

## 5 LINK MODELS

Where variables, such as stock size, can be observed directly, the concept of link models is largely redundant. However they are useful where only indirect observations can be made on these variables, and the relationships are complex, as is almost always the case. Separating link models from the population model is useful in several respects. It allows a number of indices to be incorporated into the fitting process, each with their own separate model describing how they are related to variables in the population model. Ideally link models avoid any time series effects, as unlike population models they should not represent a process. This means the link models are subject only to direct measurement or observation errors. Finally, link models are often linear in form. This has a distinct advantage in that linear parameters, which can be estimated directly, are separated from non-linear parameters, associated with the population model, which need to be estimated iteratively. This can significantly decrease the time needed to fit the model.

Link models will always add information to the population models if the number of parameters is less than the number of data points. However, clearly the fewer the parameters in the link model, the more information there will be for fitting the population model. So a biomass survey, which requires no parameters, is preferable to a biomass index which requires at least one parameter. In some cases a large number of parameters are required to account for changes having nothing to do with the stock. This will undermine the value of the data series.

### 5.1 ABUNDANCE INDICES

#### 5.1.1 Index Standardisation

For many indices there are a number of potential influences besides the variable we are interested in. For example, CPUE may be affected by the fishing area, season, gear, size of vessel, and number of crew as well as the stock size. Standardisation aims to separate these effects, in particular removing those that may bias the population size index, and to generate indices, which may be some of the fitted parameters. In order to be able to evaluate the quality of the index that is derived, a key statistic is the proportion of the variability that can be explained by the explanatory variables. This proportion is often quite low in these types of analysis, in the region of 50-60 %.

The analysis most often used is a special case of the generalized linear model approach (McCullagh and Nelder 1983). In the version applied in many fish stock assessments, the task is to find the variation in the data that can be allocated to vessel, time and area. This is done by applying a model of the form:

$$\ln( CPUE ) = Constant + Vessel + Area + Year + Season + Area.Season \quad (35)$$

where the "." operator refers to the interaction terms between factors (see McCullagh and Nelder 1983). In specific analyses, there may be fewer terms involved (e.g. if only one vessel is used in the survey and if the survey is confined to a particular season throughout the time series the vessel and season terms do not appear).

The analysis provides an estimate of the value of the explanatory parameters. These parameter values are often used as input for the subsequent stock assessment. For example, the year effect in an analysis of CPUE data from an abundance survey could be used as an estimate of relative annual abundance and the vessel effect is the relative fishing power.

The season effect is often interpreted as either difference in availability – fish concentration varies with the season – or as seasonal migrations. The migrations should make the interaction term (area.season) significant (i.e. the geographical distribution of fish varies through the year).

In many fish stock assessments, it is preferable to isolate individual analyses and investigate the data in subsets. For example, studies of the structure of the catch data, the abundance CPUE data from surveys and CPUE data from logbooks can be undertaken separately. Only when one is satisfied with the consistency of the data does it make sense to include the data in an integrated model. For this purpose, linear models are often used (Gavaris 1988, Kimura 1981, Large 1992).

### **5.1.2 Dis-aggregated Abundance Indices**

This is the class of indices most often used in fish stock assessment based on the VPA analytical model. These indices are typical CPUE estimates from either well-defined commercial fleets or from abundance surveys using research vessels. The CPUE values are expressed in numbers-by-age per effort unit. The effort unit can be days-at-sea, trawl hours, search time, etc. Commercial CPUE data are obtained through sampling the commercial fisheries for biological information and linking this information with catch and effort statistics. Abundance surveys using research vessels provide these data directly, often from bottom trawl surveys expressed as numbers caught per hour trawling.

Survey data differ from commercial CPUE data in two respects:

- Survey data are obtained through a designed sampling programme and the data often represent the stock over a short time period.
- Commercial CPUE data represent the geographical distribution of fishing activities as well as fish abundance, but the data often represent the stock over a longer time period.

Whether the CPUE data are linearly (or otherwise) related to abundance is discussed separately for the two data sources as the problems are distinct. For commercial fishing, the sampling is probably not random relative to the population and there are a variety of fishing vessels with different fishing strategies and different fishing power involved in construction of the mean CPUE value. Surveys represent few samples, the largest surveys include 500-600 trawl stations per year, but the coverage, effort allocation and standardisation of gear and fishing strategy are under the control of the researcher. In the models to be studied below we assume a linear relation between CPUE and stock abundance. Experience with non-linear models has shown marked problems with over-estimation of stock size (and corresponding under-estimation of the fishing mortality), because of the random noise in the data.

### 5.1.2.1 Commercial CPUE data

Even when it is possible to dis-aggregate commercial CPUE data, they are often only representative for a time period, e.g. a quarter or a year. The link model becomes:

$$\ln CPUE_{ay} = \ln q_a + \ln \tilde{P}_{ay} \quad (36)$$

The population model might need to be corrected to match the population relevant for the CPUE index. The model must not only account for any mortality occurring in the population before the index is measured, but also for any mortality occurring in the stock over the time period for which the index measurements are taken:

$$\ln \tilde{P}_{ay} = \ln P_{ay} - \alpha Z_{ay} + \ln \frac{1 - \exp(-(\beta - \alpha) Z_{ay})}{(\beta - \alpha) Z_{ay}}$$

with (37)

$$Z_{ay} = F_{ay} + M_a$$

The constant  $\alpha$  is the fraction of the total mortality that occurs before the index is relevant and  $\beta - \alpha$  is the fraction of the mortality occurring while the index is relevant. In practice, these fractions are not known precisely and are approximated by the fraction of the year that has past prior to the measurement starting, while  $\beta - \alpha$  refers to the period the observed CPUE applies.

The random noise is usually assumed to be log-normally distributed, but following Methot (1990) this may not be appropriate as the age composition is a mixture of a contribution of total catch – possibly log-normal – and the breakdown of this catch into age groups – possibly multinomial. The estimation is very similar to estimating the total catch in numbers. There is, however, an additional problem as the effort data available are usually the nominal effort. The efficiency of a nominal effort unit may well increase with time and the fishing strategy of a fleet may change with time (e.g. as a result of changes in the geographical distribution or abundance of the stock). These problems suggest that CPUE indices from commercial fisheries may only be applicable for shorter time periods.

Another approach could be to estimate the catch based on effort data and a linear link between effort and fishing mortality. For a set of terminal  $F$ 's, the cohort sizes are calculated back using the observed catches and the standard VPA methodology. Catches can then be estimated based on the cohort sizes at the beginning of each year and the effort. Hence, the expected catch becomes:

$$\ln C_{ay}^{\text{mod}} = \ln P_{ay} + \ln \left( \frac{1 - e^{-q_a E_y - M_a}}{q_a E_y + M_a} \right) + \ln(q_a E_y) \quad (38)$$

This method has the distinct disadvantage that the link model is non-linear, but the errors may be better behaved than using a CPUE index, particularly if  $F$  fluctuates widely during the time series to values greater than  $1.0 \text{ year}^{-1}$ . Alternatively, the  $F$ 's calculated from the VPA may be fitted to the effort directly, which should be easier (see Section 5.2).

### 5.1.2.2 Survey data

These data are often the best stock size indicators available since such data should include sampling design to control and measure errors resulting from the stock distribution, gear design etc. The survey will not necessary take place at the start of the year and the CPUE therefore should be corrected for the mortality that takes place between the start of the year and the start of the survey. The survey often lasts a short period (e.g. a month). Even so it may be relevant to correct the survey CPUE for the mortality that takes place during the time of the survey and the model is therefore the same as that presented for the commercial CPUE data (Equation 37). Note, for surveys of short duration, where effectively  $\beta - \alpha = 0$ , the last term in the model becomes zero.

### 5.1.3 Biomass Indices

Biomass indices are usually provided from two different sources: from commercial fishing where CPUE data (catch weight per trip, or per day-at-sea, or per trawl hour, etc.) may be available from logbook or landing reports. Such data may or may not be accompanied by biological sampling. It is therefore not possible in all cases to break these data down by number and by age group. Another data source that provides biomass indices are egg- and larvae surveys that provide estimates of the spawning stock biomass.

The model linking the CPUE biomass index to the stock is:

$$\ln I_y^{\text{mod}} = \ln \left( q \sum_a W_a \tilde{P}_{a,y} \right) \quad (39)$$

where the population,  $\tilde{P}_{a,y}$ , is the appropriate adjusted population corresponding to the CPUE (Equation 37). As, in this case, the age dependency is not estimable, the model is formulated with a single (average) catchability parameter.

Spawning stock biomass estimates from egg- and larvae surveys can be obtained either in absolute terms or as indices. For establishing the link between these observations and the spawning stock biomass calculated from the analytical age disaggregated model, it is necessary to include a new data item, the maturity ogive ( $mat_{ay}$ ). This is an array of proportions, which gives the fraction of each age group in numbers that is mature at spawning time. This ogive probably varies between years and maturity data should therefore ideally be available by year. However, such data are often not collected routinely and then an average maturity ogive is used for a series of years. The expected spawning stock size index is calculated as:

$$\ln I_y^{\text{mod}} = \ln \left( q \sum_a W_a mat_{ay} P_{ay} e^{-\alpha F_{ay} - \beta M_a} \right) \quad (40)$$

where  $\alpha$  is the proportion of the fishing mortality and  $\beta$  the proportion of the natural mortality that is exerted on the stock before spawning (i.e. proportion of the year between 1<sup>st</sup> January and spawning time). In this case the mean weights-at-age,  $W_a$ , should be those of the spawning stock, not of the total stock nor of the catch.

These indices are often assumed to be log-normally or normally distributed. The indices may not be estimated directly from the surveys, but result from separate analysis of the survey data (Pennington 1983, 1986).

## 5.2 FISHING MORTALITY INDICES

Effort data are often provided through fisheries statistics. These data can be collected from logbooks, from landing reports or as interview surveys of skippers. The model most often included in assessments is the assumption of a linear relation between fishing mortality and nominal effort:

$$\bar{F}_y = q E_y^{obs}$$

or in logarithmic form :

(41)

$$\ln \bar{F}_y = \ln q + \ln E_y^{obs}$$

where the  $\bar{F}_y$  is the average fishing mortality of the fully recruited age groups. There may well be data from several fleets each representing a different segment of the age composition of the stock. These fleet data may all be valid stock indicators that preferably should be included in the assessment.

Nominal effort may not be linearly related to the fishing mortality. This is because fishing is not a random sampling of the stock, but all possible skills are used to find those grounds where the catch rates are highest. Another problem with the use of such data is the increase in efficiency that takes place over the years. (Squires 1994, Pascoe and Robinson 1996). Such efficiency increase would be reflected by a time dependence in the catchability  $q$ .

In Section 4.4, a separable VPA model was introduced as part of the population model. However, it can also be developed as a link model. In this case, Equation 41 is expanded to allow for different catchabilities for each age group:

$$F_{ay} = q_a E_y^{obs}$$
(42)

We can then fit these estimated  $F_{ay}$  to those in the VPA population model (Equation 7) given the terminal  $F$ 's (perhaps also derived from Equation 42). This approach assumes an error between the expected and observed fishing mortalities rather than catches, which may be considered more appropriate if catches are considered more accurate than nominal effort. Because both Equations 41 and 42 are linear, they usually add only a small cost to the fitting procedure (see Section 6.3 Finding the Least-squares Solution).

It should be noted that the effort data may already have been used to construct CPUE stock indices and in this case the effort data should not be used again as part of the estimation procedure.

## 5.3 USING LENGTH COMPOSITIONS

Although VPA methods use age, catch data is at best divided into size classes. The link between numbers-at-size and numbers-at-age is potentially a complex one. Therefore, this link is usually dealt with separately as a conversion from size to age frequency using a variety of different methods. Once the conversion is complete, the VPA proceeds as though all fish were aged.

There are several groups of species where it is not possible to age individuals, such as shrimps, nephrops, lobsters, crabs and many tropical fish species. Crustaceans do not possess bone structures that they keep throughout their life span and therefore their shells or exoskeletons cannot be used for ageing. The environment of tropical fish may not have sufficient seasonal differences to establish well-defined structures in otoliths and bones<sup>1</sup>.

In these cases, the approach is to use solely the length compositions in the population as the basis for establishing cohorts. Reproduction, even in tropical areas usually shows some seasonal pattern (e.g. based on the local rainy season) and this is reflected in the length compositions where a peak in the length composition will identify a cohort.

In length-based assessment we define a method that converts length composition into age composition without age data. This procedure is often called cohort slicing (the generic term). The basis of the procedure is to allocate an age to a proportion of the fish found in a length range. This is precisely what the ALK does with available size-age sample. The difference is that whereas ALK can identify different age groups among similarly sized fish, cohort slicing cannot, which may lead to inaccurate allocation of catches to cohorts.

Although we only use length in this discussion, other additional biological information might be used. In all cases, a good knowledge of the biology of the species being analysed can greatly assist in developing models. For example, it is possible to use the sex or location of capture of some species to assist in establishing the age structure.

The present discussion is an expansion of Lassen (1988). Several methods are implemented in the FAO/ICLARM software (Gayanilo Jr. *et al.* 1996). Common length based approaches are explained in Sparre and Venema (1998).

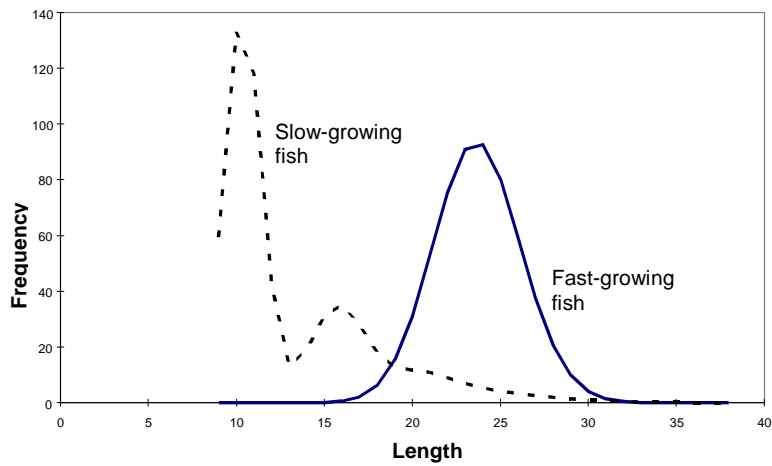
### 5.3.1 Factors Affecting Length Frequencies

Observed length frequencies depend on relative year class strength, total mortality, average growth and variation in growth. This is illustrated in the following examples where individual parameters are varied to produce simulated length frequencies. In each case, the effect of the parameter is illustrated with respect to interpreting length

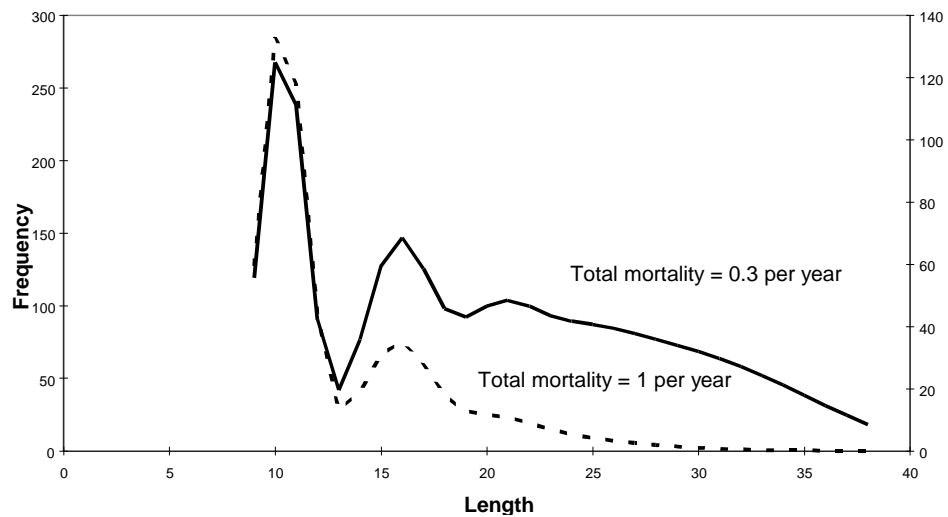
---

<sup>1</sup> This observation may very much depend on the species and local conditions. It may always be worthwhile exploring whether direct ageing techniques can be used for tropical fish for each fishery, as their use greatly enhances the scientific advice that can be given to help manage the fishery.

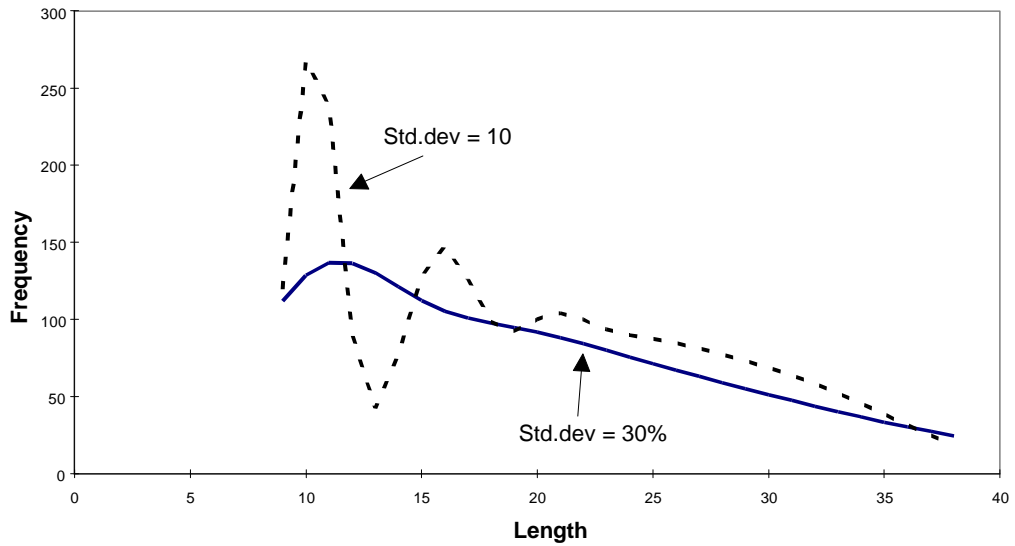
frequency distribution, and in particular identifying modes in these distributions representing individual cohorts.



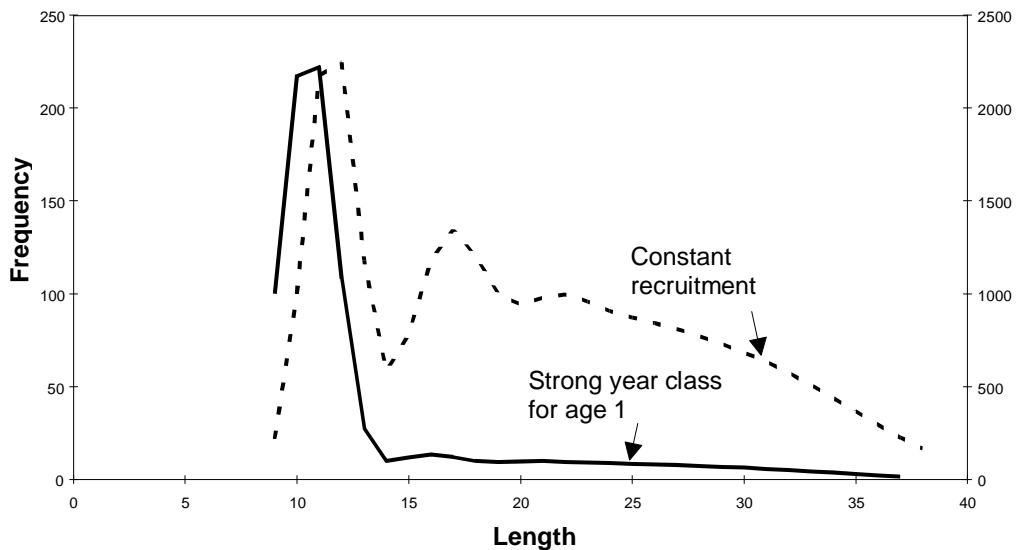
**Figure 5.1** Theoretical length distributions for two sets of growth parameters, with the same mortality ( $1.0 \text{ year}^{-1}$ ), for the Eastern Baltic cod fishery year classes 1966-1994. Modes representing cohorts can only be detected in the length frequencies for slow-growing fish.



**Figure 5.2** Theoretical length distributions for slow-growing fish, with different mortalities. Lower mortality gives a greater chance of detecting modes for older cohorts if significant differences remain between sizes at these ages.



**Figure 5.3** Theoretical length distributions for the same set of growth parameters and mortality, but with two levels of variation in growth. As variation in growth increases it becomes more difficult to detect modes as cohorts merge in size frequency.



**Figure 5.4** Theoretical length distributions for the same set of growth parameters, mortality and growth variation, but with and without variable recruitment. Recruitment variability tends to mask modes of cohorts from weak recruitments. However tracing strong year classes through time may give indications of growth rates.



The conclusion from the examples is that simple identification of peaks in the length frequencies can be impossible. It will work when there is low growth and recruitment variability, but otherwise no modes would be apparent. Analysis of a set of length frequencies under the assumption of common growth can help, as this will guide where the peaks are considered to be on the length axis. Such analysis, in many cases, is critically dependent on the growth assumption being correct.

### **5.3.2 Length to Age Conversion**

There are three approaches to decomposing size groups into ages. Each method links observations such as age and size samples to catches in numbers per age class, which are the variables used in the population models. Notice that it is hardly ever the case that catch-at-age is observed directly. The preferred method is to use age data through age-length keys (ALKs). In tropical fisheries, there is often a heavy reliance on size frequency data only, which will incur a significant penalty in accuracy. Wherever possible, ageing should be considered as part of the data collection programme.

#### **5.3.2.1 Methods Using Age-Length Keys (ALK)**

Use of size frequencies can greatly enhance the use of ageing data, as ageing is generally expensive. Using size frequencies allows improved sampling techniques reducing the amount of ageing that needs to be done by making use of the information contained in fish size to help generate the age distribution. Unlike other methods it does not depend on size however, so even larger size groups can be broken down into age categories. All the usual sampling techniques apply. So, age samples should be stratified by size, but be random within each size group.

A link model can be used to define catches in each age group as a sum of catches from the size groups:

$$C_{ay} = \sum_l p_{la} C_{ly} \quad (43)$$

where  $p_{la}$  = the proportion (i.e. a probability) of fish in length group  $l$  of being age  $a$ . The  $p_{la}$  parameters can be estimated, for example, from age-at-length data using a multinomial model (McCullagh and Nelder 1983).

#### **5.3.2.2 Methods Not Requiring a Growth Model**

These methods rely on identifying modes in length frequency samples representing cohorts. For example, the Bhattacharya method uses the first mode in a sample to fit a normal curve representing the youngest cohort. This curve is then used to remove all fish belonging to this cohort from the sample. A similar procedure is then applied to the next mode, and so on. Methods include Bhattacharya (1967), Tanaka (1956), MIX: MacDonald and Pitcher (1979) and NormSep: Hasselblad and Tomlinson (1971).

### 5.3.2.3 Methods Requiring a Growth Model

The simplest case is Jones length-based cohort analysis (Jones and Van Zalinge 1981). In this growth is assumed deterministic and the sample is sliced up according to back-transformation of the von Bertalanffy growth equation.

The method is based on re-writing the survival equation into length differences:

$$\Delta t = \frac{-\ln\left(\frac{L_\infty - l_1}{L_\infty - l_2}\right)}{k} \text{ for each size class} \quad (44)$$

$$N_{l+\Delta l} = N_l e^{-Z \Delta t} \quad (45)$$

where  $l_1$  and  $l_2$  are respectively the lower and upper bounds of the size class. Note that size classes should be chosen such that  $M\Delta t$  should be less than 0.3.

Given the change in age over each size class ( $\Delta t$ ), the population sizes within each class can be constructed in much the same way as a VPA. The method requires that the growth parameters are known. Methods such as ELEFAN may be used to estimate these. Provided appropriate averaging of the length compositions has been done so that the observed length compositions can be assumed to present the equilibrium length compositions, then a simple VPA back-calculation over length rather than age groups is possible.

The method has been investigated (Addison 1989, ICES 1995a,b) with the following conclusions:

- Cohort analysis works on a single length frequency sample assuming the population has been in a steady state. A number of length frequency samples from different times may be required to verify this.
- The model is insensitive to errors in the terminal exploitation rate, if  $F \gg M$ .
- The model is extremely sensitive to  $M$ .
- The narrowest length interval that makes data reasonably smooth should be used.
- Considerable care should be taken with the method when only poor growth data are available or when individual variation in growth is high. Ensure the terminal length interval (“plus” group) has an initial length (lower bound) of less than 70 % of  $L_\infty$ . This will minimise errors in the model’s output due to errors in estimates and variances of  $L_\infty$  and  $k$ . Any estimate of overall  $F$  should therefore cover only the smaller size interval representing the majority of the catch.
- Estimates of abundance should not be taken as absolute values. Use them only as indices to reflect relative changes.

More realistically, other methods use the growth model (usually the von Bertalanffy model) to relate the age of a fish to a parameter of a probability distribution of its size. The parameter is usually taken to be a mean, and the probability distribution is the Normal. So given growth parameters, and some parameter summarising the variation in growth-at-age, we can define the probability ( $\varphi$ ) a fish of length  $l$  is age  $a$  as:

$$l_{age} = L_{\infty} \left( 1 - e^{-k(a-a_0)} \right)$$

and (46)

$$\varphi(l; age) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(l-l_{age})^2}{2\sigma^2}}$$

The cumulative normal distribution can then be used to calculate the proportions of each length group, which should be allocated to each age group ( $p_{ia}$  in Equation 43). Other distributions, such as the log-normal may be used and in many cases may be more appropriate (see Beyer and Lassen 1994).

There are many methods to fit this and other similar functions to modes through one or more length frequency samples. They ignore modes if they do not conform to the growth models. These methods are complicated, but software is widely available, such as ELEFAN (Pauly 1987), SCLR (Shepherd 1987), MULTIFAN (Fournier *et al.* 1990).

The conversion from length to age based only on length frequency is usually subject to a variance much higher than that obtained in age readings. Most worryingly, this uncertainty is not quantified allowing a researcher to overestimate the accuracy of their assessment. Decomposition of the length distribution into age groups is less precise than ageing directly, and therefore should be used only when no ageing methods are available.