Rerum Naturalium Fragmenta

No. 404

PERECT for PC-DOS
Version 2.5
User's Guide
by
T. Jasko

Chapter 3. FACTOR MODELS

Watford
1991

.

Rerum Naturalium Fragmenta
Tamas Jasko editor
16 Melrose Place, Watford WD1 3LN, England

Chapter 3
FACTOR MODELS

The automatic validation and distribution fitting that follows the editing of a data file shows little detail of the calculations and except for errors it does not pause. If you want to see the details then select Check Model from the Model menu. Factors of a validated model can be shown in graphic form by the Factor option.

Constant factors

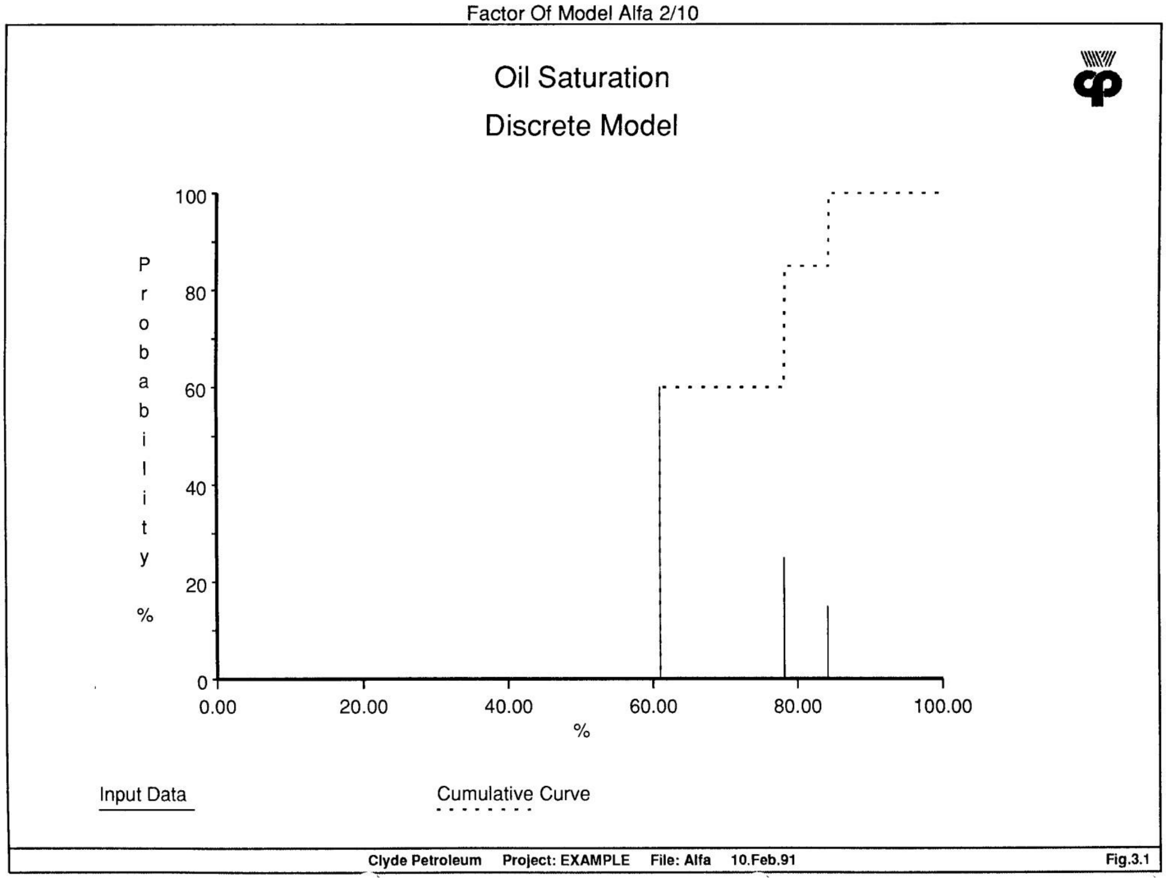
If the exact value of the factor is known it can be entered as a single value (with 100% probability). The factor is used then as a constant multiplier by the program. On data sheets it will be displayed as 'Constant:'. For a constant it does not matter what you specify as 'distribution model'.

More usually, the exact value of the factor is not known. The value is known only as a random variable within a distribution range.

Factor distribution models define the way such random values can be obtained for simulation.

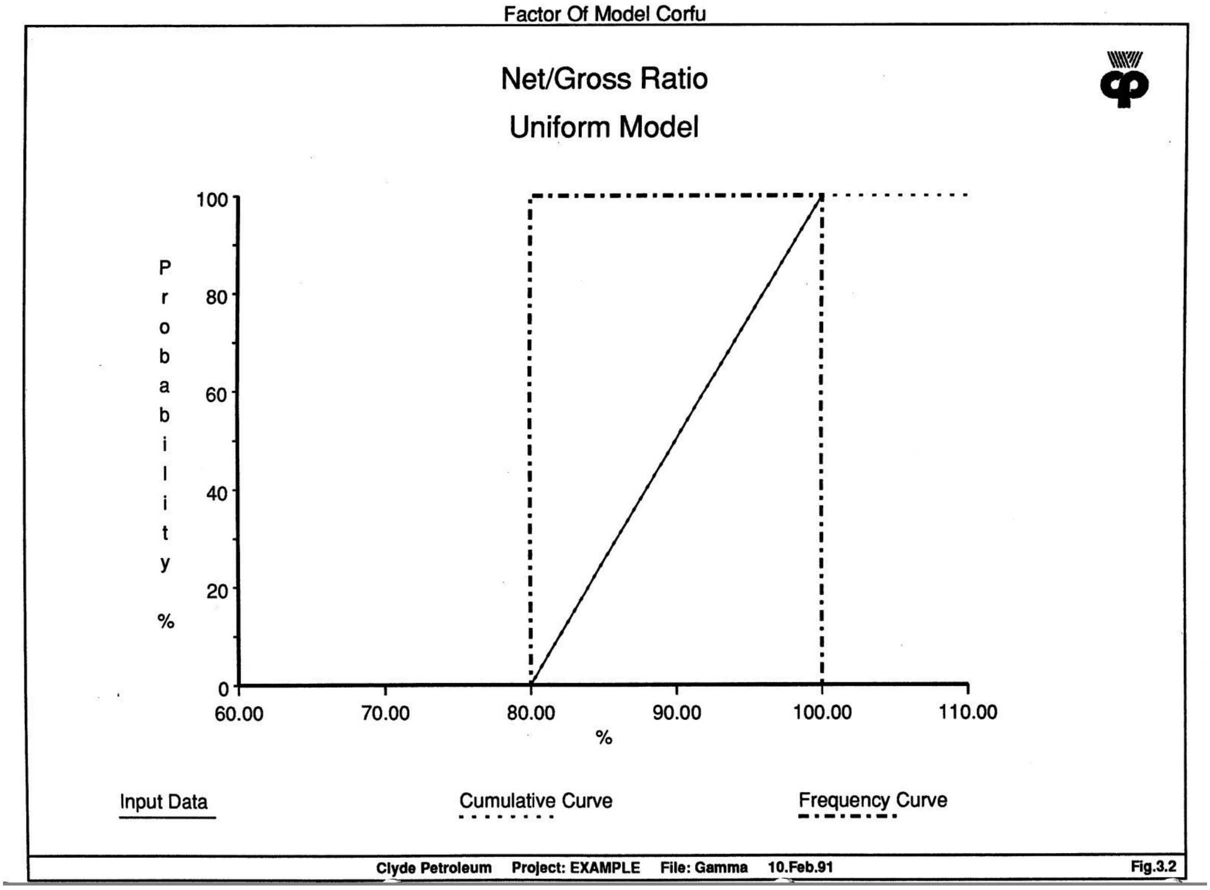
Probability distributions

When you are entering a model, for each of the geological parameters of a reservoir (the factors of a model) you can select a probability model by entering its code:



D - Discrete
U - Uniform
T - Triangular
C - Cumulative
F - Frequency
N - Normal
L - Log-normal
A - Best Fit
B - Beta
X - Explicit Normal
Y - Explicit Log-normal
Z - Explicit Beta

Enter one of these letters, in upper or lower case, in the box. Each code defines a probability distribution of the values that the factor can take. The nature of the factor and the available data are determining the best choice of model e.g. gross rock volume is generally considered to have log-normal distribution.

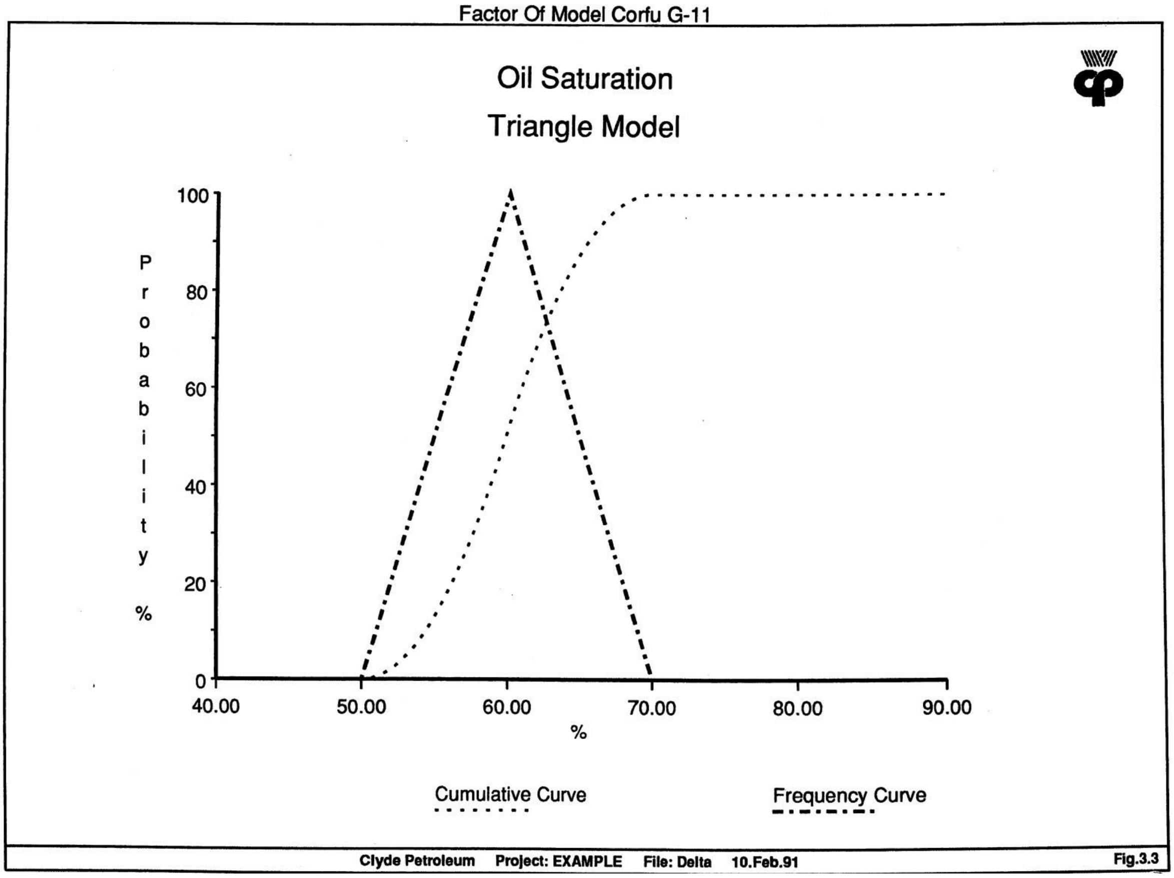


Discrete distribution

Factors that may take a number of discrete values, rather than having a continuous distribution, can be modelled using model D. The model is defined by the actual discrete values and their probabilities. Factor validation plots for discrete distributions display the input data as vertical bars (the height of the bar is the probability of the value) and the corresponding stepped cumulative graph.

Cumulative models

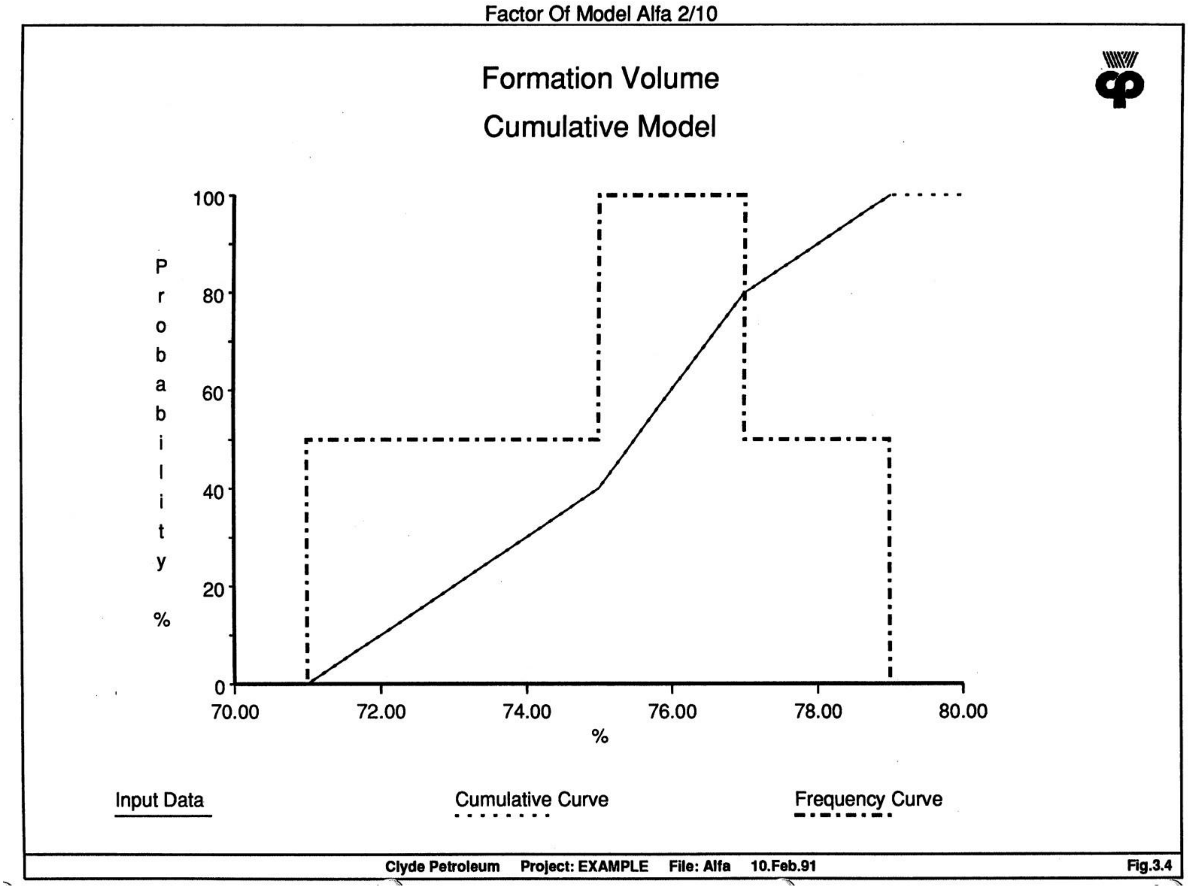
Continuous distributions can be specified in these models by entering values read off hand drawn curves rather than referring to the normal/log-normal fits. This way the model fitting stage is bypassed.



This is particularly useful if the distribution is known to have an unusual shape that cannot be well approximated by normal or log-normal distribution. E.g. if the distribution is bimodal, having more than one peak.

Cumulative distributions (model C) are entered as a series of values with their cumulative probabilities. The first value should be the minimum value and its probability is 0; the last value should be the maximum value with 100 % probability.

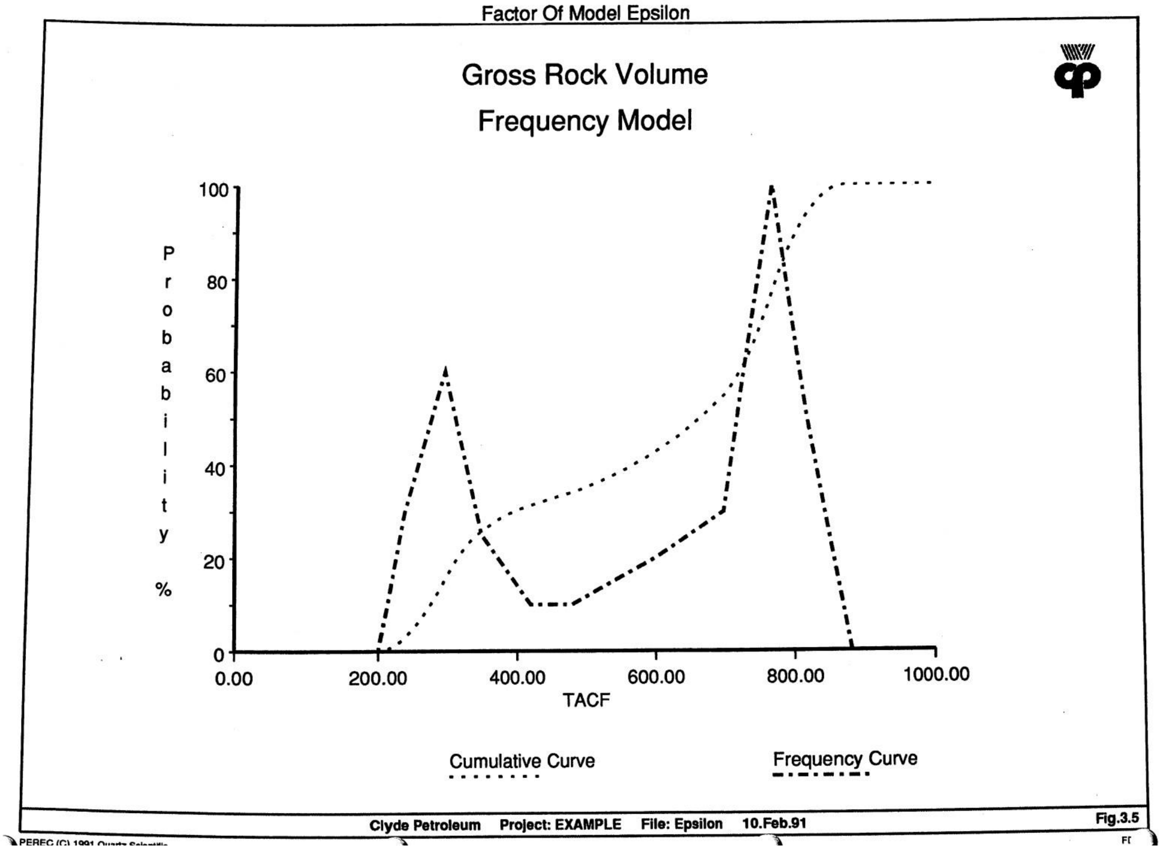
Up to 10 intermediate values can be entered to show the course of the curve in between the extremes, these should be ordered by increasing value and probability.



Frequency distributions

Frequency distributions are similarly given by pairs of values; with the first and last ones indicating minimum and maximum, respectively. Both extremes should have 0 probability, the probability of the intermediate values, ordered by value, is given on a relative 0 to 100 % scale. The example plot shows a bimodal distribution entered as a frequency distribution (model F).

Definition by frequency distribution is usually to be preferred to cumulative models as the curves are more realistic (less jerky) approximations to actual continuous distributions. Factor validation plots for either model will show both cumulative and frequency curves. One of these is obviously the input itself - the other is computed by the program. Also, the average (mean) of the distribution is computed and displayed.

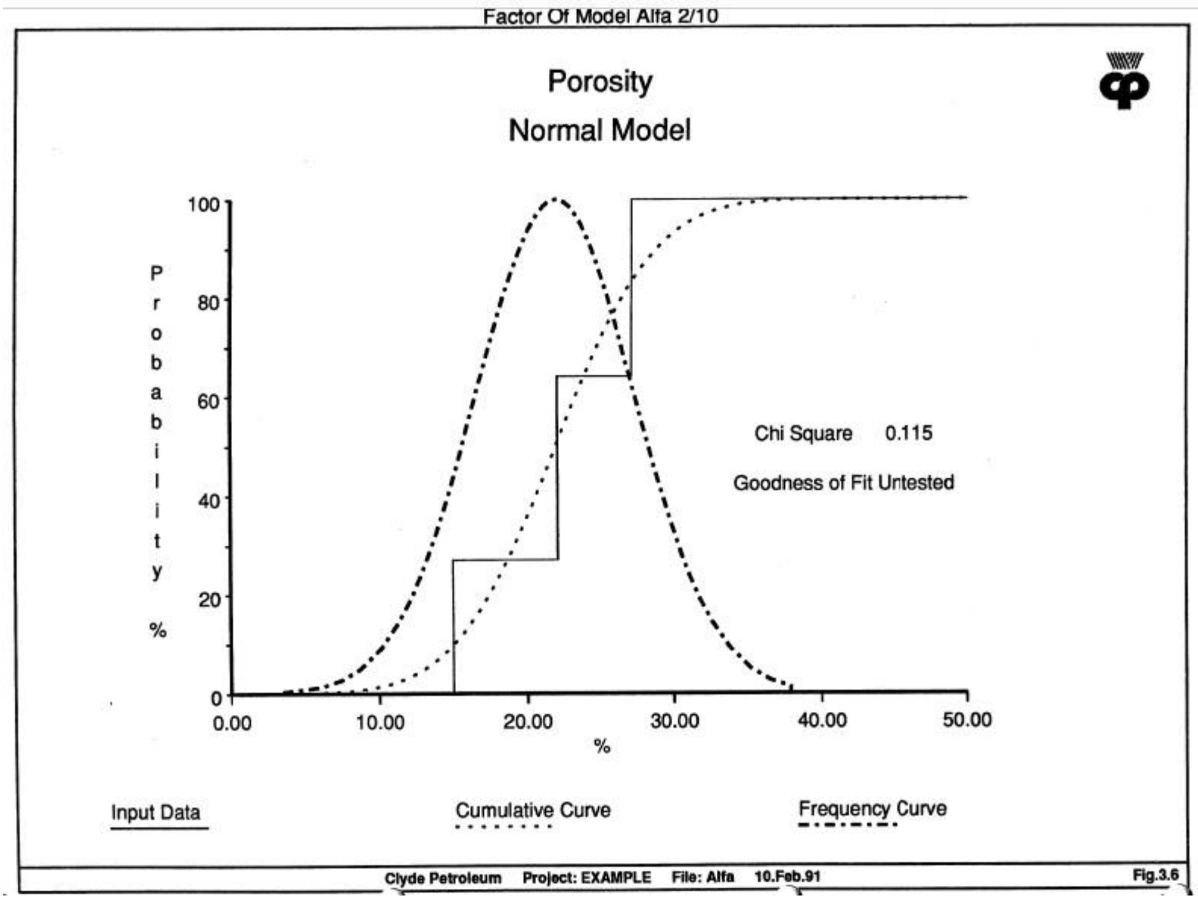


Triangular distribution

Often the only data available about a geological factor are the minimum, maximum and most likely value. For such cases the frequency distribution defined by the three values is known as the triangular distribution (See example). The minimum,most likely and maximum values should be entered in the model with 0, 100, 0 % probability, respectively.

Uniform distribution

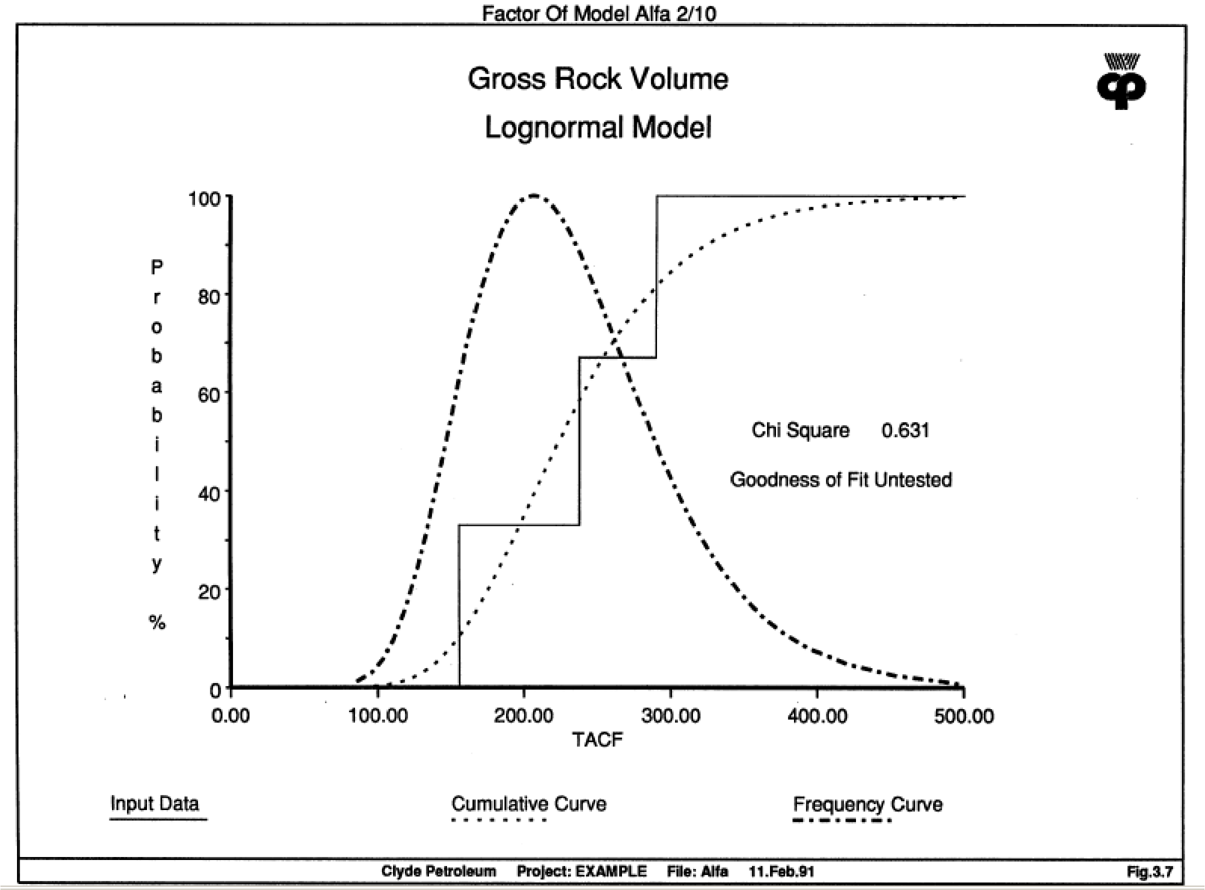
If only the minimum and maximum values are known these can be entered in the form of a uniform distribution. Enter the minimum value with 0 %, the maximum value with 100 % probability.



Normal & log-normal fits

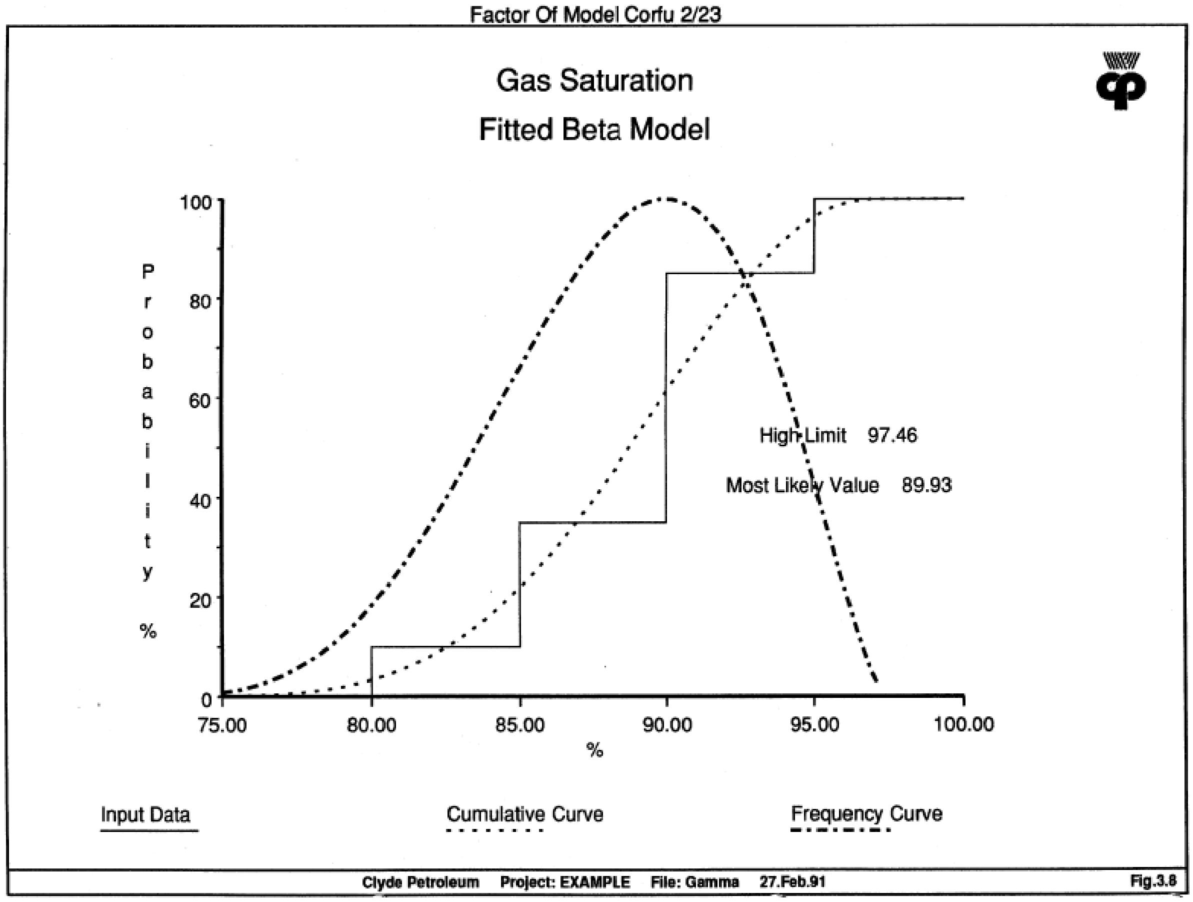
Models A, L, and N are various versions of the same procedure to fit a normal or log-normal distribution to a set of representative values. If any of these model options are used the program will fit the normal and log-normal distributions that approximate best the values (and their probability weightings) supplied by the user.

If model A is used, the program computes both fits and decides between normal and log-normal distribution automatically, choosing the one with smaller residual error. This is overruled by model choices 0, L and N.



Models N and L always result in using normal, or log-normal distributions, respectively. By using model 0 (or leaving the entry blank) the choice between normal and log-normal is left to the user to decide when the distribution fits are done. After the residuals are displayed the user can enter N for normal, L for log-normal or press Return for the (default) automatic choice.

Based on the principle of the equal area rule normal distribution is fitted by the least squares method. Log-normal fit is obtained by iterative optimisation by the method of steepest descent. The residual error is calculated as the average difference between the stepped cumulative curve given by the input data and the corresponding values of the fitted distributions. In the plots produced by the factor validation program both the input and the fitted curves are displayed.



The quality of the fit can be judged from the calculated absolute differences, and chi square measure of fitness. The chi square test value is compared to tables of 95 % confidence level critical values for a given degree of freedom.

The results are expressed as 'adequate' or 'inadequate' goodness of fit. If the number of input values is less than four, the test will not give meaningful results and the program reports 'not tested'.

Beta fit

If you specify model B then the program will try to fit a beta distribution to the data.

The procedure followed first determines the Pearson distribution class of the data and only proceeds to fit a Beta distribution if the Pearson class is appropriate.

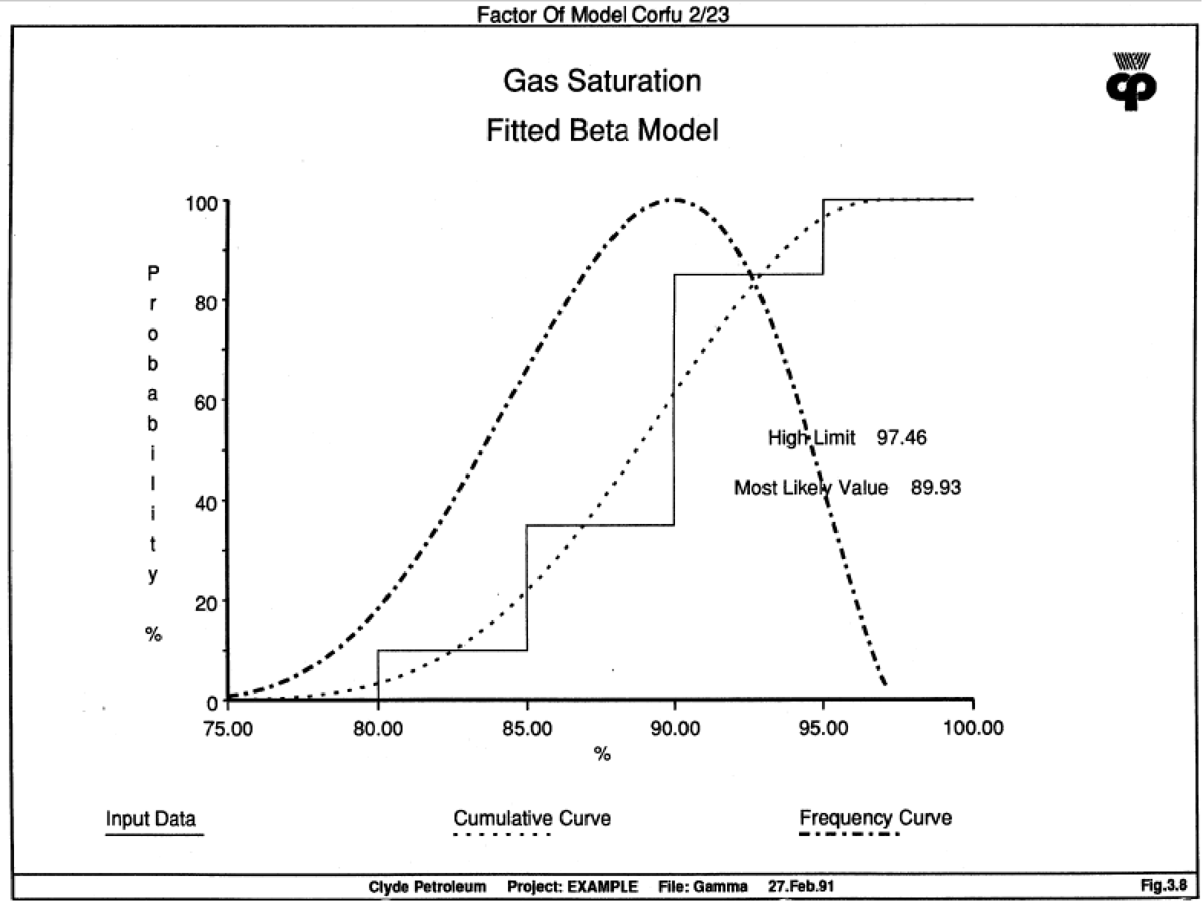
Explicit models

Factors that have Normal, Log-normal, or Beta distribution can be modelled directly by entering the distribution parameters, e.g. the mean (Mu) and variance (Sigma). Use model X for normal, Y for log-normal, Z for beta. For explicit normal models enter Mean and Variance in the units given.

For explicit log-normal, enter the Mean and Variance of the (natural) logarithm of the factor values corresponding to the Mu and Sigma parameters shown on summary data sheets. For beta, enter the values of the parameters Gamma, Delta, U, V. No 'probability' values are needed.

Explicit models can be used to test the effect of varying the parameters of distributions e.g. changing the mean and/or variance obtained by normal or log-normal fits.

Plots of explicitly defined model distributions show no 'input' curve, the factor validation plots display computed cumulative and frequency curves only (See examples).

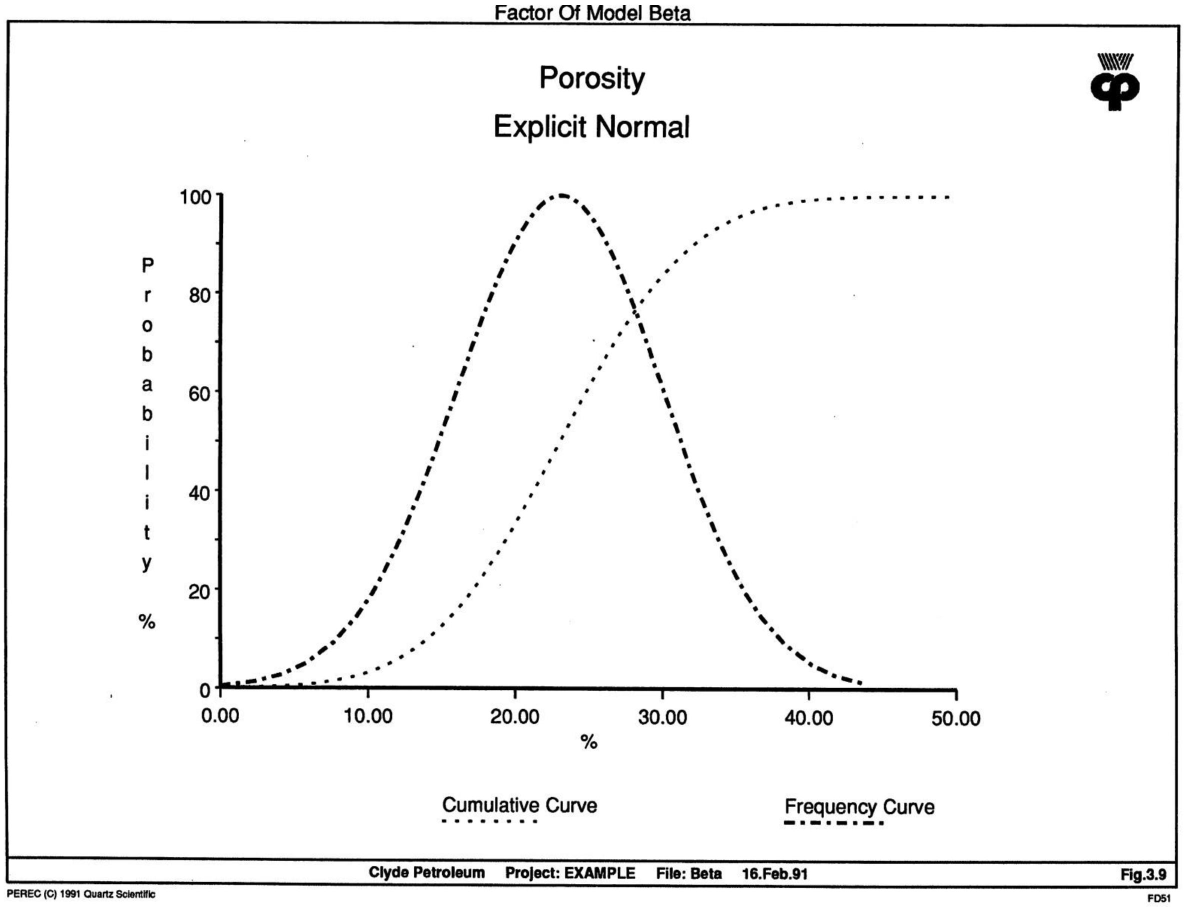


Extreme values

The random values produced for the samples are always within the intended range of acceptable values if the factor distribution is given by discrete, cumulative or frequency models.

On the other hand, normal and log-normal models, whether fitted or explicitly specified by the user, may pose a problem. The tails of the distribution may yield unacceptable values in certain cases. The factor validation program will check the probability of such excesses and, if necessary, warn the user.

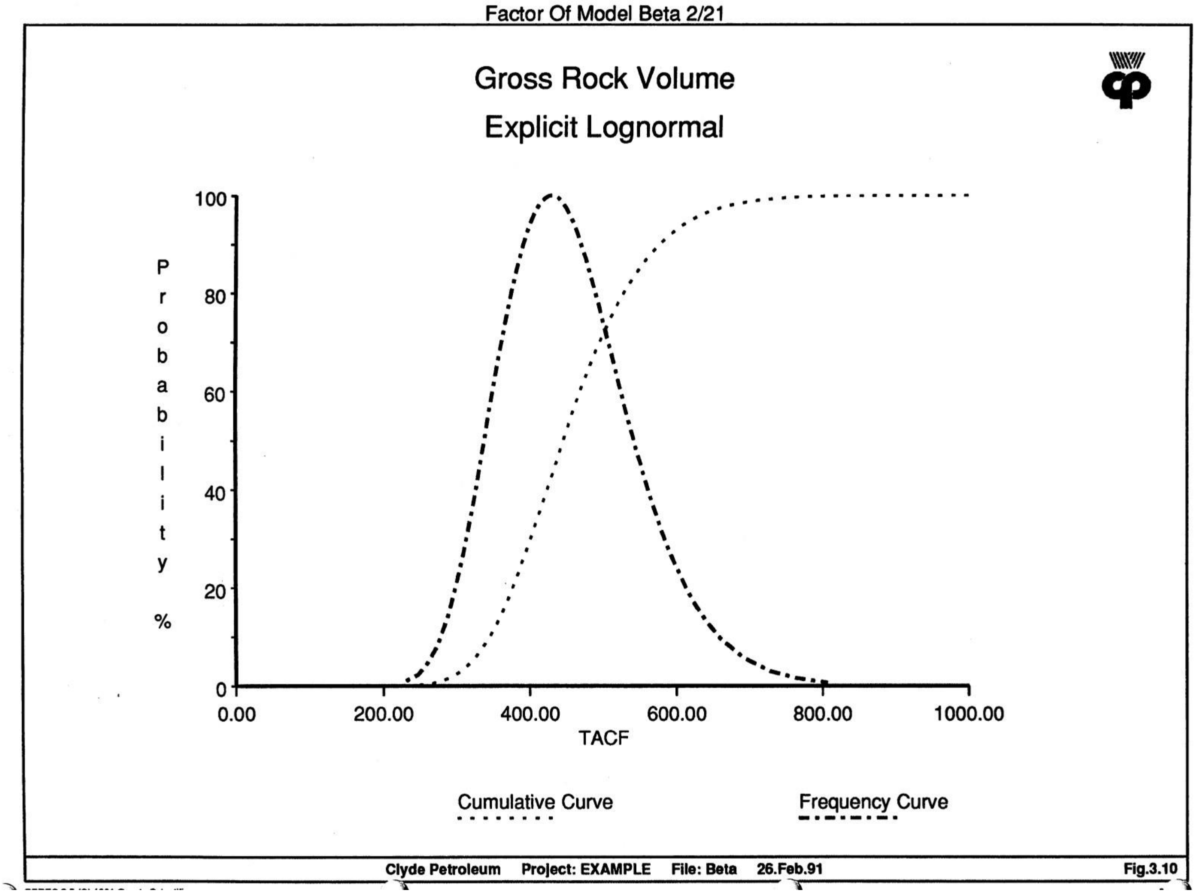
For each factor of normal or log-normal distribution the range of values produced by the model is checked against the limits of minimum and maximum acceptable values.



The limits may be specified by the user in the data file or assumed by the program. The program computes the probability of the factor values falling outside the limits and displays both probabilities i.e. values under minimum, over maximum.

Levels of severity

A WARNING is issued if the probability of falling outside of an assumed limit is greater than 1 %. An ERROR flag is raised if the probability of overflow or underflow exceeds 5 % (or 1% if the limit was explicitly set by the user.)



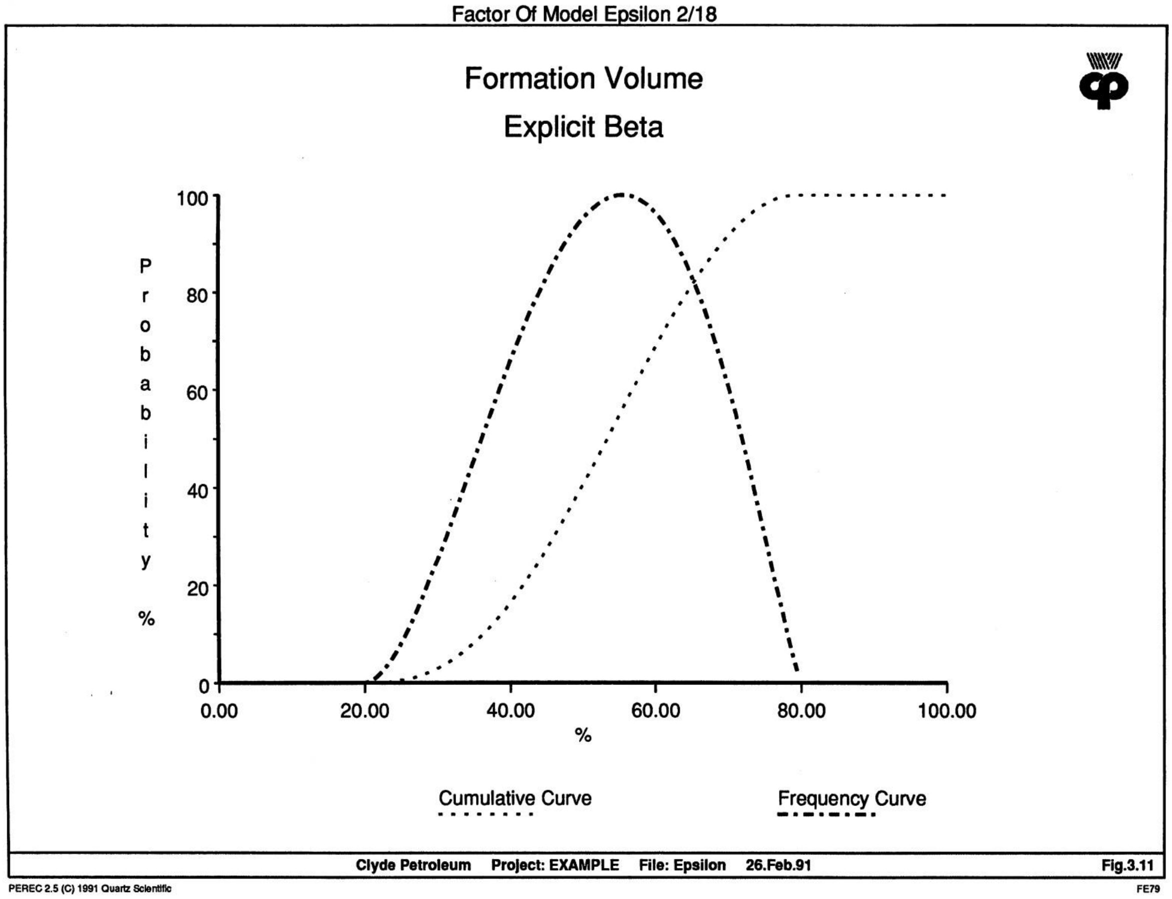
How to specify range limits

For fitted models of types A, L, and N the user can enter low and/or high limits by entering values with zero probability. Either or both can be entered.

If the factor model is explicit normal/log-normal (types X and Y) the data entry program allows the entry of minimum and maximum limiting values. If explicit limits are required then enter the limiting values in the boxes under the values of Mu and Sigma.

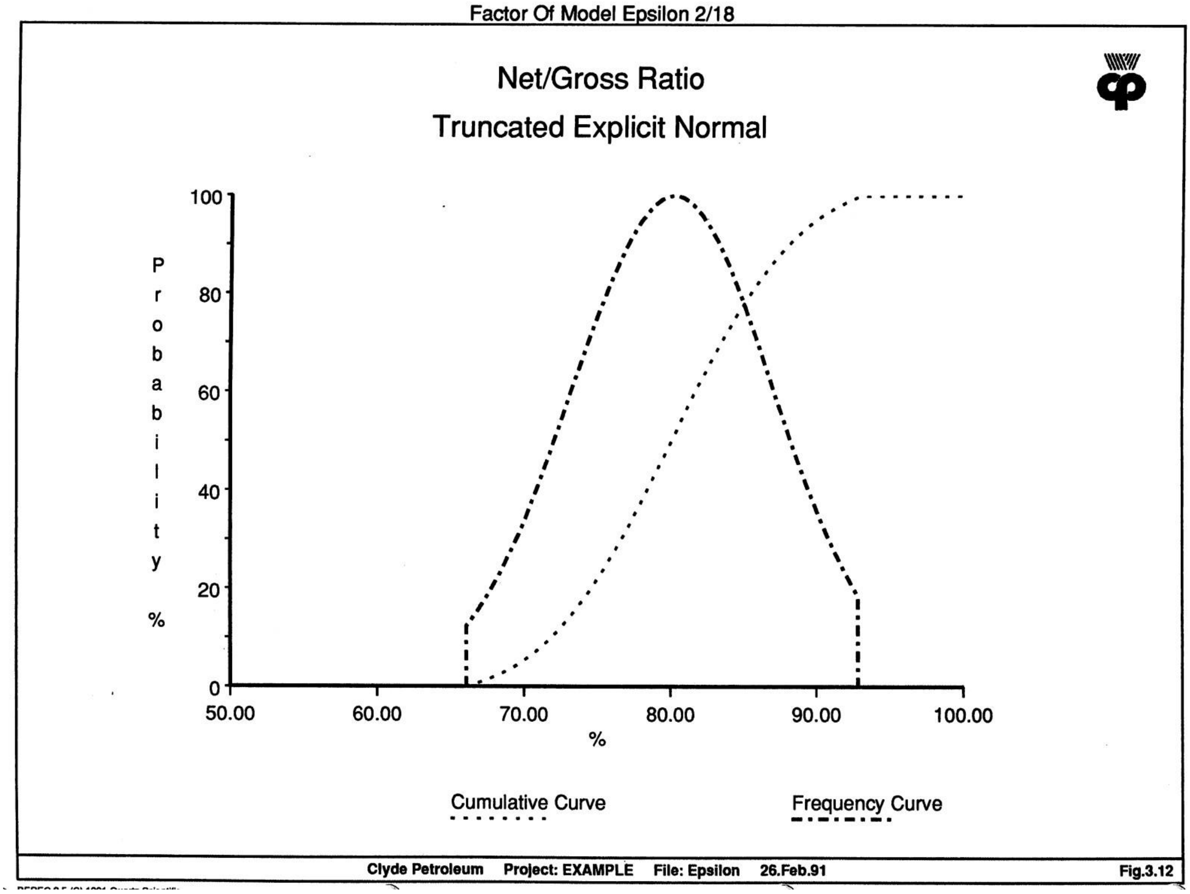
Assumed limits

In the absence of a specified minimum value a lower limit of zero is assumed for all factors as values must be positive at all times. If no upper limit is given, the program is not checking for a maximum value, except if the factor has percentages as values. For percentages, an upper limit of 100 % is assumed.



Action taken

If factor validation indicates a problem, the user may choose to ignore it and let the simulation program to deal with it. Alternatively, the user can change the data file. Either use the same model type with adjusted parameter values or change the model type. E.g. take a plot of the factor distribution curve. Redraw the frequency and enter the values read off the plot as a frequency model.



Truncated model

If validation indicates an error condition, then a truncated model is fitted. Truncated normal and log-normal distribution models are normal/log-normal curves cut at the low, high or both tails.

Such models have 2 further parameters for low and high limits in addition to mean and variance. Given the low and high limits the program finds the best fitting values of mean and variance. These are determined in a similar way as for the unconstrained models. E.g. optimising by the 'equal area' rule. The example plot shows a truncated normal distribution.

The effect of dependencies

The prospect evaluation program provides for a degree of dependency between geological factors. The level and nature are described by a single number on a scale between -10 and +10.

The + or - sign indicates positive or negative correspondence with a strength of magnitude 1 to 10, where 10 is the strongest level of correlation. A value of 0, indicating independence, is the usually assumed default.

The treatment of dependency by Monte-Carlo simulation is based on the statistical model of correlated, linearly dependent multivariate data. The joint distribution of two linearly correlated variables is ellipsoid with a sloping main axis that can be determined by regression or factor analysis. The unconditional distribution of the variables can be derived from the joint distribution.

PEREC addresses the inverse problem: given the unconditional distribution of each variable, it generates random samples for the variables with the desired level of correlation.

In the first instance, let us consider two correlated variables. These may be factors to be multiplied in calculating single reservoir expectations or two reservoirs to be added by the Monte-Carlo prospect addition program.

Input to modelling consists of the cumulative distribution curve (any given distribution can be expressed in cumulative form).

Dependency between variables is given as number, representing, in statistical terms, the correlation coefficient of the square transform of the variables.

Correlated pairs of random values are generated as follows: a random number is generated between 0 and 100 % and projected to the cumulative curve of the first variable. This gives the value for the first variable. Then, another random number between 0 and 100 % is generated to obtain a value for the second variable.

Now, if the two variables have no dependency then their correlation coefficient is zero. Random numbers are then generated independently if this is the case and the projected values are not related.

On the other hand, if variable ***b*** is dependent on variable ***a***, then the choice of the first random number r1 restricts the choice of r2. This is achieved by generating r1 and r2 as independent random numbers some of the time and letting r2 = r1 at other times.

If negative dependency is desired, then the non-random choices of r2 will be set to r2= 1 - r1. For example, let the level of dependency between a and b be -30%. Then for each pair of random values v1 and v2 three random numbers r0, r1, r2 are generated in the range of 0 to 100 %. The random number r0 determines whether r2 is generated independently (70% of the time) or set to 1-r1 (which happens 30% of the time).

The projection of r1 to curve 1 gives v1 in the range of variable a; r2 projected to curve 2 gives v2 in the range of variable b. The pairs (r1, r2) will be linearly correlated with a correlation coefficient exactly matching the dependency specified.

The pairs (v1, v2) show a correlation that is a function of the unconditional distribution functions. It will be exact for dependencies of -10, 0, +10 but the intermediate values may produce correlation coefficients differing from the dependency parameter in the same way as the covariance of the bivariate distribution differs from the covariance of its square transform.

Exact correspondence in the above context means that the expected value of the correlation coefficient is equal to the specified dependence.

If the prospect has more than two reservoir properties (geological factors) that are interdependent, data selection is performed using the same principles as for dependency with two variables.

Account is taken of the fact that a particular variable may become indirectly related to other properties as a result of interdependence within the group of variables. E.g. if B depends on A and C is specified as dependent on B then this determines a dependency between A and C. So the program will not allow entering a further specification for dependency between C and A.

Factor display

The validated factor distributions can be displayed in graphic form by the Factor display option on the Modelling menu. This program asks for the name of a data file and then presents the list of the factors in the file as a menu. The factors can be plotted by typing the number on the list.

The default horizontal scale for factor distribution plots is displayed by the program. This can be changed by specifying the desired range when prompted by the program, after displaying summary results and just before plotting starts.

Curves drawn

The plots show the factor distribution as three superimposed curves. The stepped graph of input data is shown by a thin continuous line, while the fitted distribution is shown by heavier dashed (cumulative curve) or dotted lines (frequency curve). The curves are all scaled to 100 % vertically - this means that the frequency curve is multiplied by a scaling factor. On colour screens and printers each curve is shown in a different colour.
(Perec Version 2.5 User Guide, Quartz Scientific, Watford, 1991, p.7-12)