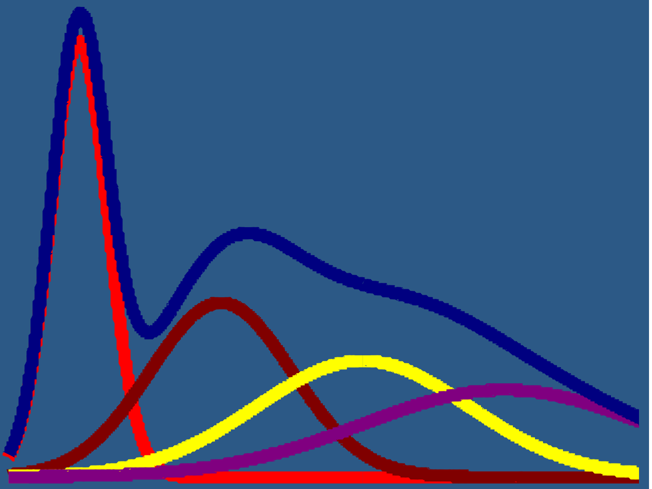
CMIX

User's Manual & Specifications



Australian Antarctic Division

Version Information

Citation

de la Mare, W. (1994) Estimating Krill recruitment and its variability. CCAMLR Science. 1:55-69.
Copies available from authors, or from CCAMLR Secretariat.

Current Version

CMIX.EXE 1997
© Marine and Ecological Research
Australian Antarctic Division

CMIX User Interface

Ronai, B. and Lamb, T.
© Australian Antarctic Division

User's Manual

Last Modified 28/9/02
Written by de la Mare, W., Constable, A., van Wijk, E., Lamb, T., Hayes, D. and Ronai, B.
© Australian Antarctic Division

Contact Person

Andrew J. Constable
Antarctic Marine Living Resources Group
Australian Antarctic Division
Channel Highway
Kingston, Tasmania 7050
Australia
Ph: 61 3 6232 3558
Email andrew.constable@aad.gov.au

CMIX

User's Manual & Specifications

Australian Antarctic Division

Contents

Section 1 Introduction

1.1 Introduction

This manual outlines the use of CMIX which is a FORTRAN program designed to fit mixture distributions to length-density data obtained from net surveys using maximum likelihood estimation. The current version of CMIX is distributed with a CMIX Excel Add-In to streamline data input and visualise output.

The user's manual introduces CMIX and the analytical method that it employs (see Section 2). The manual describes the use of the CMIX Excel Add-In and the format of both the input and output files. CMIX can be used with/without the CMIX Excel Add-In.

1.2 CMIX

CMIX fits a mixture distribution to length-density distributions (see Section 1) derived from net-survey data. The function of the program is similar to the well known method of MacDonald and Pitcher (1979). However, the mixture distribution is fitted using a maximum likelihood estimator that assumes that the length-density data have an Aitcheson delta distribution (Aitcheson, 1955) (see equation 4.4). This distribution is more suitable for describing densities estimated from net haul surveys because it provides for the possibility that a given survey haul will be empty; The delta distribution includes a log-normal distribution for the non-zero density observations and a finite probability for a zero density estimate. The current version of CMIX allows only for a mixture of normal distributions each with characteristic means and standard deviations of length.

The mixture distribution is parameterised so as to allow;

1. Estimation of the density of fish in each mixture component (cohort)
2. Estimation of the proportion of recruits in the sample, where recruits are taken to be represented by the first mixture component[1].

The program reads a data file which includes a specification of the distribution mixture to be fitted, in terms of the number of mixture components, boundary values on the means of those components, and possible restrictions on the standard deviations for the mixture components. The standard deviations can be specified to be independent, or linearly related. Example By restricting the linear relationship, it is possible to include constant standard deviations and constant coefficients of variation for the mixture components. The data must be given as the haul by haul densities in each length interval. Any hauls which included zero densities for all length intervals still need to be included in the data file, because the zeros still contribute information about the mean density for each class.

Estimates are obtained by non-linear minimisation of a residual function, which is -(log-likelihood). The minimisation involves a search over the parameter values for the set of values which gives the minimum value of the residual function. The minimisation routine

[1] The procedure to implement R1 calculations will be detailed in the second version.

requires for each parameter of the mixture distribution a starting value, a step length to be used in searching, and a set of bounds in which the search

will be confined. In the case where the mixture distribution standard deviations are linearly related, all these values are specified directly by the user. In the case of the mixture component means, only the search bounds are specified; suitable starting points and step lengths are calculated by the program from the bounds. The program also allows for various parameters to be held at fixed values. For those parameters where the user does not have direct control of the step length, such as the means and standard deviations, the fixed values can be achieved by directly specifying those values in the data file. For those parameters where the user does have direct control of the step length such as the parameters of a linear relationship between the mean and standard deviation, fixed values can be achieved by setting the step length to zero.

Be warned, this program calculates estimates by brute force, computations can take several hours for large data files.

1.3 Excel Add-In

A visual basic Add-In for Excel (CMIX_Excel_Add-In.xla) is available which can be used to create CMIX Input files, run CMIX and display CMIX output. CMIX requires a detailed and precise input file format, which users can find difficult to build. CMIX Wizard, a part of the excel Add-In, can be used to create the CMIX input files, run CMIX and display the CMIX output.

The Excel Add-In tool bar is shown in Figure 1 along with a schematic diagram of its options. The toolbar contains three options, ‘CMIX Wizard’, ‘Run CMIX’ and ‘Display CMIX Output’. For a more detailed description see Section 4.

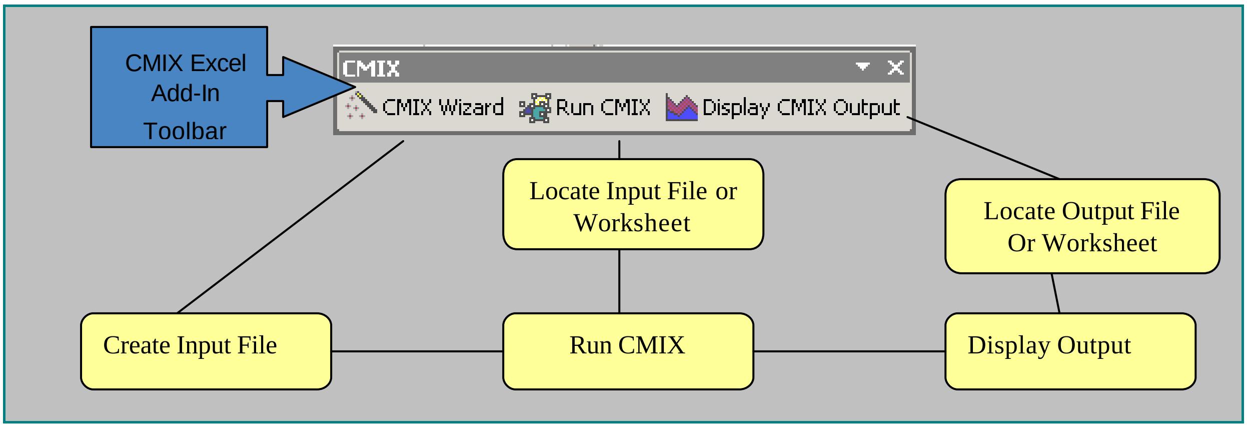

CMIX Excel Add-In Toolbar
CMIX
CMIX Wizard
Run CMIX
Display CMIX Output
Locate Input File or Worksheet
Locate Output File Or Worksheet
Create Input File
Run CMIX
Display Output

Figure 1 CMIX Excel Add-In Toolbar Schematic

Section 2 Understanding The CMIX Model

2.1 Introduction

This Section aims to give a brief overview of the theory behind the model used in the CMIX routines.

2.2 The Model (See de la Mare 1994a)

The aim of the method is to estimate the proportion of recruits in samples from populations. The proportion of recruits, also known as the gross recruitment rate, *R(t)*, is the ratio of numbers in age class *t*, to the numbers in that age class and above, that is:

$$R(t) = \frac{A_0}{\sum_{i=t}^{n} A_i} \tag{1.1}$$

where A_i is the number of animals in age class *i*, and *n* is the age of the oldest animals in the population present in non-negligible numbers. Thus, we need only be able to separate one young age class from all the others; it is not necessary to be able to distinguish between the older age classes

Figure 2 shows a schematic mixture generated from four nominal age classes with length-at-age distributions C1 to C4. It is clear that there is little prospect of accurately decomposing the mixture for age classes 3 and above. However, this is not necessary for calculating *R(t)*.

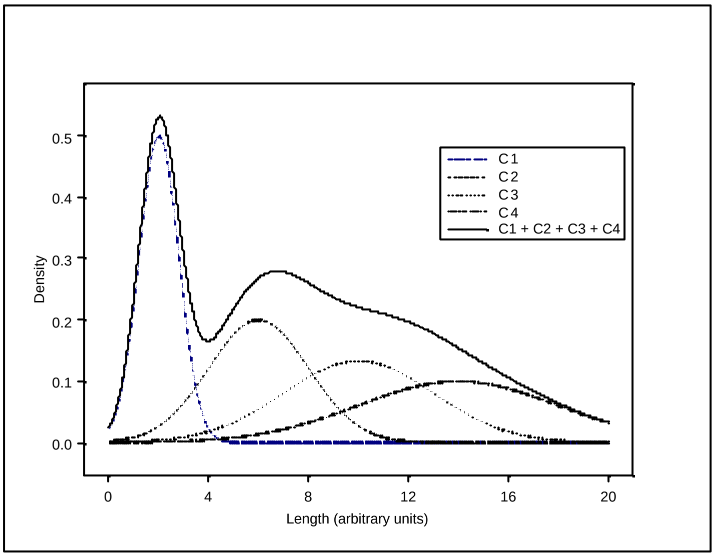


Figure 2 Schematic diagram of a mixture distribution generated from four length-at-age distributions C1 to C4 and their sum.

The model assumes that the length distributions follow a normal distributions with a constant coefficient of variation k, the expected density in length class j is the sum of n distributions in the length interval of the j^{th} class, given by

$$d_j = \sum_{i=t}^{n} D_i \left[\Phi\left(\frac{\mu_i - l_{j+1}}{k\mu_i} \right) - \Phi\left(\frac{\mu_i - l_j}{k\mu_i} \right) \right] \tag{1.2}$$

where l_j and l_{j+1} are the length bounds of the j^{th} length interval, D_i is the total density of animals aged i in the population, F (.) denotes the cumulative standard normal function, m_i is the mean of the length distribution for animals of age i. The values of the D_i, m_i and k are estimated by finding the values for them which result in the d_j having a good fit to the distribution of observed densities at length from surveys. The assumption of a constant coefficient of variation is reasonable since it implies that older animals exhibit a greater range of lengths. This assumption has the advantage of reducing the number of parameters to be estimated in fitting the model, and ensures an orderly relationship between the variance estimated for each mixture component. The estimated value of *R(t)* for the survey is given by:

$$R(t) = \frac{D_t}{\sum_{i=t}^{n} D_i} \tag{1.3}$$

Only D_t and the sum of the D_i need to be estimated accurately. The values of m_i, k and the individual $D_i \mid i > 0$ are 'nuisance' parameters. We need be only concerned that their values provide a good fit to the data; we are not particularly interested in their values, except that they should be consistent with what is known about the biology of your organism.

The major problem for the analysis is that the existing methods are not applicable to densities estimated from net haul surveys. Macdonald and Pitchers' (1979) method assumes that length frequency data have no unusual statistical properties. The usual method assumes that length frequency data are representative of a population, with the frequencies in each length class having Poisson distributions. This would be valid in the case where the animals in question are randomly and independently distributed, and the frequencies consist of a complete enumeration of all the samples.

Unfortunately most net haul survey densities do not have these statistical properties. The statistical distribution of net haul densities has to allow for an often substantial probability that a given haul will give a zero density estimate (ie. the net was empty). The statistics of such distributions have been examined by Aitchison (1955), and Pennington (1983) has recommended using Aitchison's delta distribution as the underlying statistical model when analysing net haul survey data. This recommendation is followed in the method developed here. The delta distribution consists of a discrete probability at the origin, and a lognormal distribution for the non-zero observations.

.

Simulation studies (de la Mare (1994b) show that the sampling distribution of the mean for delta distributions is highly skewed with the numbers of observations typical in trawl surveys. Simple transformations of the data or their mean do not lead to summary statistics which capture all the features of the sampling distribution. The likelihood for the sampling distribution of the mean cannot be expressed in terms of summary statistics, and so the full data set has to be used in calculating the likelihood of given parameter values. The delta distribution has the following probability function:

$$f(x;\, p,\lambda,\sigma^2) = (1-p)\mathrm{I}_0(x) + p\frac{1}{x\sqrt{2\pi\sigma}}\,\mathrm{e}^{-\frac{1}{2}\left(\frac{\ln x-\lambda}{\sigma}\right)^2}\,\mathrm{I}_{(0,\infty)} \qquad (1.4)$$

where p is the proportion of observations of x which are > 0, l and s^2 are the parameters of the lognormal distribution of the non-zero observations, $\mathbf{I}_0$ is an indicator function which takes the value 1 when $x = 0$ and 0 otherwise, and $\mathbf{I}_{(0,\infty)}$ takes the value 0 when $x = 0$ and 1 when $x > 0$. The first term is a discrete probability mass at the origin and the second term is a probability density. Figure 3 shows a schematic depiction of Equation 4.4.

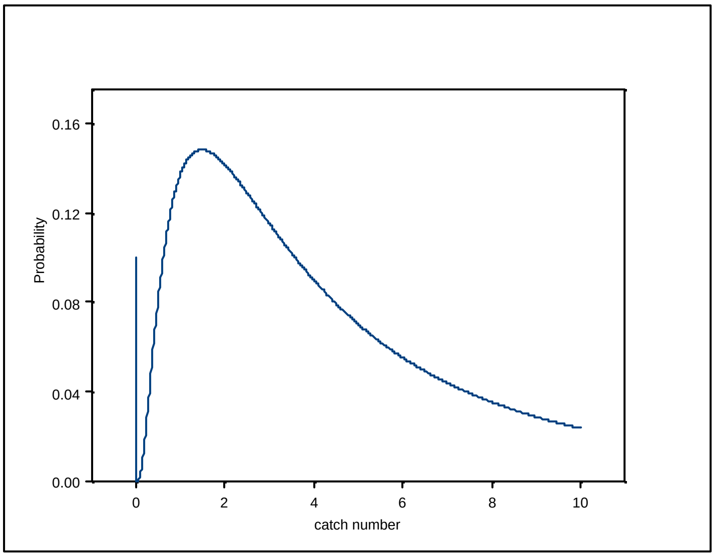


Figure 3 Aitchesons Delta consisting of the log normal distribution which shows the probability or likelihood of the number in a particular haul. This distribution includes a non-negative point at 0 which represents the case when a particular haul results in no catch. This scenario has p = 0.9.

The log-likelihood of a vector of observations $\mathbf{x} = x_1 \ldots x_N$ from a delta distribution is given by

$$\ln[L\left(x_1 \ldots x_N ; p, \lambda, \sigma^2\right) = (N-m)\ln\left(1-p\right) + m\ln p - \frac{m}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{x>0}\left(\ln x_i - \lambda\right)$$

(1.5)

$$-\sum_{x>0}\ln x_i - \frac{m}{2}\ln 2\pi$$

where *N* is the total number of observations and *m* is the number of non-zero observations. The last two terms are additive constants which can be ignored when maximising the likelihood function to calculate estimates. In the method described here, it is the densities for a given length class in each haul which constitute the x_i. Using Aitchison's (1955) formulae, the maximum likelihood estimate of the mean value of density in the j^{th} length class is:

$$\bar{d}_j = \frac{m}{N} e^{\bar{y}} G_m \left(\tfrac{1}{2} s^2\right), \qquad m > 1$$

$$\bar{d}_j = \frac{x_1}{N}\ , \qquad m = 1 \qquad (1.6)$$

$$\bar{d}_j = 0 \quad , \qquad m = 0$$

where $\bar{y}$ and s^2 are the sample mean and sample variance of the log of the non-zero observations and:

$$G_m(t) = 1 + \frac{m-1}{m}t + \sum_{j=2}^{\infty}\frac{(m-1)^{2j-1}}{m^j(m+1)(m+3)\ldots(m+2j-3)}\cdot\frac{t^j}{j!} \qquad (1.7)$$

Using a likelihood ratio approach (Cox and Hinkley, 1974), asymptotic confidence intervals on the mean density can be found as the roots of the following function:

$$\mathrm{q}(d) = \left[\ln L\left(\mathbf{x}; p, \lambda, \sigma^2\right) \middle| \, p = \frac{m}{N},\ \lambda = \frac{1}{m}\sum_{X_i>0}\ln x_i\ ,\sigma^2 = \frac{1}{m}\sum_{X_i>0}\left(\ln x_i - \lambda\right)^2\right]$$

(1.8)[2]

$$-\mathrm{Sup}\left[\ln L\left(\mathbf{x}; p, \lambda, \sigma^2\right) \middle| \, 0 < p \le 1,\ \lambda = \ln\left(\frac{d}{p\,\mathrm{G}_m\left(\frac{1}{2}\sigma^2\right)}\right), 0 < \sigma^2 < \infty\right] - \tfrac{1}{2}\chi_{1,\alpha}{}^2$$

where $\chi_{1,\alpha}{}^2$ is the critical value of the χ^2 distribution with one degree of freedom, at the α probability level. The maximum likelihood estimates of the parameters of the mixture

[2] Sup(x) (Supremum) A supremum operator will take the asymptotic maximum of a function.

distribution are obtained by maximising the sums of the log-likelihood's for each length class. It is useful to designate the parameters of the mixture distribution as:

R(t): the parameter of primary interest, and

q : a vector of the nuisance parameters consisting of D_{t+1} .. D_n, k and m_{t+1} .. m_n

The value of D_t used in calculating the mixture is derived from *R(t)* and the $D_{i>t}$ as:

$$D_t = \frac{R(t)}{1-R(t)} \sum_{i=t+1}^{n} D_i \tag{1.9}$$

The likelihood function for fitting the mixture distribution can be written as:

$$h(R_1;\mathrm{q}) = \left[\sum_{j=t}^{n} \mathrm{Sup}\left(\ln L\left(\mathbf{x}_j; p_j, \lambda_j, \sigma_j^{\,2}\right) \,\middle|\, 0 < p_j \le 1,\ \lambda_j = \ln\left(\frac{d_j}{p_j\, \mathrm{G}_m\left(\frac{1}{2}\sigma_j^{\,2}\right)} \right),\ 0 < \sigma_j^{\,2} < \infty \right) \right] \tag{1.10}$$

where d_j is the expected value of the density in length class *j* derived from equation (2), with the mixture distribution with parameters *R(t)* and q. Note that estimating *R(t)* and q requires maximising *h*(*R(t)*,q) which in turn requires maximising the likelihood function for the delta distribution in each length class. All these maximisation's have to be carried out numerically. The parameters p_j and $\sigma_j^{\,2}$ are also nuisance parameters. The maximisation's are carried out subject to the following constraints:

$$0 \le R(t) < 1$$

$$\mu_t^{\,-} \le \mu_t \le \mu_t^{\,+} < \mu_{t+1}^{\,-} \le \mu_{t+1} \le \mu_{t+1}^{\,+} < \ \dots\ < \mu_n^{\,-} \le \mu_n \le \mu_n^{\,+}$$

$$k^- \le k \le k^+$$

where a superscript + or - represents a numerically specified constraint. Apart from the well known advantages of statistical efficiency, working with log-likelihood allows asymptotic confidence intervals and variances to be calculated for the parameters. In particular we are interested in a variance estimate for *R(t)*. This is estimated from the second derivative of a quadratic function (Cox and Hinkley, 1974) passing through the points:

$$\left\{ \hat{R}(t)-\delta,\ h\left(R(t) = \hat{R}(t)-\delta \right) \right\},\ \left\{ \hat{R}(t),\ h\left(R(t) = \hat{R}(t) \right) \right\},\ \left\{ \hat{R}(t)+\delta,\ h\left(R(t) = \hat{R}(t)+\delta \right) \right\} \tag{1.11}$$

where d is small. In determining these points, R(t) is fixed as specified, but the vector of nuisance parameters is re-estimated by re-maximising the likelihood function. Thus the estimate obtained is for the marginal variance of R(t). Although asymptotic variance estimates are not always accurate for non-normal sampling distributions, they should be adequate for providing relative weights for the subsequent estimation of the distribution statistics of R(t) estimates.

Section 3 Installation

The package is distributed as an MS-DOS executable program called ‘CMIX.EXE’ together with the Excel Add-In ‘CMIX_Excel_Add-In.xla’. A setup program is supplied to install these files onto any windows operating system.

In order to install and operate the Excel Add-In. It is recommended to have:

- any Windows operating system except for Windows XP (as CMIX wont run under this version)
- At least a 386 processor
- 16 MB RAM

3.1 Installing CMIX and Excel Add-In

Complete the following steps to install CMIX and the Excel Add-In.

- Before continuing with the following installation procedure please make sure you have uninstalled any previous versions of the Add-In. See Section 3.2 ‘Uninstalling CMIX and Excel Add-In’.
- Double click the “setup.exe” file supplied in the base directory of the CD. This should handle the whole installation process on your PC.
- The default installation directory is C:\Program Files\CMIX\. It is recommended that you do not change the installation directory from this default directory.
- Various ActiveX controls and Dynamic-Linked Libraries (DLL’s) are supplied in conjunction with this installation. The setup process will prompt you if there are newer controls or DLL’s on your system than those being installed. Do not overwrite your system files in these cases.
- Once the required files have been installed, on some systems Excel is automatically launched and the CMIX Excel Add-In is loaded. You should see the CMIX Excel Add-In menu bar which can be dragged to any location. If Excel is not launched after completing the setup double click on the file ‘*Installation Directory*\install.xls’ to have the Add-In loaded for you.
- The Add-In will now be available to use with any Excel workbook. If you wish to turn the Add-In on or off at any stage, on the Excel file menu, choose Tools >> Add-Ins and deselect or select the CMIX Excel Add-In checkbox.

3.2 Uninstalling CMIX and Excel Add-In

If CMIX and the CMIX Excel Add-In have been installed using the 'setup.exe' installation program, use the following procedure to uninstall the software.

- In Excel, choose the menu options Tools >> Add-Ins and make sure the 'CMIX Excel Add-In is deselected and then close Excel.

- From the Windows Start menu select Settings >> Control Panel >> Add/Remove Programs.

- Select 'CMIX' from list and hit the 'Add/Remove' button. The un-installation procedure will begin. If prompted do not remove any components that are designated as shared components as this may affect the operation of other software.

- In Excel, choose the menu options Tools >> Add-Ins and click on the 'CMIX Excel Add-In' name. When prompted select 'Remove From List' and then close Excel.

If CMIX and the CMIX Excel Add-In were installed manually, use the following procedure to uninstall the software.

- In Excel, choose the menu options Tools >> Add-Ins and make sure the 'CMIX Excel Add-In is deselected and then close Excel.

- Locate the files 'CMIX.EXE' and 'CMIX_Excel_Add-In.xla' on your local system and delete them.

- In Excel, choose the menu options Tools >> Add-Ins and click on the 'CMIX Excel Add-In' name. When prompted select 'Remove From List' and then close Excel.

Section 4 Using CMIX Excel Add-In

This section illustrates how you can use the CMIX Excel Add-In easily to run CMIX and display CMIX output. The Add-In is essentially a toolbar containing 3 tools, CMIX Wizard, Run CMIX and Display CMIX Output (Figure 4).

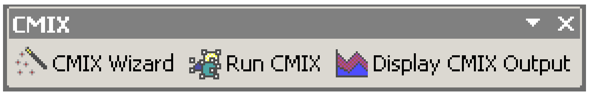


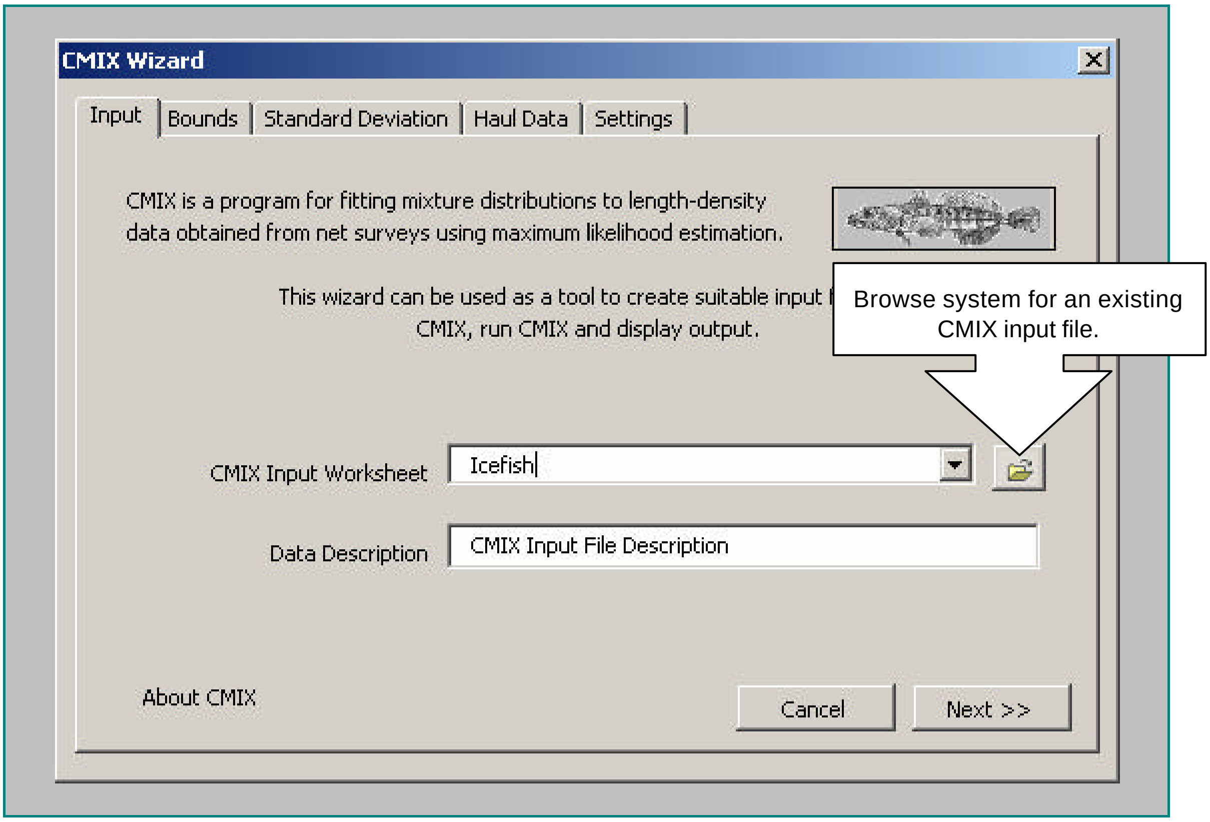
Figure 4 CMIX Excel Add-In Toolbar

4.1 Using CMIX Wizard

This section illustrates the use of CMIX Wizard to create input data suitable for running through CMIX. CMIX requires a carefully formatted input file, thus it is recommended that you use CMIX Wizard to create and edit your CMIX input files.

4.1.1 Setting Input Details

The first tab of the CMIX Wizard form (Figure 5) may be used to create a new input file to run through to CMIX or to load an existing CMIX input file.



CMIX is a program for fitting mixture distributions to length-density data obtained from net surveys using maximum likelihood estimation.

This wizard can be used as a tool to create suitable input f CMIX, run CMIX and display output.

CMIX Input Worksheet: Icefish

Data Description: CMIX Input File Description

Browse system for an existing CMIX input file.

Figure 5 CMIX Wizard Input Tab.

- To create a new input file simply type in a new name for the worksheet that the input data will be loaded into. The wizard will automatically append the name with “_Input”. Also enter a description for the input file you will be creating.

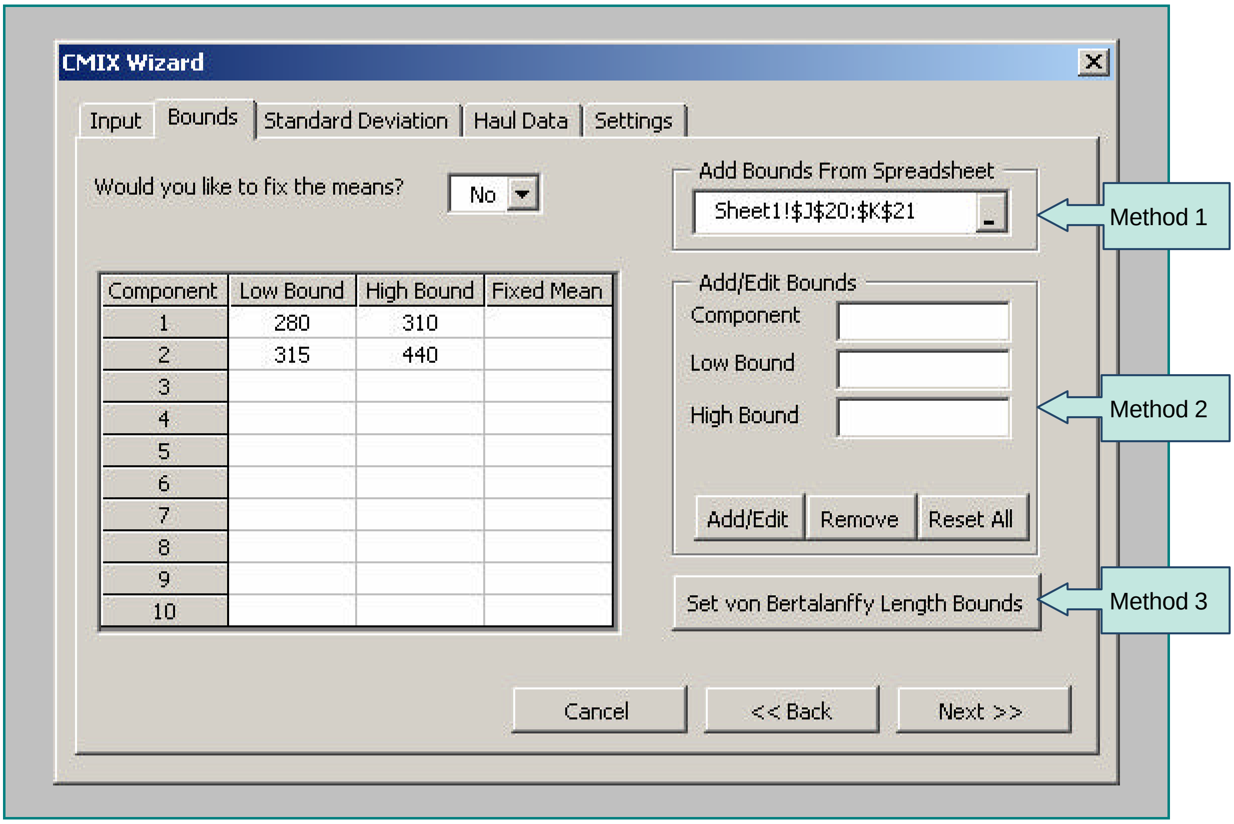
- To load an existing input file, you may use the browse button to search for the input file. The input file *MUST* contain data in the correct format suitable for running through CMIX. For information on the correct structure of a CMIX input file please see Appendix 2. All data contained in the input file will automatically be loaded into the wizard ready for editing.

- To load an existing input file from a worksheet in the current Excel workbook, select the required worksheet from the drop-down list. All data contained in the worksheet will be automatically loaded into the wizard ready for editing.

Note that the prefix used for the input worksheet will be used to name worksheets displaying output resulting from the CMIX run at a later point.

4.1.2 Setting Mixture Component Bounds

The ‘Mixture Component Bounds’ tab (Figure 6) specifies the bounds within which the means for each mixture component can be fitted. You can choose whether to fix one or more of the means. The bounds still need to be specified if you decide to fix the means. There are 3 methods for setting the mixture component bounds as seen in Figure 6.



Would you like to fix the means? No

Component	Low Bound	High Bound	Fixed Mean
1	280	310	
2	315	440	
3			
4			
5			
6			
7			
8			
9			
10			

Add Bounds From Spreadsheet: Sheet1!$J$20:$K$21

Add/Edit Bounds
Component
Low Bound
High Bound

Figure 6 Mixture Component Bounds Form

4.1.2.1 Method 1 – Add Bounds From Spreadsheet

Click on the corner of the 'Add Bounds From Spreadsheet' tool to select data from any table within the current active worksheet. Use the mouse to select the data by drawing a bounding box around the data. You may not select empty cells or cells containing non-numeric data.

4.1.2.2 Method 2 – Add/Edit Bounds

You manually enter the required bounds in the 'Add/Edit Bounds' frame. The text box for entering a fixed mean value is only visible if the 'Would you like to fix the means?' checkbox has value 'Yes'. After entering bounds for a component you need to hit the 'Add\Edit' button to have the bounds added to the table.

Bounds do not need to be entered in order, the wizard will re-order them. If bounds overlap the wizard will inform you and ask you to modify the bounds till they are correct.

To edit bounds that have already been added you need to click on the desired row of the bounds table and the selected bounds will appear in the 'Add/Edit Bounds' frame where they can be edited. Note that edits will not appear updated in the bounds table until the 'Add/Edit' button has been hit.

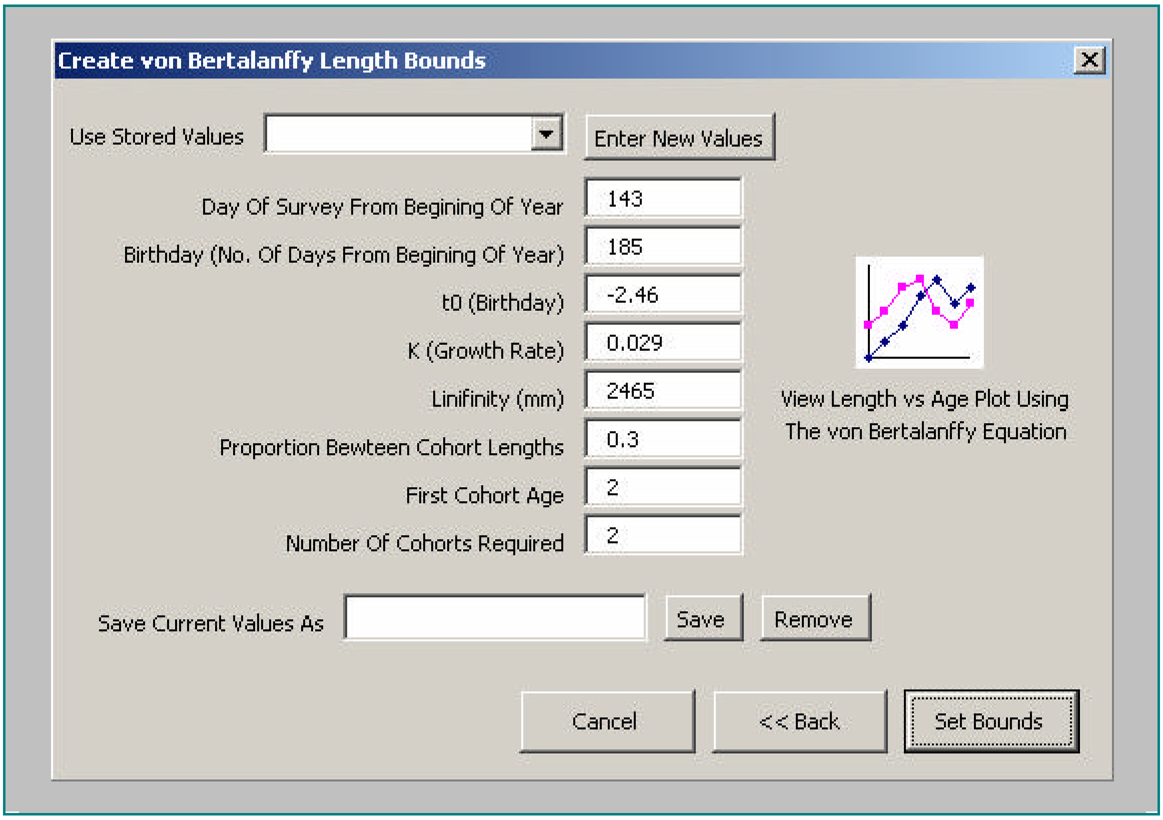
4.1.2.3 Method 3 – Set von Bertalanffy Length Bounds

Click the 'Set von Bertalanffy Length Bounds' to let the wizard automatically calculate the bounds from user specified growth parameters using the von Bertalanffy equation. Figure 7 will be displayed and needs to be completed with suitable parameters. Click on the 'Enter New Values' button to begin entering a new set of parameters.

See section 4.3 'von Bertalanffy Equation', for information about how the von Bertalanffy equation is used to estimate bounds and for information on the parameters used see Appendix 1.

The Length vs Age plot generated using the von Bertalanffy equation can be viewed also (Figure 8). Upon clicking the 'Set Bounds' button if the parameters have been set correctly, the wizard takes you back to the 'Mixture Component Bounds Form' where the von Bertalanffy generated Bounds have been set automatically. If the means on the bounds form were chosen to be fixed, the fixed means are set along with the bounds.

User defined default von Bertalanffy parameters can be stored within the wizard for subsequent uses of the Add-In. Every time the Add-In is closed or Excel is closed the default parameters are saved with the Add-In for use at a later date. To save a set of parameters give it a name in the "Save Current Values As' textbox and click the 'Save' button. To use a default set of values choose a set from the 'Use Stored Values' drop-down list. *Be careful of not losing default von Bertalanffy parameters if you re-install the software.*



Create von Bertalanffy Length Bounds

Use Stored Values	
Day Of Survey From Begining Of Year	143
Birthday (No. Of Days From Begining Of Year)	185
t0 (Birthday)	-2.46
K (Growth Rate)	0.029
Linifinity (mm)	2465
Proportion Bewteen Cohort Lengths	0.3
First Cohort Age	2
Number Of Cohorts Required	2
Save Current Values As	

Figure 7 Create von Bertalanffy Length Bounds Form

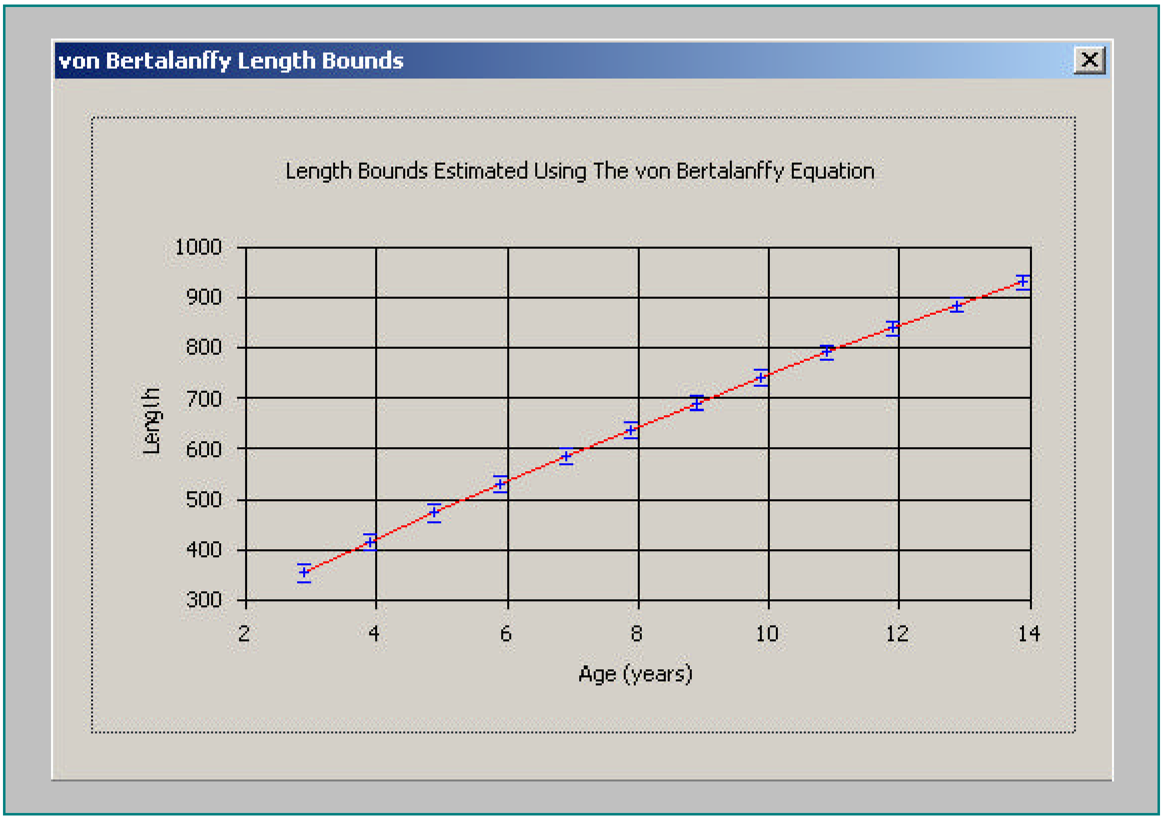


Figure 8 Plot of von Bertalanffy Length Bounds

4.1.3 Standard Deviation Settings

The ‘Standard Deviation’ tab (Figure 9) provides a means to set the data required to determine the estimates for the standard deviations. The user can specify whether the standard deviations are linearly or independent functions of the mean by selecting the relevant option from the drop-down list.

Data for both the ‘Linear’ and ‘Independent’ tabs need to be completed irrespective of whether you choose the standard deviations to be linearly related to, or independent of the mean. This is a requirement of CMIX. Both types of input are kept in the CMIX input file so that the user can change from one to the other without having to have two different versions of the data file. However, only the option selected (Linear or Independent) is evaluated. Defaults are automatically set for both forms, so the minimum amount of changes can be performed.

4.1.3.1 Linear Standard Deviation Settings

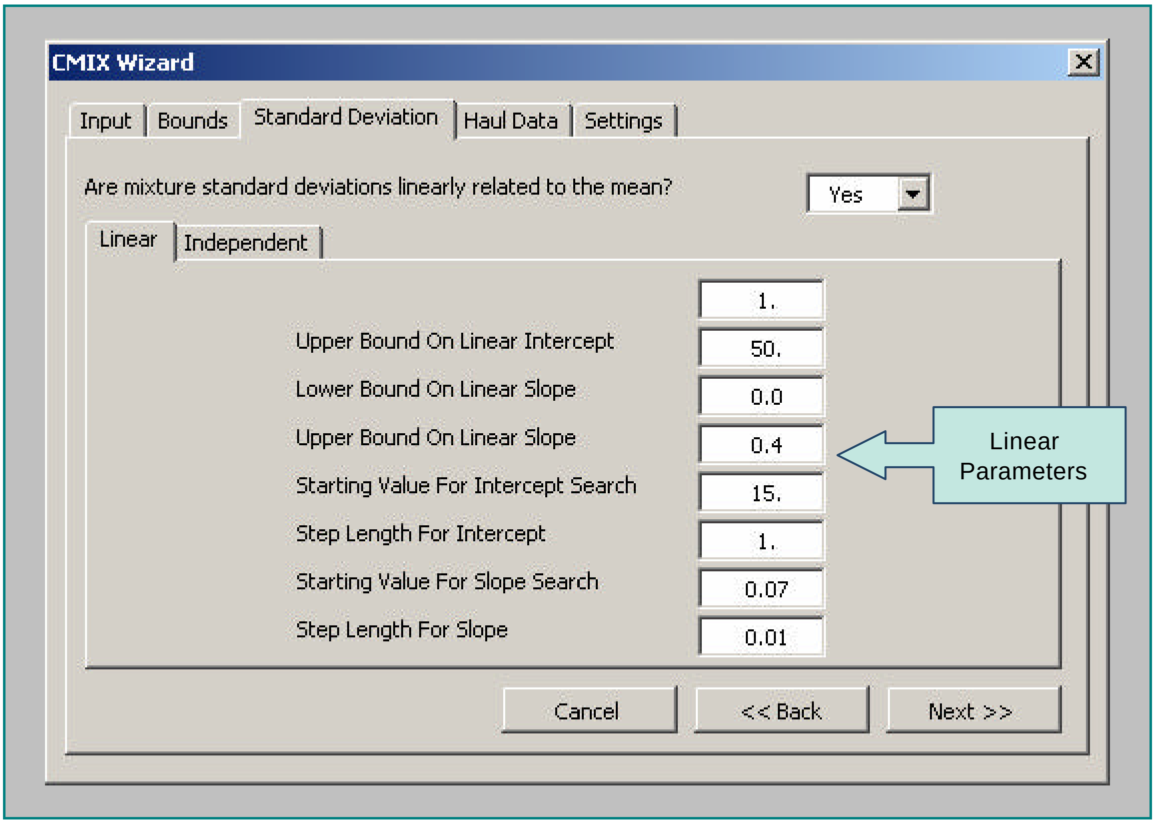
Usually we would expect that the standard deviations for the components of a length-density distribution will increase monotonically with the mean. If the linear tab is selected then the standard deviations of the mixture components are a linear function of the mean, that is:

$$s_i = a + bm_i$$

where s_i and m_i are the standard deviation and mean of component *i* respectively and *a* and *b* are constants which may be estimated from the data. If the intercept (*a*) is held fixed at 0 then the mixture components will have a constant coefficient of variation estimated by the slope (*b*). If *b* is held fixed at 0 then the mixture components will have a constant standard deviation estimated by *a*. These possibilities are allowed for when setting bounds and parameters.

Using the linear option ensures that the standard deviations behave in an orderly fashion.

Figure 9 and Figure 10 is the standard deviations setting tab with the linear and the Independent tab selected respectively. The input for both tabs is described below together with some suggestions for reasonable starting values for parameters.



Are mixture standard deviations linearly related to the mean? Yes

	1.
Upper Bound On Linear Intercept	50.
Lower Bound On Linear Slope	0.0
Upper Bound On Linear Slope	0.4
Starting Value For Intercept Search	15.
Step Length For Intercept	1.
Starting Value For Slope Search	0.07
Step Length For Slope	0.01

Figure 9 Setting Linear Standard Deviation Settings

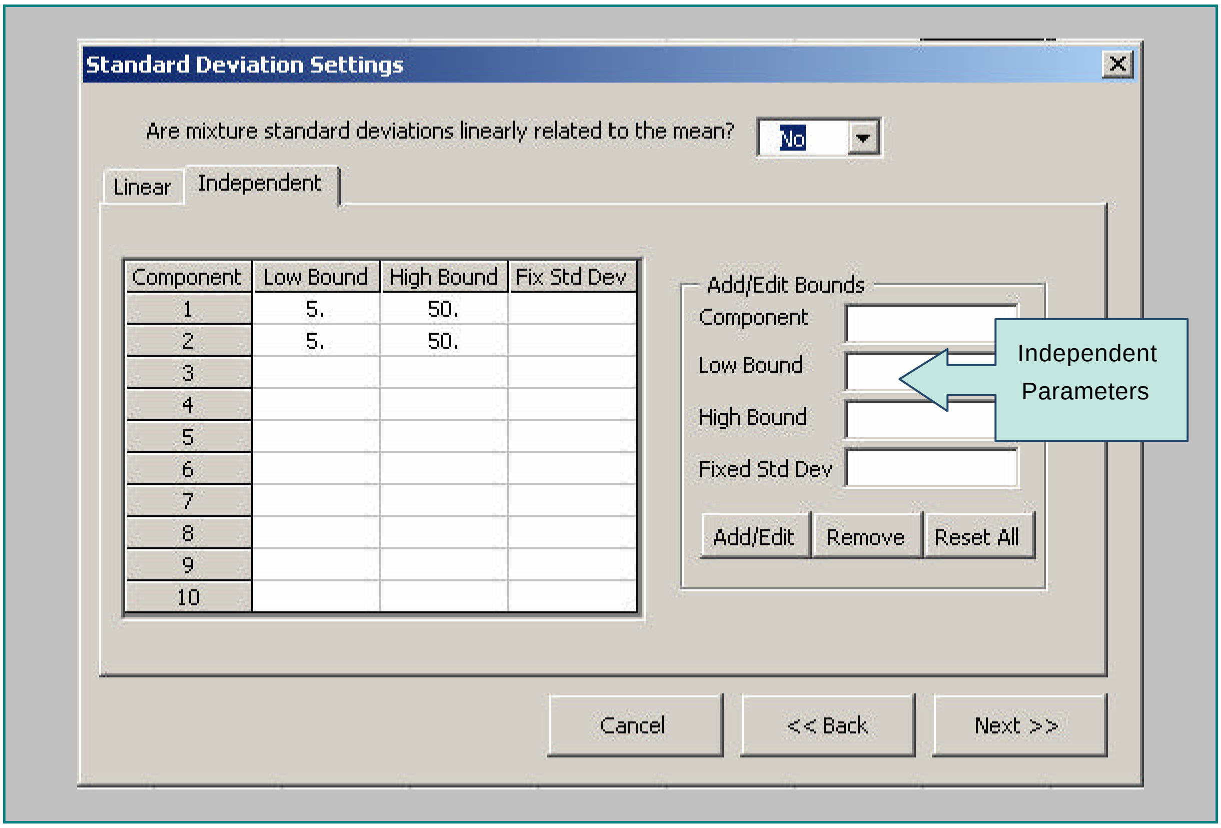
Table 1 Linear Parameters

Lower Bound on Linear Intercept	The lower bound desired for fitting the relationship between the SD and the mean.
Upper bound on Linear Intercept	The highest bound desired.
Lower Bound on Linear Slope	The lower bound on linear slope should be > 0 if the constraint is to be enforced that the mixture standard deviations increase with the mean. Set lower bound = 0 if SD is to be constant
Upper Bound on Linear Slope	Upper bound on slope must be greater than the lowest bound.
Starting Value for Intercept Search	Enter a starting value for Intercept search. A reasonable starting point is ‘1’.
Step Length for Intercept	Enter the step length for intercept. A good starting value is 0.1. If you don’t want the starting value to change set to 0.0.
Starting Value for Slope Search	Enter a starting value for slope search , a reasonable starting value is 0.05. If you don’t want the starting value to change set to 0.0.
Step Length for Slope	Enter the step length for the slope a reasonable starting value is 0.005. If you don’t want the starting value to change set to 0.0.

NOTE: When the starting value for the intercept and its step length are both set to zero, the estimated slope is a constant coefficient of variation for the mixture components. When the starting value and step length for the slope search are both set to zero, the estimated intercept is a constant standard deviation for the mixture components. The slope and intercept cannot both be simultaneously fixed at zero, although either or both can be fixed at non-zero values.

4.1.3.2 Independent of Mean Standard Deviation Settings

If the ‘Independent’ tab is selected, the standard deviations can take any values within specified ranges given in this form. Any or all of the standard deviations can be forced to specified values in this form. Figure 10 shows this form and descriptions of the input required, follows.



Standard Deviation Settings

Are mixture standard deviations linearly related to the mean? No

Component	Low Bound	High Bound	Fix Std Dev
1	5.	50.	
2	5.	50.	
3			
4			
5			
6			
7			
8			
9			
10			

Add/Edit Bounds

Component

Low Bound

High Bound

Fixed Std Dev

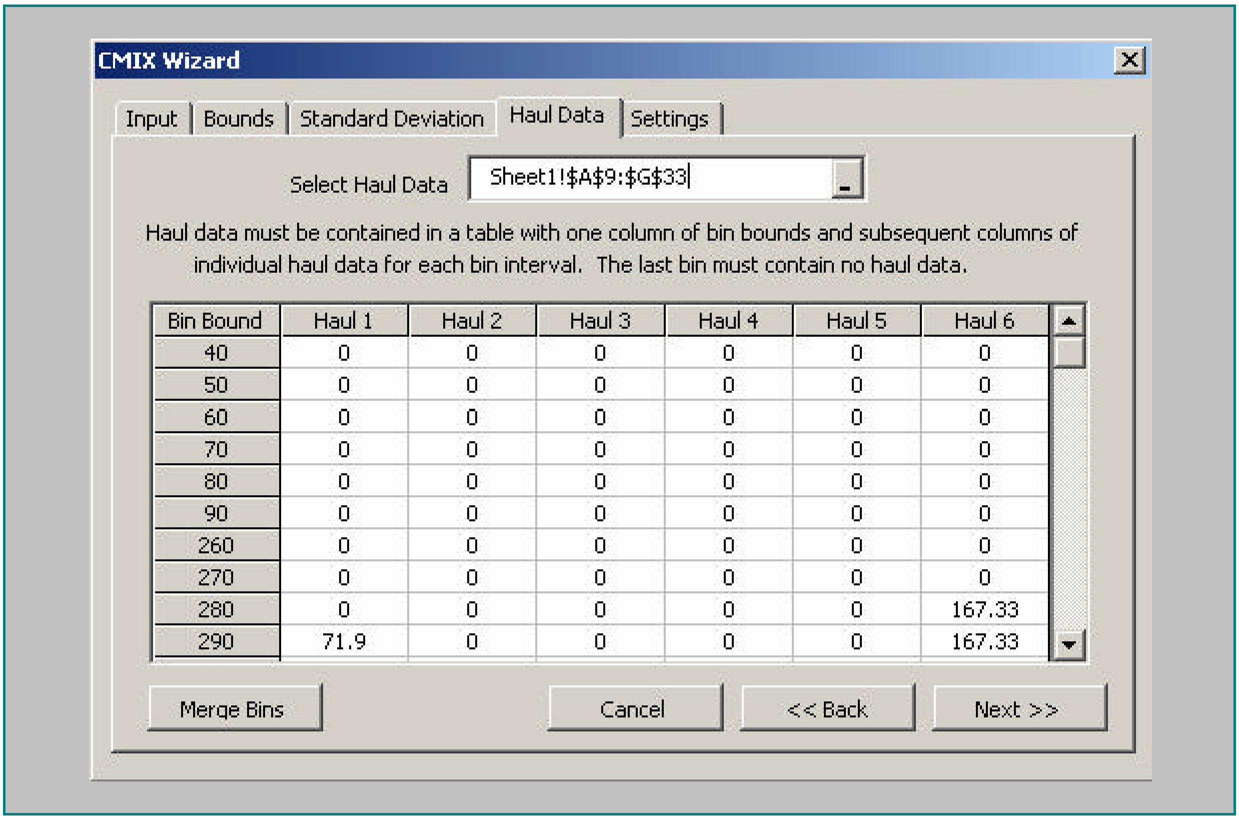
Figure 10 Setting Independent Standard Deviation Settings

Default low and high bounds are always present for the number of components specified on the ‘Bounds’ tab of 5 and 50 respectively. These may be edited as you require by selecting the relevant row from the bounds table. You may edit the values in the ‘Add\Edit’ bounds frame. The table will be updated upon hitting the ‘Add/Edit’ button.

NOTE: Unless the bounds are selected such as to impose restrictions on the values of the standard deviation estimates, it is possible that the standard deviation for a high mean will be numerically smaller than for a low mean.

4.1.4 Selecting Haul Data

The user can select the haul data using the ‘Haul Data’ form shown in Figure 11. A description of the import process follows.



Select Haul Data Sheet1!$A$9:$G$33

Haul data must be contained in a table with one column of bin bounds and subsequent columns of individual haul data for each bin interval. The last bin must contain no haul data.

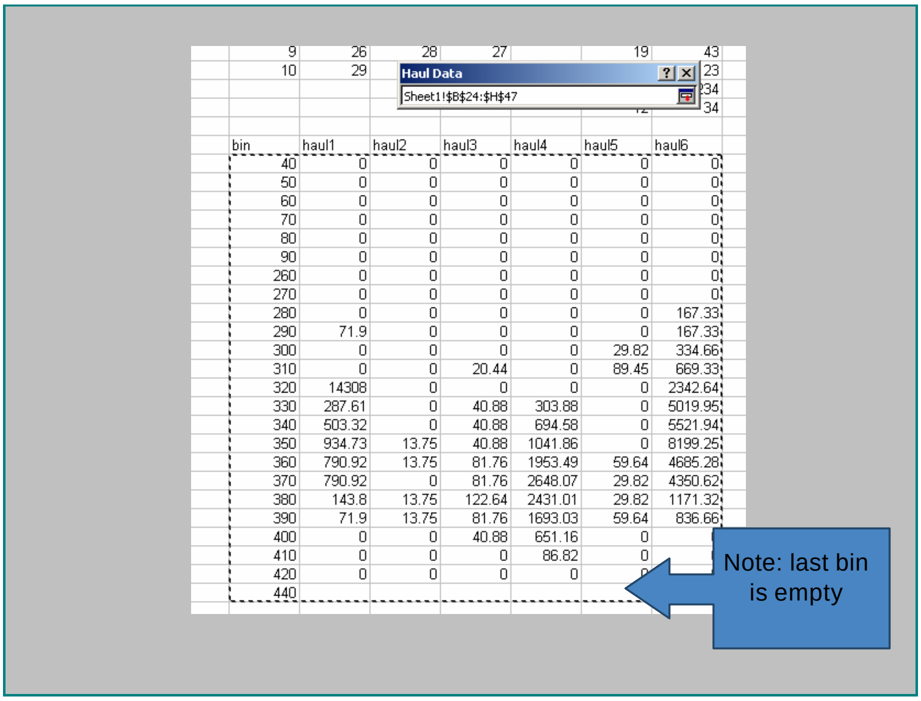
Bin Bound	Haul 1	Haul 2	Haul 3	Haul 4	Haul 5	Haul 6
40	0	0	0	0	0	0
50	0	0	0	0	0	0
60	0	0	0	0	0	0
70	0	0	0	0	0	0
80	0	0	0	0	0	0
90	0	0	0	0	0	0
260	0	0	0	0	0	0
270	0	0	0	0	0	0
280	0	0	0	0	0	167.33
290	71.9	0	0	0	0	167.33

Figure 11 CMIX Wizard Haul Data Tab

The ‘Select Haul Data’ tool allows the user to select data from any table in the current worksheet (Figure 12). The first column contains the bin interval boundary data, and subsequent columns the haul data. There is no limit on the number of bins or hauls. The bin boundaries do not need to be all the same width. The last bin however, must contain no haul data, as the bin boundaries in the table are lower bounds and the last bin bound closes off the last bin.

Haul data can be merged over several bins by using the ‘Merge Bins’ button. Select the rows of haul data in the table you wish to be merged and hit the ‘Merge Bins’ button. The haul data for each bin selected will be summed and placed in one new bin, with the bounds of the bin ranging from the lowest to the highest bin selected. This step is irreversible. It is a useful function for grouping bins comprising of zeros or very low densities which can prevent the analysis from successfully minimising.

It is important that all hauls are included in the input data, even those where no fish were caught in any of the length bins.



Sheet1!$B$24:$H$47

9	26	28	27		19	43
10	29					23
						234
						34

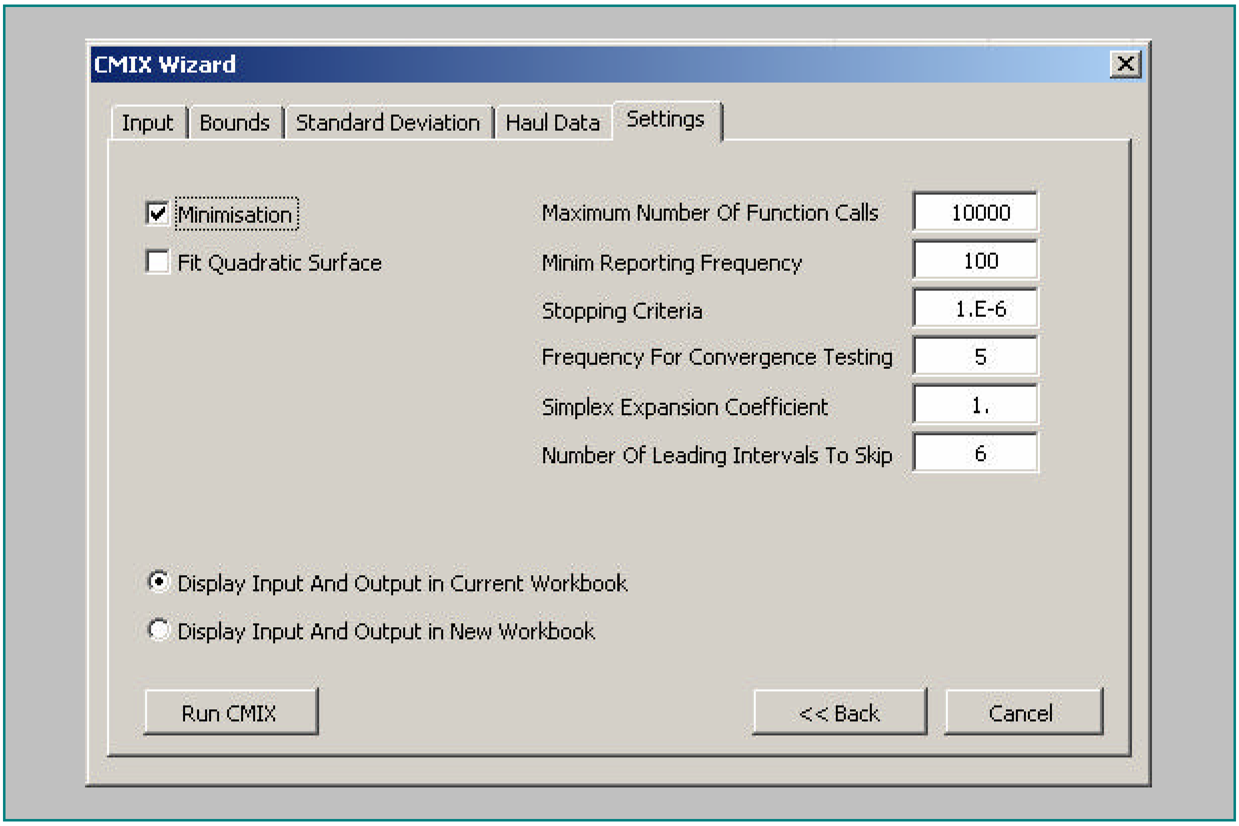
bin	haul1	haul2	haul3	haul4	haul5	haul6
40	0	0	0	0	0	0
50	0	0	0	0	0	0
60	0	0	0	0	0	0
70	0	0	0	0	0	0
80	0	0	0	0	0	0
90	0	0	0	0	0	0
260	0	0	0	0	0	0
270	0	0	0	0	0	0
280	0	0	0	0	0	167.33
290	71.9	0	0	0	0	167.33
300	0	0	0	0	29.82	334.66
310	0	0	20.44	0	89.45	669.33
320	14308	0	0	0	0	2342.64
330	287.61	0	40.88	303.88	0	5019.95
340	503.32	0	40.88	694.58	0	5521.94
350	934.73	13.75	40.88	1041.86	0	8199.25
360	790.92	13.75	81.76	1953.49	59.64	4685.28
370	790.92	0	81.76	2648.07	29.82	4350.62
380	143.8	13.75	122.64	2431.01	29.82	1171.32
390	71.9	13.75	81.76	1693.03	59.64	836.66
400	0	0	40.88	651.16	0	[illegible]
410	0	0	0	86.82	0	[illegible]
420	0	0	0	0	0	[illegible]
440						

Note: last bin is empty

Figure 12 Selecting Haul Data From An Excel Worksheet

4.1.5 Final CMIX Settings and Parameters

The Settings Form (Figure 13) accepts the instructions for visual output and sets some technical parameters (See Table 2 for parameter descriptions). From here the user can instruct CMIX Wizard to run CMIX and the output is automatically displayed upon completion. See Section 4.4 for output details.



☑ Minimisation

☐ Fit Quadratic Surface

Maximum Number Of Function Calls	10000
Minim Reporting Frequency	100
Stopping Criteria	1.E-6
Frequency For Convergence Testing	5
Simplex Expansion Coefficient	1.
Number Of Leading Intervals To Skip	6

● Display Input And Output in Current Workbook

○ Display Input And Output in New Workbook

Figure 13 Settings Tab

Table 2 Parameters

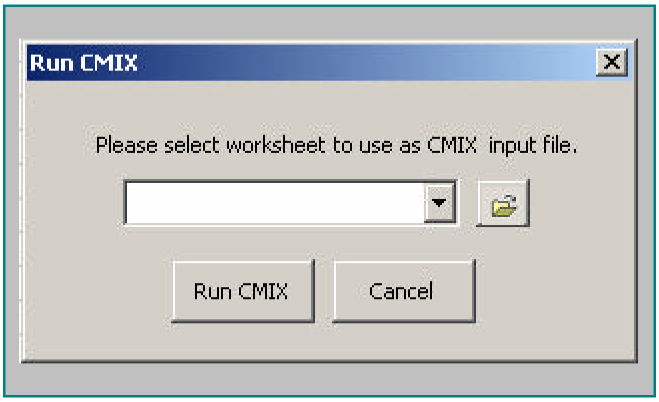
Minimisation	This parameter specifies whether estimates are to be made by minimising the residual function over the parameter space. If not selected, the residual function is evaluated once, at the starting values of the parameters.
Fit Quadratic Surface	The minimisation routine can fit a quadratic function to the region of the minimum, and hence give normal theory approximations to the information matrix pertaining to the estimated parameters. This is not necessarily reliable, and **its use is not recommended** for this program.
Maximum Number of Function Calls	This parameter specifies the maximum number of function calls allowed in searching for the best fit to the data. The value given here, of 10,000 should be adequate in most cases.
Minimum Reporting Frequency	The minimisation routine will report the parameters and residual function value at the regular intervals specified here. This output is useful for checking that the procedure has converged reliably, and on a minimum in the range covered in a given run. A specified value less than zero inhibits reporting.
Stopping Criteria	This is a technical parameter of the minimisation routine, and controls how little the values of the residual function should vary with changes in the estimated parameters before the minimisation will end. In other words, the minimisation will end when the change in the residual function is smaller than or equal to the magnitude of the stopping criterion. Failure of the procedure to converge even though the function reports indicate that convergence has occurred (i.e. the residual function values in the final function calls are more or less the same) is symptomatic that the stopping criterion is too small. Conversely, if there are few function calls before convergence is reported combined with obvious variation remaining between values of the residual function, indicates the stopping criterion is too large
Frequency for Convergence Testing	This is another minimisation technical parameter, 5 is usually satisfactory.
Simplex Expansion Coefficient	This is a minimisation routine technical parameter, which only applies if quadratic surface fitting is enabled. Adjusting it can improve the correspondence between the minima of the fitted quadratic and the minimum found by the search.
Number of Leading Intervals to Skip	This parameter can be used to specify whether the length frequency data are to be truncated to the left. In this example the specification means that the first 6 intervals are not used in fitting. This means that the intervals less than 260mm will not be used in fitting the mixture distribution. Specifying zero means that all the intervals will be used.

4.2 Run CMIX

The ‘Run CMIX’ tool can be used to run CMIX on an input file or an input worksheet if changes do not need to be made to the data.

Firstly a form will be displayed asking you to choose between displaying files in the existing workbook or in a new workbook.

Then the form shown in Figure 14 will be displayed where you can either browse for an input file using the browse button or select an input worksheet from the drop-down list. Hit the ‘Run CMIX’ button and CMIX will run on the input data selected and automatically display the output in Excel (see Section 4.4).



Run CMIX

Please select worksheet to use as CMIX input file.

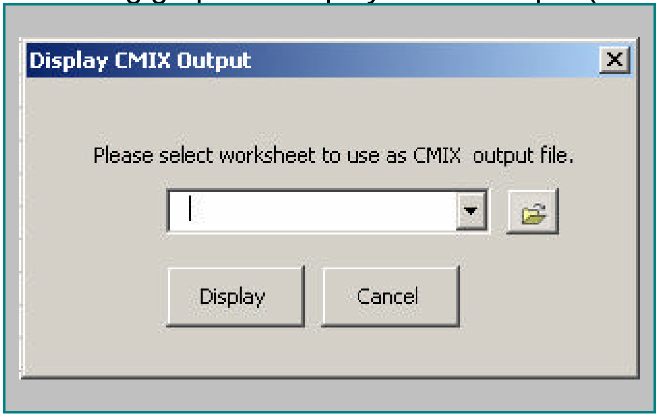
Figure 14 Run CMIX form

4.3 Display CMIX Output

The ‘Display CMIX Output’ tool can be used to display output for any CMIX output file or output worksheet.

A form will be displayed asking you to choose between displaying the files in the existing workbook or in a new workbook.

The form shown in Figure 15 will be displayed where you can either browse for an output file using the browse button or select an output worksheet from the drop-down list. Hit the ‘Display Output’ button and the output will automatically be imported into Excel as various worksheets containing graphical displays of the output (see Section 4.4).



Display CMIX Output

Please select worksheet to use as CMIX output file.

Figure 15 Display CMIX Output Form

4.4 Output

Once CMIX has finished execution and the DOS window running CMIX has been closed (either manually or automatically) the output will be imported to Excel and presented graphically. Five output worksheets will be imported into your current Excel workbook with names:

*Name*_Output

Contains the output file from CMIX loaded as an Excel worksheet.

Name _Results

Contains the results of the CMIX run such as the calculated length means of the mixture components and the density of the mixture components.

Name _Distribution

Contains a density vs length distribution plot.

*Name*_Density Plot

Contains a observed and expected density vs length plot. Note that this worksheet contains checkboxes underneath the plot to alter the series displayed. You may choose to show or hide the confidence intervals, standard error bars, normal mixture distributions, observed and expected densities.

It is essential to view the quality of the fit, particularly when there is a possibility of fitting a component in a region where there are no or few density values for a range of length classes. Length classes for which all observations have zero density make no contribution to the likelihood function. Therefore, a mixture component in such regions may be completely spurious, and hence bias the estimate. In such cases constraints on the range of length classes to be included in the fit, as well as on the mixture components are required to attempt to produce a sensible fit. The quality of fit plot is used as the principle method for performing adjustments to the values submiited as input to the program.

*Name*_Residuals

Contains a plot of the residuals.

Where *Name* is the name you assigned as the prefix to the CMIX worksheets.

You will need to check that the output from the run is valid. Please see Section 6.2 which outlines methods to check the validity of output.

Section 5 CMIX And DOS

5.1 Running CMIX From DOS Command Prompt

The CMIX executable file is located in the installation directory chosen during the setup process (See Section 3).

The command line for executing the program from a MS-DOS command prompt is:

cmix <input file> <output file>

The program prompts for input and output filenames if they are not specified in the command line.

You will need to make sure that the input file has been placed in the same directory as the CMIX executable.

For details on the format of the input and output files please see Appendices 2,3 and 4. An example input file will have been installed during the setup process and can be found in *Installation Directory*\Examples\Input.dat.

CMIX also outputs a plot file (written in HGRAPH archive format) and is given the default filename ‘PLOT05.DAT’.

5.2 Plotting CMIX Output From DOS Command Prompt

The MSDOS version of the program can also plot various graphs on a screen, HPGL compatible plotter, or IBM/EPSON graphics compatible printer. Plotting is done using subroutines from the HGRAPH library by Heartland Software Inc. 234 S. Franklin, Ames, IOWA 50010, USA.

The CMIX program produces a plot file of the mixtures with default name ‘PLOT05.DAT’. This default file is overwritten each time CMIX is run, so if you want to keep the file, rename it!

The file can be plotted using the program ‘vtrans.exe’, which is located in the CMIX installation directory.

The program prompts for the name of the file to plot and then 'iunit' which specifies the output device.

iunit = 0 (plots to screen)
iunit = 1 (plots to screen)
iunit = 2 (plots to plotter)
iunit = 3 (plots to printer in portrait mode)
iunit = 4 (plots to printer in landscape mode)

e.g

```
input file name: Plot05.dat
iunit: 1
```

Section 6 Tips When Fitting Mixtures

6.1 Input Data

- Be very careful to format the input file correctly. Using the Excel Add-In will assist in producing the correct input for CMIX.

- Include all valid samples (data blocks), including those with no catch.

6.2 Output data

- Always check that the program has run to completion by carefully checking the results in the output file which is displayed by the Excel Add-In as an appended worksheet.

- Check that in the Excel worksheet '*Name*_Results' the sum of the observed densities should not differ by a large amount to the sum of the expected densities. If you are getting a large difference between your observed and expected densities then alter the expected configuration of the mixtures.

- Check that the program has been able to converge successfully during the minimisation procedure. You can check this by scrolling up a number of pages in the Excel worksheet 'name_Output' until you find the line:

  END OF SEARCH
  ****
  This means the program has run successfully. If the program couldn't converge it will tell you something along the lines of

  ERROR! COULD NOT CONVERGE
  ****
  If this occurs then change the expected configuration of the mixtures or increase the maximum number of function calls that the program can use.

6.3 Achieving a Good Fit

Experience has shown that a good fit can often be obtained if you have a good understanding of the growth parameters. Using the von Bertalannfy growth curve (remember to take into account the difference between your sampling time and the birthday of the fish) (see Section 4.1.2.3 and Appendix 1).expected mean lengths at age can be determined. Try fixing the mean lengths and allow the standard deviations to be linearly related to the means. This often results in a better fit than setting the bounds around the expected mean. The standard deviations around the mean should increase with increasing mean length.

When fixing means and allowing the standard deviations to be linearly related, the following settings are a good starting point.

STARTING VALUE FOR INTERCEPT SEARCH	1.
STEP LENGTH FOR INTERCEPT	0.
STARTING VALUE FOR SLOPE SEARCH	0.05

6.4 Bias in length-densities arising from gear selectivity or sampling pattern

The sampling program may sample only a portion of the population. If that is the case then some cohorts may not be sampled at all and/or gear may selectively sample a specific size range. Such information can be used to restrict the data used in the input file (truncate the lower or upper ranges). In this case, the lower range can be truncated by skipping a number of length bins. The upper range should be truncated by excluding those bins not required from the input file.

6.5 Missing cohorts

A better fit will result by leaving out cohorts known to be missing or in very low abundance. This can be done simply by excluding them when nominating the characteristics of the components. Although the components need to be numbered consecutively, the characteristics do not have to represent consecutive cohorts (e.g. age classes).

6.6 Size of length bins and Poorly sampled length bins

The time taken to fit a mixture and whether or not a successful minimisation is obtained can be affected by the amount of zero data present, particularly if there is a vast array of length bins with zero data across many of the samples. In this case, reducing the range of the data (discussed under ‘bias’) or increasing the size of the length bins may yield a better result. The size of length bins does not need to be constant across the range (see the discussion under data input concerning bin intervals).

6.7 Excluding messy data

Messy data at each end of the range of lengths may disrupt the fit. Restricting the range as described under ‘bias’ may help in this case. Another method for dealing with messy data is to allow for a component to have a larger or restricted standard deviation than might be expected from ordered growth (eg. such as arising from biases) combined with a different mean size to take account of this. Changing bin sizes as well as collapsing all the messy data into a single bin might help this as well (see the discussion under data input concerning bin intervals).

6.8 R1 – The proportion of sample in first mixture component

PLEASE NOTE: This version of CMIX does not explicitly return a value for R1 (The proportion of sample in first mixture component) nor its standard error. If these values are required it is recommended that you use an earlier version of the mixtures program MLMIX. These values will be included in the next version of CMIX.

6.9 Error Messages

Most of the errors arising from using CMIX result from errors in formatting the input file (see above). It is hoped that the use of 'CMIX Wizard' will dramatically reduce the number of errors users encounter.

Some points to note:

- Input files must have **'.DAT'** extension.

- If you look in the output file it will show you when the program stopped and give you a clue where to make changes in the input file, either formatting or input values.

- Error 151 means that CMIX can't find the input file, it could be missing, named incorrectly or have more than 8 characters in the name.

- If you get the error **'Root bracketed for ZBRENT'** then part of the minimisation was unable to be completed. It is likely that you are asking too much of the program given the expected mixture distributions set up in the input file combined with the amount of data available **-** you are likely to have too many zeros in your input data and you may not get a reliable result. Although you can press the **[ENTER]** key several times to force the program continue, this is not recommended as the result may be unreliable.

APPENDIX 1 von Bertalanffy Equation

1.1 The Equation

The von Bertalanffy Equation (Equation (1.11)) is a commonly used to describe growth in fisheries modelling.

$$L_t = L_\infty\left(1-e^{-K(t-t_0)}\right) \tag{1.12}$$

where L_t is the expected size at age *t*, L_∞ is the asymptotic length (the length at which growth rate is theoretically zero), *K* is the Brody growth rate parameter (rate of growth towards asymptote), t_0 is the projected time (age) when length would have been zero on the modelled growth trajectory.

CMIX Wizard uses an adjusted form of the von Bertalanffy equation (see Equation (4.12) to generate length bounds for input into CMIX (see Section 4.1.2.3). The modified equation is used to correct for the time difference between the time of survey and the animals arbitrary birthday.

$$L_t = L_\infty\left(1-e^{-K\left(t-\left(t_0+t_{adj}\right)\right)}\right)\text{where } t_{adj} = t_s - t_b \tag{1.13}$$

Here t_0 is the projected time (age) when length would have been zero on the modelled growth trajectory, and t_{adj} is the difference between t_b and t_s. Where t_b is time between arbitrary birthday and January 1st (arbitrary birthday is date when animals are assumed to enter the next age year), and t_s is time between date of survey and January 1st.

1.2 The Parameters Used By CMIX Excel Add-In

Day of Survey From Beginning of Year	Time between date of survey and January 1st.
Birthday (No. of Days from Beginning of Year)	Time between birthday (can be arbitrary) and January 1st. Arbitrary Birthday is date when animals are assumed to enter the next age year.
T_0 (Birthday)	Enter the time at which length is zero on the modelled growth trajectory.
K (Growth Rate)	Enter the von Bertalanffy growth coefficient
$L_{inifinity}$ (mm)	Enter the maximum size in sample or population?
Proportion between Cohort Lengths	Enter the value for the proportion between cohort lengths for use in generating bounds.
First Cohort Age	Place here the age of the first cohort.
Number of Cohorts Required	Enter the number of cohorts required.

APPENDIX 2 CMIX Input File Format

CMIX requires a carefully formatted input file. Sample data files are included on the distribution CD. It is suggested that CMIX Excel Add-In and CMIX Wizard be utilised in creating input files for CMIX as it already handles the complex formatting of the input file.

The input file must have the format of an MS-DOS filename consisting of at most 8 characters followed by "**.DAT**" extension.

Range checking is carried out on input data. It is essential to have the correct number of data items, and in the correct order. With the combination of range checking and type checking, it is extremely unlikely that the program will run if the number and order of items is incorrect. In some places, the types of items depend on the selection of an option.

The input is copied across to the output file, and if you strike problems with getting the data file accepted, examining the output will usually reveal the nature of the problem.

Following is an annotated data file for the CMIX program. Each line of a data file is printed in **BOLD** type. Where the function of the parameter is not obvious, explanatory notes about it are given.

Mixtures input file for C. gunnari at Heard Island 1993

> The first line of the input file is available for an annotation identifying the data file, and any other details the user would like for identification purposes. (Free format text).

[blank]

All lines marked as blank must be included in the data file as blank lines

```
NUMBER OF COMPONENTS IN MIXTURE        2
```

This is the number of components to be included in the mixture distribution (maximum = 10)

```
 [blank]
Bounds on the means of the components
Component  Low bound  High bound
    1        280.       310.
    2        315.       440.
```

The first two lines in this block are explanatory header lines. The following numeric lines specify the bounds within which the means for each mixture component can be fitted. Each line consists of three items, separated by spaces. The first item is the mixture component number. The next two items are the upper and lower bounds respectively. There must be one line for each component, in ascending order. The bounds for the components must not overlap. The bounds are not allowed to be negative or greater than 10000. The program checks that these conditions are complied with.

```
[blank]
Components with means to be held constant (-1) is end of list
Component  Fixed mean
   -1        0.
```

This block enables the user to specify if any of the component means are to be held at a fixed value. The first element on the line is the component number and the second element is the value of the mean. The component number has to fall in the range of 1 to the number specified in the number of components line at the start of the file. The fixed value has to fall within the bounds specified for that component. In this particular example, no means are to be held fixed because the only entry in the list is the end of list marker (-1 0.).

There is a bug in the program. If you have the same number of components in the fixed mean block as you have in the block above (i.e. the bounds around the mean) then you need to delete the blank line after the last line in the fixed mean block. If you have an uneven number of components in the two blocks then you do not need to delete this line.

```
[blank]
MIX Std. Devs LINEARLY RELATED          TRUE
```

This line specifies whether the standard deviations for the mixture components are to be restricted to being linearly related. Valid responses are “YES”, “NO”, “TRUE” or “FALSE” and depend only on the first letter and are case independent. Thus responses “yes”, “true”, “y” and “TRUE” for example all evaluate to TRUE. Similarly, “NO”, “f” and “FALSE” all evaluate to FALSE.

Usually we would expect that the standard deviations for the components of a length-density distribution will increase monotonically with the mean. If the response field is TRUE the standard deviations of the mixture components are a linear function of the mean, that is:

$$s_i = a + bm_i$$

where s_i and m_i are the standard deviation and mean of component *i* respectively and *a* and *b* are constants which may be estimated from the data. If the intercept (*a*) is held fixed at 0 then the mixture components will have a constant coefficient of variation estimated by the slope (*b*). If *b* is held fixed at 0 then the mixture components will have a constant standard deviation estimated by *a*. These possibilities are allowed for in setting bounds and parameters in the next block.

Using the linear option ensures that the standard deviations behave in an orderly fashion.

If the response is FALSE, the standard deviations can take any values within specified ranges given in the block after the next. Any or all of the standard deviations can be forced to specified values in the block after that.

Both types of block are kept in the data file so that the user can change from one to the other without having to have two different versions of the data file. However, only one type of the following blocks is in effect depending on whether the response was TRUE or FALSE.

```
[blank]
The following 8 parameters are only used if LINEARLY RELATED is TRUE
LOWER BOUND ON LINEAR INTERCEPT        1.
UPPER BOUND ON LINEAR INTERCEPT        50.
LOWER BOUND ON LINEAR SLOPE            0.0
UPPER BOUND ON LINEAR SLOPE            0.4
STARTING VALUE FOR INTERCEPT SEARCH     15.
STEP LENGTH FOR INTERCEPT              1.
STARTING VALUE FOR SLOPE SEARCH        0.07
STEP LENGTH FOR SLOPE                  0.01
```

This block controls the estimation of the parameters for the linear relationship describing the standard deviations of the mixture distribution. Obviously they only have any effect when the response to the linear relationship was TRUE.

The lower bound on the linear slope should be > 0 if the constraint is to be enforced that the mixture standard deviations increase with the mean.

When the starting value for the intercept and its step length are both set to zero, the estimated slope is a constant coefficient of variation for the mixture components. When the starting value and step length for the slope search are both set to zero, the estimated intercept is a constant standard deviation for the mixture components. The slope and intercept cannot both be simultaneously fixed at zero, although either or both can be fixed at non-zero values.

```
[blank]
Bounds on the standard deviations of the components (Only if LINEAR is false)
Component  Low bound  High bound
    1         5.        50.
    2         5.        50.
```

This block specifies the bounds to be respected in searching for the estimates of the standard deviations of the mixture components when linear is FALSE. Unless the bounds are selected such as to impose restrictions on the values of the standard deviation estimates, it is possible that the standard deviation for a high mean will be numerically smaller than for a low mean.

```
 [blank]
Components with standard deviations to be held constant (-1) is end of list
Component  Fixed std dev.  (Only if LINEAR is false)
   -1
```

As was the case for the means, specified mixture components can have their standard deviations fixed.

There is a bug in the program so that the component number and the SD to be held constant must be put on separate lines. Also, if you have the same number of components in the fixed mean block as you have in the block above (i.e. the bounds around the mean) then you need to delete the blank line after

the last line in the fixed mean block. If you have an uneven number of components in the two blocks then you do not need to delete this line.

```
 [blank]
MINIMISATION                    YES
```

This parameter specifies whether estimates are to be made by minimising the residual function over the parameter space. If FALSE, the residual function is evaluated once, at the starting values of the parameters.

```
PLOT FITTED FUNCTION AND DATA          YES
```

The program will plot the fit of the mixture distribution to the data. It is essential to view the goodness of fit, particularly when there is a possibility of fitting a component in a region where there are no or few density values for a range of length classes. Length classes for which all observations have zero density make no contribution to the likelihood function. Therefore, a mixture component in such regions may be completely spurious, and hence bias the recruitment proportion. In such cases constraints on the range of length classes to be included in the fit, as well as on the mixture components are required to attempt to produce a sensible fit. The goodness of fit plot is the principle method for using the program for these interactive adjustments.

PLOT RESIDUAL FUNCTION OVER P1 NO

If required, the program will plot out the value of the residual function against the various values of Ri. This plot can be used to determine approximate 95% confidence intervals for the estimate of Ri. The asymptotic 95% confidence interval is defined by the line at the residual function value which is 1.92 above the residual at the minimum (see de la Mare, 1994).

PLOT ON SCREEN	**YES**
PLOT ON PLOTTER	**NO**
PLOT ON PRINTER	**NO**
PLOT TO FILE	**YES**

Output devices for graphs (If plots have been enabled). If plots are saved to files, they are automatically named PLOT??.DAT, with the ?? representing a two digit number which increases from 05 as the number of plots generated by the program during a fit increases. **If these files are to be retained they must be renamed to prevent them being overwritten by the next run of the program**. They can be viewed with the VTRANS program (see below).

MAXIMUM NUMBER OF FUNCTION CALLS 10000

This parameter specifies the maximum number of function calls allowed in searching for the best fit to the data. The value given here should be adequate in most cases.

MINIM REPORTING FREQUENCY 100

The minimisation routine will report the parameters and residual function value at the regular intervals specified here. This output is useful for checking that the procedure has converged reliably, and on a minimum in the range covered in a given run. A specified value less than zero inhibits reporting.

STOPPING CRITERIA 1.E-6

This is a technical parameter of the minimisation routine, and controls how little the values of the residual function should vary with changes in the estimated parameters before the minimisation will end. In other words, the minimisation will end when the change in the residual function is smaller than or equal to the magnitude of the stopping criterion. Failure of the procedure to converge even though the function reports indicate that convergence has occurred (i.e. the residual function values in the final function calls are more or less the same) is symptomatic that the stopping criterion is too small. Conversely, if there are few function calls before convergence is reported combined with obvious variation remaining between values of the residual function, indicates the stopping criterion is too large.

FREQUENCY FOR CONVERGENCE TESTING 5

This is another minimisation technical parameter, 5 is usually satisfactory.

FIT QUADRATIC SURFACE NO

The minimisation routine can fit a quadratic function to the region of the minimum, and hence give normal theory approximations to the information matrix pertaining to the estimated parameters. This is not necessarily reliable, and its use is not recommended for this program.

SIMPLEX EXPANSION COEFFICIENT 1.

This is a minimisation routine technical parameter, which only applies if quadratic surface fitting is enabled. Adjusting it can improve the correspondence between the minima of the fitted quadratic and the minimum found by the search.

[blank]
NUMBER OF LEADING INTERVALS TO SKIP 6

This parameter can be used to specify whether the length frequency data are to be truncated to the left. In this example the specification means that the first 6 intervals are not used in fitting. This means that the intervals less than 260mm will not be used in fitting the mixture distribution. Specifying zero means that all the intervals will be used.

Bin interval boundaries (mm)
40. 50. 60. 70. 80. 90. 260. 270. 280. 290.

```
300.  310.  320.  330.  340.  350.  360.  370.  380.  390.
400.  410.  420.  440.  450.
```

This block specifies the boundaries between the length intervals in the length-density distribution. The boundaries do not need to be all the same width, but a length interval is defined between each pair of values. In the example above, there is one very wide interval between 90 and 260mm. This feature can be used to save entering a large number of zeros in the haul by haul data

in cases where there are large gaps between mixture components. Obviously, there will be one more datum entered in this block of interval boundaries than there are intervals in the data blocks below. There is no requirement to use measurements in mm.

```
[blank]
Gunnari Ridge haul no 19                   1
0     0     0     0     0     0     0     0     0     71.90
0     0     143.80  287.61  503.32  934.73  790.92  790.92  143.80  71.90
0     0     0     0
```

This block and those that follow are the basic length-density data from each single haul in the survey. The first line in the block can be used for a haul identifier. After the 40th character is the data block number. **This starts at 1** and increases strictly sequentially for each haul. This is used to help ensure that any formatting errors in the data blocks will be detected. The data can be placed over as many lines as required. The first blank line terminates the individual haul block.

```
[blank]
Gunnari Ridge haul no 55                   2
0     0     0     0     0     0     0     0     0     0
0     0     0     0     13.75   13.75   0     13.75   13.75   0
0     0     0     0
[blank]
Gunnari Ridge haul no 56                   3
0     0     0     0     0     0     0     0     0     0
0     20.44   0     40.88   40.88   40.88   81.76   81.76   122.64  81.76
40.88   0     0     0
[blank]
Gunnari Ridge haul no 57                   4
0     0     0     0     0     0     0     0     0     0
0     0     0     303.88  694.58  1041.86 1953.49 2648.07 2431.01 1693.03
651.16  86.82   0     0
[blank]
Gunnari Ridge haul no 58                   5
0     0     0     0     0     0     0     0     0     0
29.82   89.45   0     0     0     0     59.64   29.82   29.82   59.64
0     0     0     0
[blank]
Gunnari Ridge haul no 114                   6
0     0     0     0     0     0     0     0     167.33  167.33
334.66  669.33  2342.64 5019.95 5521.94 8199.25 4685.28 4350.62 1171.32 836.66
0     0     0     0
```

[blank]
End of haul data -1

This line must be present to indicate the end of the haul by haul data.

After this line, the user may type any additional information they wish to record about the data file, or anything else for that matter. These lines are not read by the program, e.g.:

[blank]
Data are numbers of fish per square kilometre

APPENDIX 3 CMIX Output File Format

The following annotated output file is not presented in complete form. The following line will be inserted at the points where lines of the file are skipped because the loss of information will not reduce the interpretation of the annotations:

……[skipped]……

```
Mixtures input file for C. gunnari at Heard Island 1993

NUMBER OF COMPONENTS IN MIXTURE                 2

Bounds on the means of the components
Component  Low bound  High bound
  1   280.000      310.000
  2   315.000      440.000

……[skipped]……
End of haul data                          -1
```

The output file reports the input data so that the user can check that the inputs have been interpreted correctly. This also provides a means of checking where the program might stop as a result of incorrect input data.

```
Table of densities for each length class


Length      Density       S.E.     Lower C.I.     Upper C.I.



265.000    0.000000    0.000000    0.000000    0.100000E+36

275.000    0.000000    0.000000    0.000000    0.100000E+36

285.000     27.8883     27.8883   0.000000    0.100000E+36

295.000     39.8717     28.0657    7.18659     223.849

305.000     60.7467     54.9987    6.52327     13772.4

315.000     123.821     100.170    18.4901     10894.5

325.000     414.407     386.361    37.0117     222581.

335.000     798.009     629.461    127.341     63714.4

345.000     1244.24     1023.69    168.259     158029.

355.000     2166.27     1852.64    242.037     466673.

365.000     1467.45     1077.89    292.180     62100.2

375.000     1773.21     1422.86    267.309     155819.

385.000     678.933     488.063    145.398     24076.7
```

```
395.000     431.856     278.559     113.983     7893.69

405.000     115.340     107.372     10.4226     58818.9

415.000     14.4700     14.4700     0.000000    0.100000E+36

430.000     0.000000    0.000000    0.000000    0.100000E+36

445.000     0.000000    0.000000    0.000000    0.100000E+36
```

The densities of each length class is estimated (with standard error and lower and upper confidence intervals) and is the first table to be provided.

```
  First minimisation

   PROGRESS REPORT EVERY 100 FUNCTION EVALUATIONS


EVAL. NO.  FUNC. VALUE         PARAMETERS
1     81.7253     287.598     360.445     19.1433     0.347428E-12  178.321     9156.41

2     81.7270     290.598     360.445     19.1433     0.347428E-12  178.321     9156.41
             ……[skipped]……

INITIAL EVIDENCE OF CONVERGENCE
CENTROID OF LAST SIMPLEX    292.078     360.412     14.0210     0.143387E-01   157.671     9090.62

FUNCTION VALUE AT CENTROID     81.7237
*
INITIAL EVIDENCE OF CONVERGENCE
CENTROID OF LAST SIMPLEX     292.074     360.412     14.0215    0.143405E-01   157.671     9091.31

FUNCTION VALUE AT CENTROID     81.7237
             ……[skipped]……



   MINIMUM AT     294.001     360.379     1.00941     0.504853E-01   135.183     9102.06



   MINIMUM FUNCTION VALUE     81.7234



   END  OF  SEARCH
   ***************
```

Summary results during the minimisation procedures are reported at the rate specified in the input file. These can be reviewed.

```
Standard Error in estimate  5   200.170
             ……[skipped]……
```

```
Standard Error in estimate  6   5799.84

Compare minimum found by quadratic fit with that found by minimisation.
If the the difference is large, the standard error estimate is not reliable

Minimum from minimisation routine
R1        =  9102.06
Func. value =   81.7234

Minimum from quadratic fit routine
R1        =  9115.06
Func. value =   81.7234

Means of mixture components
294.001      360.379


Parameters of linear standard deviations
Intercept =   1.00941
Slope     =  0.504853E-01


Standard deviations of mixture components
15.8521      19.2033


Total density of each mixture component
135.183      9102.06


SD of each mixture component density
200.170      5799.84
```

Summary results of the mean length and standard deviation for the component. If the standard deviations have been found through a linear relationship between the SD and the mean then the parameters for that relationship are given. Otherwise, only the mean and standard deviation of the lengths in each component are given. In all executions of the program, the total density and standard deviation of the density for each component will be given.

```
Table of observed and expected mean densities

Interval     Observed      Expected      Lower C.I.    Upper C.I.

265.000     0.000000      6.63820     0.000000      0.100000E+36
275.000     0.000000      16.8203     0.000000      0.100000E+36
285.000      27.8883     29.6301     0.000000     0.100000E+36
295.000      39.8717     39.8543      7.18659      223.849
305.000      60.7467     58.5455      6.52327      13772.4
315.000      123.821     136.221      18.4901      10894.5
325.000      414.407     360.987      37.0117      222581.
335.000      798.009     797.416      127.341      63714.4
345.000      1244.24     1366.78      168.259      158029.
355.000      2166.27     1799.44      242.037      466673.
365.000      1467.45     1817.60      292.180      62100.2
```

```
375.000    1773.21    1408.41    267.309    155819.
385.000    678.933    837.154    145.398    24076.7
395.000    431.856    381.673    113.983    7893.69
405.000    115.340    133.453    10.4226    58818.9
415.000    14.4700    35.7814    0.000000   0.100000E+36
430.000   0.000000    4.25730    0.000000   0.500000E+35
445.000   0.000000   0.139999    0.000000   0.100000E+36

Sum of the observed densities =   9356.51
Sum of the expected densities =   9235.07
If the sums differ by a large amount, the fit may be unreliable
```

APPENDIX 4 Trouble Shooting

If the program ends before providing a full output then the following steps might help:

Check the Output file to determine at what stage in the process the program stopped.

If the problem is with the input file then check:

i) the formatting of the input file (see Appendix 2)

ii) that the 'number of components in mixture' specified in line 2 of the file matches the number of components in the lists below.

iii) that the number of components for the means equals the number of components for the standard deviation. NB: even if the standard deviations are linearly related to the means (i.e. TRUE) and the block listing the standard deviations components is not used, the number of components for the mean and standard deviations must be equal. This is because the two methods for estimating the standard deviations of the components was retained in the one input file to minimise the changes to the input file necessary to explore different options (see annotations to the input file).

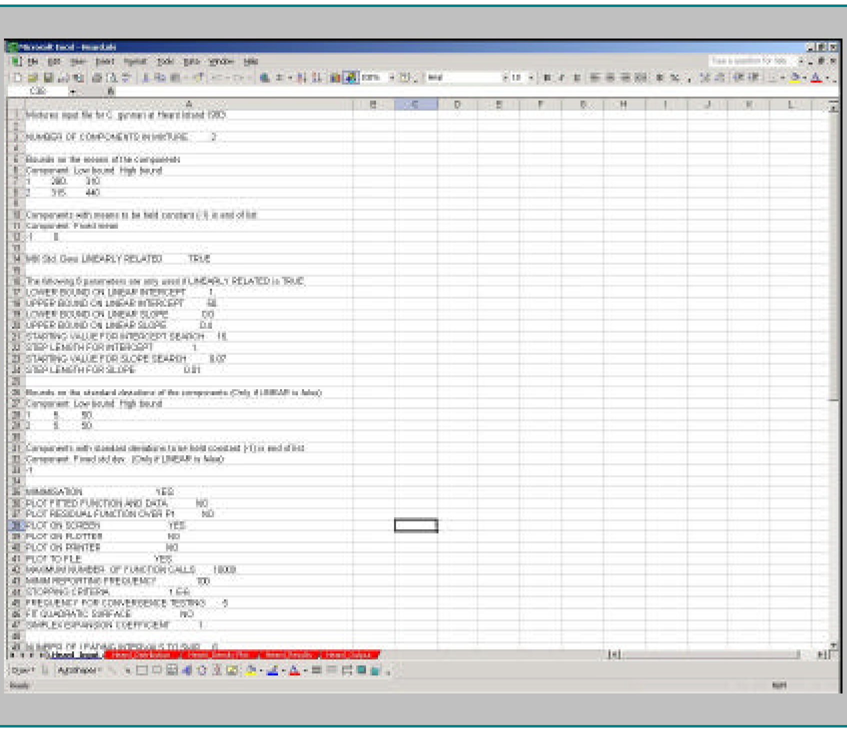
iv) that there is one extra bin-boundary than there are data points.

v) that there is a blank line between the last block of data and the 'End of data -1' line.

vi) that the number of leading zeros to skip is correct.

vii) that the number of each data block is given consecutively beginning at ‘**1**’

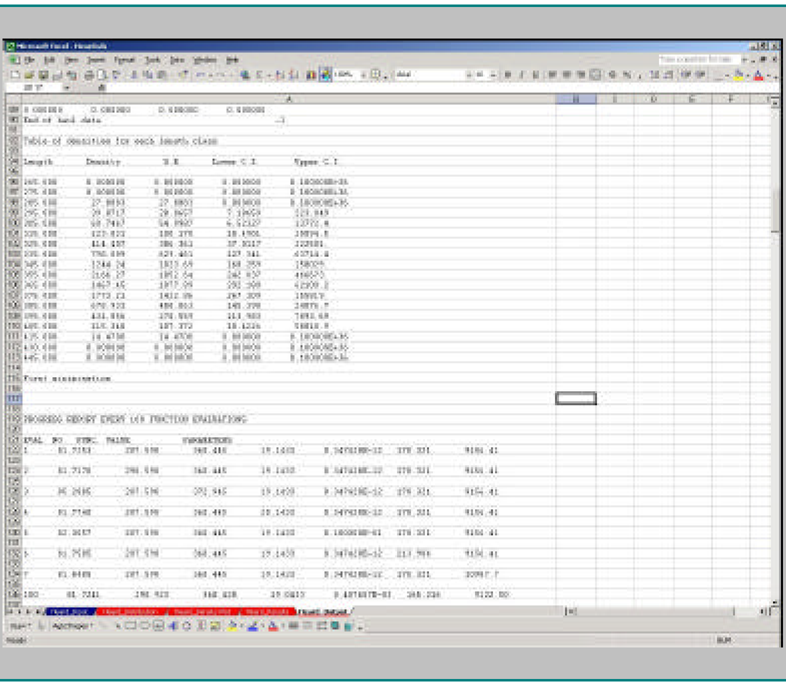
APPENDIX 5 CMIX Excel Add-In Worksheets



```
1  Mixtures input file for C. gunnari at Heard Island 1990
2
3  NUMBER OF COMPONENTS IN MIXTURE      2
4
5  Bounds on the means of the components
6  Component  Low bound  High bound
7  1       280     310
8  2       315     440
9
10 Components with means to be held constant (-1 is end of list
11 Component  Fixed mean
12 -1      0
13
14 MIX Std. Devs LINEARLY RELATED       TRUE
15
16 The following 6 parameters are only used if LINEARLY RELATED is TRUE
17 LOWER BOUND ON LINEAR INTERCEPT      1
18 UPPER BOUND ON LINEAR INTERCEPT      50
19 LOWER BOUND ON LINEAR SLOPE         0.0
20 UPPER BOUND ON LINEAR SLOPE         0.4
21 STARTING VALUE FOR INTERCEPT SEARCH   15
22 STEP LENGTH FOR INTERCEPT          1
23 STARTING VALUE FOR SLOPE SEARCH      0.07
24 STEP LENGTH FOR SLOPE            0.01
25
26 Bounds on the standard deviations of the components (Only if LINEAR is false)
27 Component  Low bound  High bound
28 1       5      50
29 2       5      50
30
31 Components with standard deviations to be held constant (-1) is end of list
32 Component  Fixed std dev.  (Only if LINEAR is false)
33 -1
34
35 MINIMISATION                YES
36 PLOT FITTED FUNCTION AND DATA       NO
37 PLOT RESIDUAL FUNCTION OVER P1       NO
38 PLOT ON SCREEN               YES
39 PLOT ON PLOTTER              NO
40 PLOT ON PRINTER              NO
41 PLOT TO FILE                YES
42 MAXIMUM NUMBER OF FUNCTION CALLS     10000
43 MINIM REPORTING FREQUENCY         100
44 STOPPING CRITERIA             1.E-6
45 FREQUENCY FOR CONVERGENCE TESTING    5
46 FIT QUADRATIC SURFACE           NO
47 SIMPLEX EXPANSION COEFFICIENT      1
48
49 NUMBER OF LEADING INTERVALS TO SKIP   0
```

Microsoft Excel - Heardtab
File Edit View Insert Format Tools Data Window Help
Heard_Input
Heard_Distribution
Heard_Density Plot
Heard_Results
Heard_Output
Ready

Figure 16 CMIX Input File Worksheet



```
End of hard data

Table of densities for each length class

Length      Density      S.E.      Lower C.I.      Upper C.I.
[illegible]

Final minimisation

PROGRESS REPORT EVERY 100 FUNCTION EVALUATIONS

EVAL. NO.   FUNC. VALUE        PARAMETERS
[illegible]
```

Microsoft Excel - Heardtab
File Edit View Insert Format Tools Data Window Help
Heard_Input
Heard_Distribution
Heard_Density Plot
Heard_Results
Heard_Output
Ready

Figure 17 CMIX Output File Worksheet

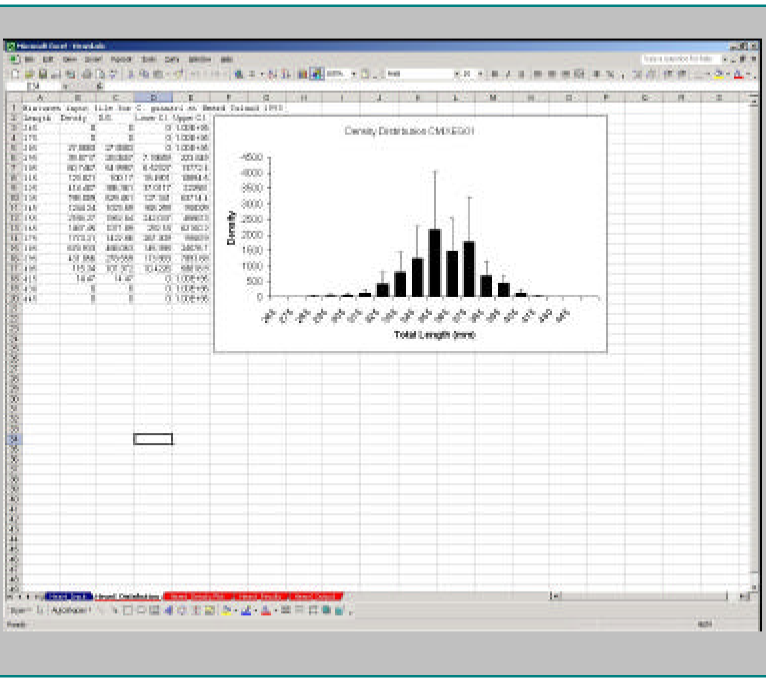


Figure 18 CMIX Distribution Worksheet

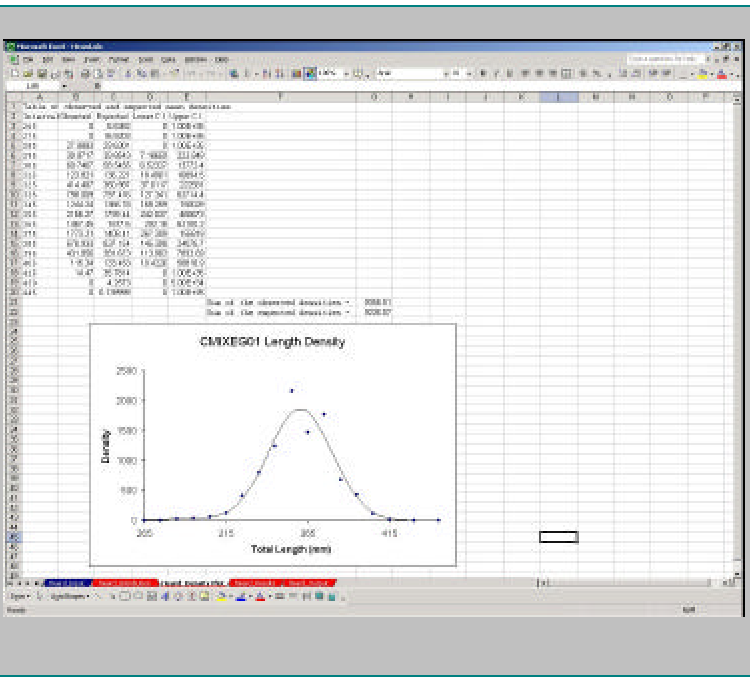
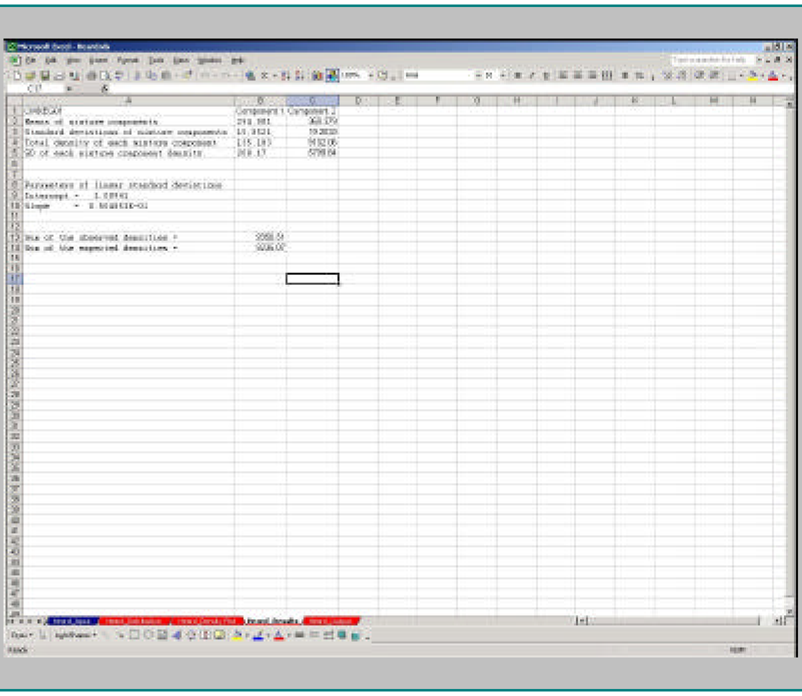


Figure 19 CMIX Density Plot Worksheet



	A	B	C
1	[illegible]	Component 1	Component 2
2	Means of mixture components	[illegible]	[illegible]
3	Standard deviations of mixture components	[illegible]	[illegible]
4	Total density of each mixture component	[illegible]	[illegible]
5	SD of each mixture component density	[illegible]	[illegible]
6			
7			
8	Parameters of linear standard deviations		
9	Intercept = [illegible]		
10	Slope = [illegible]		
11			
12			
13	Sum of the observed densities =	[illegible]	
14	Sum of the expected densities =	[illegible]	

Figure 20 CMIX Results Worksheet

- Your Notes -

References

Aitchison, J. (1955) On the distribution of a positive random variable having a discrete probability mass at the origin. *J. Am. Stat. Assoc.,* **50:901-908**

de la Mare, W. K. (1994a) Estimating krill recruitment and its variability. *CCAMLR Science* **1:55-69**.

de la Mare, W. K. (1994b) Modelling krill recruitment. *CCAMLR Science* **1:**

MacDonald, P. D. M. and Pitcher, T. J. (1979) Age-groups from size-frequency data: a versatile and efficient method for analysing distribution mixtures. *J. Fish. Res. Board Can.* 36:987-1001.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics.* **Chapman and Hall, London.**

Pennington, M. (1983) Efficient estimators of abundance for fish and plankton surveys. *Biometircs* **39:281-286**