APPENDIX B10: Technical documentation for the CASA Length Structured Stock Assessment Model[12]

The stock assessment model described here is based on Sullivan et al.'s (1990) CASA model[13] with a number of additional features. Many aspects are similar to aspects of the Stock Synthesis Model (Methot 2000) although CASA assumes a single set of life history characteristics within a single stock area. 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 although "effective age" (years since recruitment to the model) calculations have been implemented experimentally. 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 on maximum likelihood. CASA can incorporate prior information and constraints on parameters such as survey catchability in a quasi-Bayesian fashion. The implementation described here was programmed in AD-Model Builder (Otter Research Ltd.).[14]

Population dynamics

Time steps in the model are years, which are also used to tabulate catch and other data. Recruitment occurs at the beginning of each time step. 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) by making a single change to the input data file. Millimeters are the units for length data. Length-weight relationships should generally convert millimeters to grams. The units for catch and biomass are usually metric tones. Model input data include a scalar that is used to convert the units for length-weight parameters (e.g. grams) to the units of the biomass estimates and landings data (e.g. mt).

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 in CASA by their lower bound. Calculations requiring information about length (e.g. length-weight) use the mid-length ℓ_j of each bin. The user specifies the first length included in model calculations (L_{min}) and the size of length bins (L_{bin}). Based on these specifications, the model determines the number of length bins to be used in modeling as $n_L = 1 + \text{int}\left[(L_\infty - L_{\min})/L_{bin}\right]$, where L_∞ is maximum asymptotic size based on a von Bertalanffy growth curve supplied by the user and int[x] is the integer part of x. The last length bin in the model is always a "plus-group" containing individuals L_∞ and larger. Specifications for length data used in tuning the model are entirely separate (see below).

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

[12] Documentation last updated on May 11, 2007 as file CASA-Appendix-NC-describe57.doc.

[13] 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.

[14] 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.

is used in defining length bins prior to the parameter estimation phase and in determining the largest "plus-group".[15] 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 terms in the growth transition matrix *P(b,a)* which measures 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. When growth increments are based on parametric probability distributions (e.g. gamma distributions following Sullivan et al. 1990), probability calculations assume that 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 size-specific cumulative distribution function for growth increments. In parametric growth model calculations, 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 parametric gamma distributions. The gamma distributions for growth increments are initial size dependent and are specified in terms of mean increments and CV's. Mean increments $\bar{\iota}_a$ are from the von Bertalanffy growth curve:

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

where $K=e^{\chi}$ is the von Bertalanffy growth coefficient and χ is an estimable parameter.[16] Under Option 1, CVs are a log-linear function of length:

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

[15] "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.

[16] 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.

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 is nonparametric and 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.

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:

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

where $\vec{N}_y$ is a vector (length n_L) of abundance in each length bin during year y, $\mathbf{P}$ 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, $\circ$ 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:

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

where $Z_{y,L}$ is the total instantaneous mortality rate and $M_{y,L}$ is the instantaneous rate for natural mortality (see below). Length-specific fishing mortality rates are $F_{y,L}= F_y s_{y,L}$ where $s_{y,L}$ is the size-specific selectivity[17] for fishing in year y (scaled to a maximum of one at fully recruited size groups), F_y is the fishing mortality rate on fully selected individuals. 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.[18] The instantaneous rate for "incidental" mortality ($I_{y,L}$) accounts for mortality due to contact with the fishing gear that does not result in any catch on deck (see below).[19] The degree of variability in dev parameters for fishing mortality, natural mortality and for other variables can be controlled using variance constraints described below.

Natural mortality rates $M_{y,L} = u_L e^{\zeta+\xi_y}$ may vary from year to year and by length. Variability among length groups is based on a user-specified vector $\vec{u}$ that describes the relative natural mortality rate for each length group in the model. The user supplies a value for each length group which the model rescales so that the average of all of the values is one (i.e. $\vec{u}$ is set by the user and cannot be estimated). Temporal variability in natural morality rates are modeled in the same manner as temporal variability in fishing mortality. In particular, ζ is an estimable parameter measuring the mean log natural mortality rate during all years and ξ_y is an estimable year-specific dev parameter. Several approaches are available for estimating natural mortality parameters (i.e. natural mortality covariates and surveys that measure numbers of dead individuals, see below).

Incidental mortality $I_{y,L} = F_y u_L i$ is the product of fully recruited fishing mortality (F_y, a proxy for effective fishing effort, although nominal fishing effort might be a better predictor of incidental mortality), relative incidental mortality at length (u_L) and a scaling parameter i, which is supplied by the user and not estimable in the model. Mortality at length is supplied by the user as a vector ($\vec{u}$) containing a value for each length group in the model. The model rescales the relative mortality vector so that the mean of the series is one.

Given abundance in each length group, natural mortality, and fishing mortality, 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 $C_y = \sum_{j=1}^{n_L} C_{y,L}$. Catch data (in weight, numbers or as length composition data) are understood to include landings (L_y) and discards (d_y) but to exclude losses to incidental mortality (i.e. C_y=L_y+d_y).

Discard data are supplied by the user in the form of discarded biomass in each year or a discard rate for each year (or a combination of biomass levels and rates). It is important to remember that discard rates in CASA are defined the ratio of discards to landings (d/L). The user may also specify a mortal discard fraction between zero and one if not discards are expected to die. If the discard fraction is less than one, then the discarded biomass and

[17] 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.

[18] 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.

[19] . See the section on per recruit modeling below for formulas used to relate catch, landings and indicental mortality.

discard rates in the model are reduced correspondingly. See the section on per recruit modeling below for formulas used to relate catch, landings and incidental mortality.

Recruitment (the sum of new recruits in all length bins) at the beginning of each year after the first is calculated:

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

where ρ is an estimable parameter that measures the geometric mean recruitment and the γ_y are estimable dev parameters that measure interannual variability in recruitment. As with natural mortality devs, a variance constraint can be used to help estimate recruitment deviations (see below).

Proportions of recruits in each length group are calculated based on a beta distribution *B(w,r)* over the first n_r length bins that is constrained to be unimodal.[20] Proportions of new recruits in each size group are the same from year to year. Beta distribution coefficients must be larger than one for the shape of the distribution to be unimodal. Therefore, w=1+e^{ω} and r=1+e^{ρ}, where ω and ρ are estimable parameters. It is probably better to calculate the parameters in this manner than as bounded parameters because there is likely to be less distortion of the Hessian for *w* and *r* values close to one and parameter estimation is likely to be more efficient.

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. In future versions of the CASA model, surplus production will be more accurately calculated by projecting populations at the beginning of the year forward one year assuming only natural mortality. ***[NOTE: surplus production calculations are being updated and were not available for the 2007 sea scallop stock assessment.]***

Population summary variables

Population summary variables described above are calculated for the entire stock (all length groups) and two user specified ranges of length bins. One set of bins is typically used for "stock" statistics that may, for example, exclude the smallest size groups. The other set of bins is typically used for exploitable sizes that may be vulnerable to the fishery. Several statistics are calculated for the beginning (January 1) and middle (July 1) of the year.

Estimated 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:

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

[20] Standard beta distributions used to describe recruit size distributions and in priors are often constrained to be unimodal in the CASA model. Beta distributions *B(w,r)* with mean $\mu = w/w + r$ and variance $\sigma^2 = wr/\left[(w+r)^2(w+r+1)\right]$ are unimodal when w > 1 and r >1. See http://en.wikipedia.org/wiki/Beta_distribution for more information.

CASA assumes that weight-at-length relationships for the stock (on January 1) and the fishery may differ and that mean fishery weight-at-length may change interannually. 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 for the population on January 1 computed at the midpoint of each length bin using the length-weight relationship for the fishery specified by the user. Total catch weight is:

$$W_y = (1+\omega_y)\sum_{L=1}^{n_L} C_{y,L} w'_L$$

where w'_L is weight at length in the fishery and ω_y is an annual anomaly input by the user to describe changes in fishery length weight that may occur from year to year due, for example, to changes in seasonal distribution of fishing. Model input data include a scalar that is used to convert the units for length-weight parameters (e.g. grams) to the units of the biomass estimates and landings data (e.g. mt).

F_y estimates for two years are comparable only if the fishery selectivity in the model was the same in both years. A set of simpler exploitation indices may be more useful when fishery selectivity changes over time. For example:

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

where x is a user-specified length bin (e.g. just below the first bin that is fully selected during all fishery selectivity periods) and the term $\tilde{N}$ is predicted abundance at the middle (July 1st) of the year. Similar statistics are calculated based on stock and catch weights and for January 1st was well as July 1st. Exploitation indices from different years with different selectivity patterns may be readily comparable if x is chosen carefully.

Spawner abundance in each year is (T_y) is computed:

$$T_y = \sum_{L=1}^{n_L} N_{y,L} e^{-\tau Z_y} g_L$$

where $0 \leq \tau \leq 1$ is the fraction of the year elapsed before spawning occurs (supplied by the user). Maturity at length (g_L) is from an ascending logistic curve:

$$g_L = \frac{1}{1+e^{a-bL}}$$

with parameters a and b supplied by the user.

Spawner biomass or egg production (S_y) in each year is computed:

$$S_y = \sum_{L=1}^{n_L} T_{y,L} x_L$$

where:

$$x_L = cL^v$$

Using parameters (c and v) for fecundity- or body weight at size supplied by the user.

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. All selectivity curves are rescaled to a maximum value of one before used in further calculations.

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. δ_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.

Per recruit recruit modeling

A complete per recruit output table is generated in all model runs that can be used for evaluating the shape of YPR and SBR curves, including the existence of particular reference points. The output table summarizes a wide range of per recruit information in terms of fully recruited fishing mortality F and a number of exploitation indices (U) specified by the user. Per recruit calculations in CASA use the same population model and code as all other model calculations under conditions identical to the last year in the model. It is a standard length-based approach except that discard and incidental mortality are accommodated in all calculations.

In per recruit calculations, fishing mortality rates and associated yield estimates are understood to include landings and discard mortality, but to exclude incidental mortality. Thus, landings per recruit is:

$$L = \frac{C}{(1+\Delta)}$$

where C is total catch (yield) per recruit and Δ is the ratio of discards D to landings in the last year of the model. Discards per recruit are calculated:

$$D = \Delta L$$

Losses due to incidental mortality (G) are calculated:

$$G = \frac{I\left(1-e^{-Z}\right)B}{Z}$$
$$= IK$$

where $I = F\,u$ is the incidental mortality rate, u is a user-specified multiplier (see above) and B is stock biomass per recruit. Note that C=FK so that K=C/F. Then,

$$G = \frac{FuC}{F}$$
$$G = uC$$

In addition to generating a per recruit output table, the model will estimate key ($F_{\%SBR}$, F_{max} and $F_{0.1}$) per recruit model reference points as parameters. For example,

$$F_{\%SBR} = e^{\theta_j}$$

where $F_{\%SBR}$ is the fishing mortality reference point that provides a user specified percentage of maximum SBR. θ_j is the model parameter for the j^{th} reference point.

Per recruit reference points are time consuming to estimate and it is usually better to estimate them after other more important population dynamics parameters are estimated. Phase of estimation can be controlled individually for %SBR, *Fmax* and $F_{0.1}$ so that per recruit calculations can be delayed as long as possible. If the phase is set to zero or a negative integer, then the reference point will not be estimated. As described below, estimation of F_{max} always entails an additional phase of estimation. For example, if the phase specified for F_{max} is 2, then the parameter will be estimated initially in phase 2 and finalized the last phase (phase >= 3). This is done so that the estimate from phase 2 can be used as an initial value in a slightly different goodness of fit calculation during the final phase.

Per recruit reference points should have no effect on other model estimates. Residuals (calculated – target) for %SBR, $F_{0.1}$ and F_{max} reference points should always be very close to zero. Problems may arise, however, if reference points (particularly F_{max}) fall on the upper bound for fishing mortality. In such cases, the model will warn the user and advise that the offending reference points should not be estimated. *It is good practice to run CASA with reference point calculations turned on and then off to see if biomass and fishing mortality estimates change*.

The user specifies the number of estimates required and the target %SBR level for each. For example, the target levels for four %SBR reference points might be 0.2, 0.3, 0.4 and 0.5

to estimate $F_{20\%}$, $F_{30\%}$, $F_{40\%}$ and $F_{50\%}$. The user has the option of estimating F_{max} and/or $F_{0.1}$ as model parameters also but it is not necessary to supply target values.

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-h*oc 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.

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:

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

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 $\sigma = \sqrt{\ln(1+CV^2)}$.

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 make L easier to interpret. 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 β>0 and γ>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.[21] Means, CV's and distributional parameters may, depending on the situation, be estimated in the model or specified by the user.

Robust methods

Goodness of fit for survey data may be calculated using a "robust" maximum likelihood method instead of the standard method that assumes lognormal measurement errors. The robust method may be useful when survey data are noisy or include outliers.

Robust likelihood calculations in CASA 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 is reduced, the tails of the t distribution become fatter so that outliers have higher probability and 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:

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

Catch weight data

Catch data (landings plus discards) 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 and errors in catch are assumed to be normally distributed. The standardized residual used in computing NLL for a single catch observation and in making residual plots is $r_Y = (C_Y - \hat{C}_Y)/\sigma_Y$.

Specification of landings, discards, catch

Landings, discard and catch data are in units of weight and are for a single or "composite" fishery in the current version of the CASA model. The estimated fishery selectivity is assumed to apply to the discards so that, in effect, the length composition of catch, landings and discards are the same.

[21] Parameters for standard beta distributions *B(w,r)* with mean $\mu = w/w + r$ and variance $\sigma^2 = wr/\left[(w+r)^2(w+r+1)\right]$ are calculated from user-specified means and variances by the method of moments. In particular, $w = \mu\left[\mu(1-\mu)/\sigma^2 - 1\right]$ and $r = (1-\mu)\left[\mu(1-\mu)/\sigma^2 - 1\right]$. Not all combinations of μ and σ^2 are feasible. In general, a beta distribution exists for combinations of μ and σ^2 if $0 < \mu < 1$ and $0 < \sigma^2 < \mu(1-\mu)$. Thus, for a user-specified mean μ between zero and one, the largest feasible variance is $\sigma^2 < \mu(1-\mu)$. These conditions are used in the model to check user-specified values for μ and σ^2. See http://en.wikipedia.org/wiki/Beta_distribution for more information.

Discards are from external estimates (d_t) supplied by the user. If $d_t \geq 0$, then the estimates are treated as the ratio of discard to landed catch so that:

$$D_t = L_t \Delta_t$$

where $\Delta_t = D_t/L_t$ is the ratio of discard and landings (a.k.a. *d/K* ratios) for each year. If $d_t < 0$ then the data are treated as discard in units of weight:

$$D_t = abs(d_t).$$

In either case, total catch is the sum of discards and landed catch ($C_t = L_t + D_t$). It is possible to use discards in weight $d_t < 0$ for some years and discard as proportions $d_t > 0$ for other years in the same model run.

If catches are estimated (see below) so that the estimated catch $\hat{C}_t$ does not necessarily equal observed landings plus discard, then estimated landings are computed:

$$\hat{L}_t = \frac{\hat{C}_t}{1 + \Delta_t}$$

Estimated discards are:

$$\hat{D}_t = \Delta_t \hat{L}_t.$$

Note that $\hat{C}_t = \hat{L}_t + \hat{D}_t$ as would be expected.

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 $(\hat{c}_Y)$ 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. Typical conventions make ${}^{c}N_Y \leq 200$ (Fournier and Archibald, 1982) or set ${}^{C}N_Y$ equal to the number of trips or tows sampled (Pennington et al., 2002). Effective sample sizes are sometimes chosen based on goodness of fits in preliminary model runs (Methot, 2000; Butler et al., 2003).

Survey index data

In CASA model calculations, "survey indices" are data from any source that reflect relative proportional changes in an underlying population state variable. In the current version, surveys may measure stock abundance at a particular point in time (e.g. when a survey was carried out), stock biomass at a particular point in time, or numbers of animals that dies of natural mortality during a user-specified period. For example, the first option is useful for bottom trawl surveys that record numbers of individuals, the second option is useful for bottom trawl surveys that record total weight, and the third option is useful for survey data that track trends in numbers of animals that died due to natural mortality (e.g. survey data for sea scallop "clappers"). Survey data that measure trends in numbers dead due to natural mortality can be useful in modeling time trends in natural mortality. In principle, the model will estimate model natural mortality and other parameters so that predicted numbers dead and the index data match in either relative or absolute terms.

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). Survey index and fishery length composition data are handled separately from trend data (see below). Survey data may or may not have corresponding length composition information.

In general, survey index data give one number that summarizes some aspect of the population over a wide range of length bins. Selectivity parameters measure the relative contribution of each length bin to the index. 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.

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 stock available to the survey. Scaling factors are calculated $q_s = e^{\varpi_s}$ where ϖ_s is estimable and survey-specific.

Available stock for surveys measuring trends in abundance or biomass is calculated:

$$A_{k,y} = \sum_{L=1}^{n_L} s_{k,L} N_{y,L} e^{-Z_{y,L}\tau_{k,y}}$$

where $s_{k,L}$ is size-specific selectivity of the survey, $\tau_{k,y}$=$J_{k,y}$/365, $J_{k,y}$ is the Julian date of the survey in year y, and $e^{-Z_y \tau_{k,y}}$ is a correction for mortality prior to the survey. Available biomass is calculated in the same way except that body weights w_L are included in the product on the right hand side.

Available stock for indices that track numbers dead by natural mortality is:

$$A_{k,y} = \sum_{L=1}^{n_L} s_{k,L} \tilde{M}_{y,L} \overline{N}_{y,L}$$

where $\overline{N}_{y,L}$ is average abundance during the user-specified period of availability and $\tilde{M}_{y,L}$ is the instantaneous rate of natural mortality for the period of availability. Average abundance during the period of availability is:

$$\overline{N}_{y,L} = \frac{\tilde{N}_{y,L}\left(1 - e^{-\tilde{Z}_{y,L}}\right)}{\tilde{Z}_{y,L}}$$

where $\tilde{N}_{y,L} = N_{y,L} e^{-Z\Delta}$ is abundance at elapsed time of year Δ=$\tau_{k,y}$- v_k, v_k=j_k /365, and j_k is the user-specified duration in days for the period of availability. The instantaneous rates for total $\tilde{Z}_{y,L} = Z_{y,L}\left(\tau_{k,y} - v_k\right)$ and natural $\tilde{M}_{y,L} = M_{y,L}\left(\tau_{k,y} - v_k\right)$ mortality are also adjusted to correspond to the period of availability. In using this approach, the user should be aware that the length based selectivity estimated by the model for the dead animal survey ($s_{k,L}$) is conditional on the assumed pattern of length-specific natural mortality ($\vec{u}$) which was specified as data in the input file.

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

Length bins for fishery and survey length composition data are flexible and the flexibility affects goodness of fit calculations in ways that may be important to consider. The user specifies the starting size (bottom of first bin) and number of bins used for each type of fishery and survey length composition. The input data for each length composition record identifies the first/last length bins to be used and whether they are plus groups that should include all smaller/larger length groups in the data and population model when calculating goodness of fit. Goodness of fit calculations are carried out over the range of lengths specified by the user. Thus length data in the input file may contain data for size bins that are ignored in goodness of fit calculations. As described above, the starting size and bin size for the population model are specified separately. In the ideal and simplest case, the minimum size and same length bins are used for the population and for all length data. However, as described below, length specifications in data and the population model may differ.

Care is required in some cases because the implicit definitions of plus groups in the model and data may differ. If the first bin used for length data is a plus group, then the first bin will contain the sum of length data from the corresponding and smaller bins of the original length composition record. However, the first bin in the population model is never a plus group. Thus, predicted values for a plus group will contain the sum of the corresponding and smaller bins in the population. The observed and predicted values will not be perfectly comparable if the starting sizes for the data and population model differ. Similarly, if the last bin in the length data is a plus group, it will contain original length composition data for the corresponding and all larger bins. Predicted values for a plus group in the population will be the sum for the corresponding bin and all larger size groups in the population, implicitly including sizes > L_∞.. The two definitions of the plus group will differ and goodness of fit calculation may be impaired if the original length composition data does not include all of the large individuals in samples.

In the current version of the CASA model, the size of length composition bins must be ≥ L_{bin} in the population model (this constraint will be removed in later versions). Ideally, the size of data length bins is the same or a multiple of the size of length bins in the population. However, this is not required and the model will prorate the predicted population composition for each bin into adjacent data bins when calculating goodness of fit. With a 30-34 mm population bin and 22-31and 32-41 mm population bins, for example, the predicted proportion in the population bin would be prorated so that 2/5 was assigned to the first data bin and 3/5 was assigned to the second data bin. This proration approach is problematic when it is used to prorate the plus group in the population model into two data bins because it assumes that abundance is uniform over lengths within the population group. The distribution of lengths in a real population might be far from uniform between the assumed upper and lower bounds of the plus group.

The first bin in each length composition data record must be ≥ L_{min} which is the smallest size group in the population model. If the last data bin is a plus group, then the *lower* bound of the last data bin must be ≤ the upper bound of the last population bin. Otherwise, if the last data bin is not a plus group, the *upper* bound of the last data bin must be ≤ the upper bound of the population bin.

NLL calculations for survey length composition data are similar to calculations for fishery length composition data. Surveys index data may measure trends in stock abundance or biomass but survey length composition data are always for numbers (not weight) of individuals in each length group. Survey length composition data represent a sample from the true stock which is modified by survey selectivity, sampling errors and, if applicable, errors in recording length data. For example, with errors in length measurements, individuals belonging to length bin j, are mistakenly assigned to adjacent length bins j-2, j-1, j+1 or j+2 with some specified probability. 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).

Prior to any other calculations, observed survey length composition data are converted to multinomial proportions:

$$i_{k,y,L} = \frac{n_{k,y,L}}{\sum_{j=L_{k,y}^{first}}^{L_{k,y}^{last}} n_{k,y,j}}$$

where $n_{k,y,j}$ is an original datum and $i_{k,y,L}$ is the corresponding proportion. As described above, the user specifies the first $L_{k,y}^{first}$ and last $L_{k,y}^{last}$ length groups to be used in calculating goodness of fit for each length composition and specifies whether the largest and smallest groups should be treated as "plus" groups that contain all smaller or larger individuals.

Using notation for goodness of fit survey index data (see above), predicted length compositions for surveys that track abundance or biomass are calculated:

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

Predicted length compositions for surveys that track numbers of individuals killed by natural mortality are calculated:

$$A_{k,y} = \frac{s_{k,L} \widetilde{M}_{y,L} \overline{N}_{y,L}}{\sum_{L=L_{k,y}^{first}}^{L_{k,y}^{Last}} s_{k,L} \widetilde{M}_{y,L} \overline{N}_{y,L}}$$

Considering the possibility of structured measurement errors, the expected length composition $\vec{A}'_{k,y}$ for survey catches is:

$$\vec{A}'_{k,y} = \vec{A}_{k,y} \text{Ɇ}_k$$

where Ɇ_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 Ɇ_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 Ɇ_k add to one to account for all possible outcomes in assigning individuals to observed length bins. Ɇ_k is the identity matrix if there are no structured measurement errors.

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 (e.g. 3 bins on either side of the true length bin).

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 the effective sample size ${}^{I}N_{Y}$ supplied by the user. Residuals are not used in computing NLL for length composition data but are available for use in checking model fit.

Residuals for goodness of fit to length data

Three types of residuals are calculated automatically for all of the length composition data used in the model and are written to a special output file that can be used to make residual plots and other diagnostics. The output file contains one record for set of length composition data and length bin used in goodness of fit calculations. Each record contains the name of the survey, survey id number, length, length bin id number, observed proportion, predicted proportion and three types of residuals (simple, Pearson and deviance).

For length composition type *t*, in year *y* and length *L*, the simple residual is

$$r_{t,y,L} = p_{t,y,L} - \hat{p}_{t,y,L}$$

where $p_{t,y,L}$ and $\hat{p}_{t,y,L}$ are observed and predicted proportions at length. The Pearson residual is

$$\rho_{t,y,L} = \frac{r_{t,y,L}}{\hat{p}(1-\hat{p}) / {}^{I}N_{t,y,L}}$$

where the denominator is the expected standard deviation given the predicted proportion and the effective sample size used in goodness of fit calculations. The deviance residual is basically the contribution of the length composition observation to the total likelihood:

$$\delta_{t,y,L} = sign(r_{t,y,L})^{I}N_{t,y,L}\left[p_{r,y,L}\ln\left(\hat{p}_{r,y,L}\right) - p_{r,y,L}\ln\left(p_{r,y,L}\right)\right]$$

Note that the deviance residual is zero if the observed and predicted proportions match exactly and that the deviance and simple residuals have the same sign.

Effective sample size for length composition data

Effective sample sizes that are specified by the user are used in goodness of fit calculations for survey and fishery length composition data. A post-hoc estimate of effective sample size can be calculated based on goodness of fit in a model run (Methot 1989). Consider the variance of residuals for a single set of length composition data with N bins used in calculations. The variance of the sum based on the multinomial distribution is:

$$\sigma^2 = \sum_{j=1}^{N}\left[\frac{\hat{p}_j\left(1-\hat{p}_j\right)}{\varphi}\right]$$

where φ is the effective sample size for the multinomial and $\overline{p}_j$ is the predicted proportion in the j^{th} bin from the model run. Solve for φ to get:

$$\varphi = \frac{\sum_{j=1}^{N}\left[\hat{p}_j\left(1-\hat{p}_j\right)\right]}{\sigma^2}$$

The variance of the sum of residuals can also be calculated:

$$\sigma^2 = \sum_{j=1}^{N}\left(p_j - \hat{p}_j\right)^2$$

This formula is approximate because it ignores the traditional correction for bias. Substitute the third expression into the second to get:

$$\varphi = \frac{\sum_{j=1}^{N}\left[\hat{p}_j\left(1-\hat{p}_j\right)\right]}{\sum_{k=1}^{N}\left(p_j-\hat{p}_j\right)^2}$$

which can be calculated based on model outputs. The assumed and effective sample sizes will be similar in a reasonable model when the assumed sample sizes are approximately correct. Effective sample size calculations can be used iteratively to manually adjust input vales to reasonable levels (Methot 1989).

Variance constraints on dev parameters

Variability in dev parameters (e.g. for natural mortality, recruitment or fishing mortality) can be limited using variance constraints that assume the deviations are either independent or that they are autocorrelated and follow a random walk. When a variance constraint for independent deviations is activated, the model calculates the NLL for each log scale residual γ_y / σ_γ, where γ_y is a dev parameter and σ is a log-scale standard deviation. If the user supplies a positive value for the arithmetic scale CV, then the NLL is calculated assuming the variance is known. Otherwise, the user-supplied CV is ignored and the NLL is calculated with the standard deviation estimated internally. Calculations for autocorrelated deviations are the same except that the residuals are $\left(\gamma_y - \gamma_{y-1}\right) / \sigma_\gamma$ and the number of residuals is one less than the number of dev parameters.

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.[22] 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:

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

where the weights at length w_L^f are for the fishery rather than the population. Predicted values for LPUE data are calculated:

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

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

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

Per recruit (SBR and YPR) reference points

The user specifies a target %SBR value for each reference point that is estimated. Goodness of fit is calculated as the sum of squared differences between the target %SBR and %SBR calculated based on the reference point parameter. Except in pathological situations, it is always possible to estimate %SBR reference point parameters so that the target and calculated %SBR levels match exactly. Reference point parameters should have no effect on other model estimates and the residual (calculated – target %SBR) should always be very close to zero.

Goodness of fit for $F_{0.1}$ estimates is calculated in a manner similar to %SBR reference points. Goodness of fit is calculated as the squared difference between the slope of the yield curve at the estimate and one-tenth of the slope at the origin. Slopes are computed numerically using central differences if possible or one-sided (right hand) differences if necessary.

F_{max} is estimated differently in preliminary and final phases. In preliminary phases, goodness of fit for F_{max} is calculated as $(1/Y)^2$, where Y is yield per recruit at the current estimate of F_{max}. In other words, yield per recruit is maximized by finding the parameter estimate that minimizes its inverse. This preliminary approach is very robust and will find F_{max} if it exists. However, it involves a non-zero residual ($1/Y$) that interferes with calculation of variances and might affect other model estimates. In final phases, goodness of fit for F_{max} is calculated as (d^2) where d is the slope of the yield per recruit curve at F_{max}. The two approaches give the same estimates of F_{max} but the goodness of fit approach used in the final phases has a residual of zero (so that other model estimates are not affected) and gives more reasonable variance estimates. The latter goodness of fit calculation is not used during initial phases because the estimates of F_{max} tend to "drift down" the right hand side of the yield curve in the direction of decreasing slope. Thus, the goodness of fit calculation used in final phases works well only when the initial estimate of F_{max} is very close to the best estimate.

Per recruit reference points should have little or no effect on other model estimates. Problems may arise, however, if reference points (particularly F_{max}) fall on the upper bound for fishing mortality. In such cases, the model will warn the user and advise that the offending reference points should not be estimated. *It is good practice to run CASA with and without reference point calculations to ensure that reference points do not affect other model estimates including abundance, recruitments and fishing mortality rates.*

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 $\mathcal{P}$. 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 $\mathcal{P}$. 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 coincides so that:

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

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

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

$$K_t = \frac{A}{a_k u_k}$$

$$e_k = K_t q_t$$

where $I_{k,y}$ is a survey observation in units equivalent to biomass (or numerical) density (e.g. kg per standard tow), $B_{k,y}$ is the biomass (or abundance) available to the survey, A is the area of the stock, a_k is the area swept during one tow, $0<e_k\leq 1$ is efficiency of the survey gear, and u_k is a constant that adjusts for different units.

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, a_k and u_k. At each iteration if 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 unimodal beta distribution with the user-specified mean and CV.

If efficiency estimates are used as prior information (if the likelihood weight $\lambda > 0$), then it is very important to make sure that units and values for the survey data (I), biomass or abundance (B), stock area (A), area per tow (a), and adjustments for units (u) are correct (see Example 1). The units for biomass are generally the same as the units for catch data. In some cases, incorrect specifications will lead to implied efficiency estimates that are ≤ 0 or ≥ 1 which have zero probability based on a standard beta distribution used in the prior. The program will terminate if $e \leq 0$. If $e \geq 1$ during an iteration, then e is set to a value slightly less than one and a penalty is added to the objective function. In some cases, incorrect specifications will generate a cryptic error that may have a substantial impact on estimates.

Implied efficiency estimates are useful as a model diagnostic even if very little prior information is available because some model fits may imply unrealistic levels of implied efficiency. The trick is to down weight the prior information (e.g. λ=1e^{-6}) so that the implied efficiency estimate has very little effect on model results as long as $0 < e < 1$. Depending on the situation, model runs with e near a bound indicate that estimates may be implausible. In addition, it may be useful to use a beta distribution for the prior that is nearly a uniform distribution by specifying a prior mean of 0.5 and variance slightly less than 1/12=0.083333.

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. In effect, efficiency estimates from field studies tend to be biased high as estimates of efficiency relative to the entire stock.

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:

$$d_t = \left| \begin{array}{l} Ft - \Phi \ \mathit{if} \ Ft > \Phi \\ 0 \ \mathit{otherwise} \end{array} \right.$$

and

$$q_t = \left| \begin{array}{l} \ln(Ft / \Phi) \ \mathit{if} \ Ft > \Phi \\ 0 \ \mathit{otherwise} \end{array} \right.$$

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