows with large numbers of scallops require more time to sort and shuck and therefore reduce LPUE from fishing trips when abundance is high. The effect is exaggerated when the catch is composed of relatively small individuals. In other words, at any given level of stock biomass, LPUE is reduced as the number of individuals in the catch increases or, equivalently, as the mean size of individuals in the catch is reduced.

Average available abundance in LPUE calculations is:

$$ {}^{a}\overline{N}_{y} = \sum_{L=1}^{n_L} s_{y,L}\overline{N}_{y,L} $$

and average available biomass is:

$$ {}^{a}\overline{B}_{y} = \sum_{L=1}^{n_L} s_{y,L} w_L \overline{N}_{y,L} $$

Predicted values for LPUE data are calculated:

$$ \hat{L}_{y} = \frac{{}^{a}\overline{B}_{y}\eta}{\sqrt{\varphi^2 + {}^{a}\overline{N}^{2}{}_{y}}} $$

Measurement errors in LPUE data are assumed normally distributed with standard deviations $\sigma_y = CV_y\hat{L}_y$. Standardized residuals are $r_y = \left(L_y - \hat{L}_y\right)\big/\sigma_y$.

[11] D. Hart, National Marine Fisheries Service, Northeast Fisheries Science Center, Woods Hole, MA, pers. comm.

Growth data

Growth data in CASA consist of records giving initial length, length after one year of growth, and number of corresponding observations. Growth data may be used to help estimate growth parameters that determine the growth matrix $\mathbf{P}$. The first step is to convert the data for each starting length to proportions:

$$P(b,a) = \frac{n(b,a)}{\sum_{j=n_L-b+1}^{n_L} n(j,a)}$$

where *n(b,a)* is the number of individuals starting at size *that* grew to size *b* after one year. The NLL is computed assuming that observed proportions *p(a|b)* at each starting size are a sample from a multinomial distribution with probabilities given by the corresponding column in the models estimated growth matrix $\mathbf{P}$. The user must specify an effective sample size ${}^{P}N_j$ based, for example, on the number of observations in each bin or the number of individuals contributing data to each bin. Observations outside bin ranges specified by the user are ignored. Standardized residuals for plotting are computed based on the variance for proportions.

Survey gear efficiency data

Survey gear efficiency for towed trawls and dredges is the probability of capture for individuals anywhere in the water column or sediments along the path swept by the trawl. Ideally, the area surveyed and the distribution of the stock coincide so that:

$$I_{k,y} = q_k A_{k,y}$$

$$q_k = \frac{A e_k}{a_k}$$

$$e_k = \frac{a_k q_k}{A_k}$$

where A is the area of the stock, a_k is the area swept during one tow and $0<e_k\leq 1$ is efficiency of the survey gear. Efficiency estimates from studies outside the CASA model may be used as prior information in CASA. The user supplies the mean and CV for the prior estimate of efficiency, along with estimates of A_k and a_k. Then, at each iteration of the model, the gear efficiency implied by the current estimate of q_k is computed. The model then calculates the NLL of the implied efficiency estimate assuming it was sampled from a beta distribution with the user-specified mean and CV. Alternatively, in Bayesian jargon, the prior probability of the implied efficiency estimate is computed and added to the overall objective function.

Care should be taken in using prior information from field studies designed to

estimate survey gear efficiency. Field studies usually estimate efficiency with respect to individuals on the same ground (e.g. by sampling the same grounds exhaustively or with two types of gear). It seems reasonable to use an independent efficiency estimate and the corresponding survey index to estimate abundance in the area surveyed. However, stock assessment models are usually applied to the entire stock, which is probably distributed over a larger area than the area covered by the survey. Thus the simple abundance calculation based on efficiency and the survey index will be biased low for the stock as a whole.

Maximum fishing mortality rate

Stock assessment models occasionally estimate absurdly high fishing mortality rates because abundance estimates are too small. The NLL component used to prevent this potential problem is:

$$L = \lambda \sum_{t=0}^{N} \left(d_t^2 + q^2 \right)$$

where:

$$d_t = \begin{cases} Ft - \Phi \; if \; Ft > \Phi \\ 0 \; otherwise \end{cases}$$

and

$$q_t = \begin{cases} \ln\left(Ft / \Phi\right) if \; Ft > \Phi \\ 0 \; otherwise \end{cases}$$

with the user-specified threshold value Φ set larger than the largest value of F_t that might possibly be expected (e.g. $\Phi=3$). The weighting factor λ is normally set to a large value (e.g. 1000).