MIXFIT: A MICROCOMPUTER-BASED ROUTINE FOR FITTING HETEROGENEOUS PROBABILITY DISTRIBUTION FUNCTIONS TO DATA

R. A. Cooke, S. Mostaghimi

ABSTRACT. *Some engineering data represent a mixture of data from two distinct populations or result from two or more concomitant phenomena. Homogeneous probability density functions are not always suitable for representing such data. This article describes the development of a procedure for fitting two types of heterogeneous density functions to mixed population data. The benefits derived from using these heterogeneous functions are evaluated with conventional goodness-of-fit tests.* ***Keywords.*** *Heterogeneous, Microcomputer, Probability distributions, Statistics.*

In engineering practice, data that represent the influence of two or more populations, are not uncommon. Mixed population data are encountered in areas as diverse as hydrology (Singh, 1974) and signal processing (Jiang and Kececioglu, 1992). In soil and water engineering, for example, it has long been recognized that there are two distinct processes governing the movement of water and aqueous solutions throughout the soil, namely flow through the soil matrix and flow through preferential (by-pass) flow paths. Flow and transport through the soil matrix, and flow and transport through preferential flow paths are not necessarily dependent on each other.

In response to the existence of mixed population data, the objective of this study was to develop two-parent (heterogeneous) probability distribution functions. The term ***heterogeneous distribution*** was coined by Hald (1952) to refer to distributions which represent the effects of two or more distinct populations arising from distinct phenomena. By way of contrast, a homogeneous distribution is one that is fitted to data from a single population, or representing a single phenomenon. Although heterogeneous distributions are sometimes referred to as mixed distributions, Hawkins (1974) stressed the necessity of differentiating between mixed-functions distribution and mixed-variables distribution.

A mixed-variables distribution (MVD) is applicable to mixed population data, in which each observation is the combined effects of two or more processes. For this case, Hawkins (1974) uses the example of peak streamflow, which in many cases is made up of direct runoff, interflow and groundwater flow. A given flow sample would reflect the combined effect of one or more of these three flow processes.

A mixed-functions distribution (MFD) refers to the case where the data being fitted represents sampling from two or more distinct populations that result from distinct processes. However, each datum is the result of one process, to the exclusion of all others. One typical example is rainfall. Rainfall can result from convective, frontal, or orographic processes. However, each individual rainstorm normally results from only one of these three processes. Ostensibly, a given distribution for annual rainfall could be decomposed into three overlapping distributions. The distributions overlap since a rainstorm with a given rainfall amount, say 25 mm, may result from any of these processes. The decomposition process for rainfall distributions might lead to more accurate simulation. Rojiani et al. (1985) found that summer and winter rainstorms in Blacksburg, Virginia, are best represented by different distributions. Cooke and Mostaghimi (1992) suggested that the use of a different semivariogram for each process might lead to more accurate results when using the kriging method to spatially interpolate rainfall. A schematic diagram indicating the difference between a MFD and a MVD is shown in figure 1.

Mixed population data occur in many areas of research. Titterington et al. (1985) compiled a list of 110 instances in the past century in which heterogeneous distributions have been fitted to obvious mixed population data. In well over 75% of these cases, the data were represented by mixtures of Gaussian or exponential distributions. The use of these component distributions is due more to convenience than to goodness-of-fit considerations. Apart from procedures for fitting mixtures of two or more Gaussian or log Gaussian distributions, mixtures of exponentials, and a few others, there are no existing procedures for fitting most heterogeneous distributions. The procedures introduced in this article allow for more than an order of magnitude increase in the number of heterogeneous distributions that can be fitted.

FITTING PROCEDURES

The fitting routine (referred to as MIXFIT) was developed to fit a wide range of heterogeneous probability functions to sample data and to provide both visual and formal statistical interpretation of the goodness-of-fit. The procedures in the routine were encoded in Microsoft QUICKBASIC with some additional FORTRAN routines. These procedures do not utilize any moment above the

Article was submitted for publication in February 1994; reviewed and approved for publication by the Soil and Water Div. of ASAE in August 1994.

The authors are **Richard A. Cooke**, Assistant Professor, Dept. of Agricultural Engineering, University of Illinois, Urbana-Champaign; and **Saied Mostaghimi**, ***ASAE Member Engineer***, Associate Professor, Dept. of Biological Systems Engineering, Virginia Tech, Blacksburg, Va.

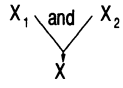
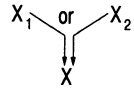
	Mixed-Variables Distributions	Mixed-Functions Distributions
Variables :	$X_1 : \mu_1, \sigma_1$ ρ $X_2 : \mu_2, \sigma_2$	$X_1 : \alpha_1, \mu_1, \sigma_1$ $X_2 : \alpha_2 = 1 - \alpha_1, \mu_2, \sigma_2$
Supply Diagram:		
Sample:	$X = X_1 + X_2$	$X = X_1$ or X_2
Moments:	$\mu = \mu_1 + \mu_2$ $\sigma^2 = \mu_1^2 + \mu_2^2 + 2\rho\sigma_1\sigma_2$ $E(X^k) = E(X_1 + X_2)^k$	$\mu = \alpha_1\mu_1 + \alpha_2\mu_2$ $\sigma^2 = \alpha_1\sigma_1^2 + \alpha_2\sigma_2^2 + \alpha_1\alpha_2 (\mu_1 - \mu_2)^2$ $E(X^k) = \alpha_1 E(X_1^k) + \alpha_2 E(X_2^k)$
Density Function:	$f(x) = \int f_1(X-Z).f_2(Z)dZ$	$f(X) = \alpha_1 f_1(X) + \alpha_2 f_2(X)$
Example:	Monthly stream flows from ground water **plus** surface water flow	Annual flood peaks from thunderstorms **or** from hurricanes

Figure 1–Distinction between mixed-variables and mixed-functions distributions (Hawkins, 1974).

second order, and are markedly different from the use of regression methods for fitting multiple parameter distributions.

MIXFIT was modeled after VTFIT, a routine developed for fitting homogeneous distribution functions (Cooke et al., 1993). The VTFIT model was developed to fit probability distribution functions to data, using the maximum likelihood method. The routine can be used to fit 18 distributions. The maximum likelihood estimates of the distribution parameters are obtained from the sample moments for those distributions where these estimates are simple functions of the sample moments. For the others, with the exception of the Log-Pearson type III, the parameters are estimated by direct maximization of the log likelihood function, using an optimization algorithm developed by Rosenbrock and Storey (1966). If, after a fixed number of user specified iterations, the maximization procedure fails to converge, then the parameters are obtained by maximizing the log multinomial distribution function. The parameters of the Log-Pearson type III are evaluated using the method of mixed moments developed by Rao (1980). This fitting procedure has been shown to be superior to the maximum likelihood procedure when applied to this distribution (Arora and Singh, 1989). For those distributions in which the location parameter is estimated, the method presented by Kline and Bender (1990) is used to obtain an initial estimate of the location parameter. The forms of the distributions that can be fitted with VTFIT, and the methods used to evaluate the parameters of these distributions are summarized in table 1. The formulae for evaluating the test statistics of the goodness-of-fit tests used in VTFIT are given in table 2. Test statistics are provided for the chi-square, Kolmogorov-Smirnov, Kuiper, Cramer-von Mises, Anderson-Darling and maximum likelihood goodness-of-fit tests. Graphs of the frequency histogram with the fitted probability distribution function superimposed, and of the empirical distribution function for the sample data, together with the fitted cumulative density function, are provided for visual assessment.

In MIXFIT, the procedures used in VTFIT are used to obtain initial estimates of the parameters for the component distributions. These parameters are then adjusted until a selected goodness-of-fit function is optimized. Unlike homogeneous distributions, the maximum likelihood function cannot be used as an object function since it is not always well behaved. Policello (1981) demonstrated that the maximum likelihood function is unbounded at every observation for a mixture of two Gaussian distributions and Jiang and Kececioglu (1992) showed that for some mixtures of Weibull distributions the likelihood functions have multiple local maxima. Attempts to use the maximum likelihood function as an object function for many of the other mixtures possible with this routine, revealed that the likelihood function is, in general, not reliable for heterogeneous distributions.

Mixed-Variables Distributions (MVDs)

In the literature, there are very few examples of the development of a distribution function for mixtures. In one example, Houghton (1978) introduced what was then a new five parameter distribution which he called the Wakeby distribution. The form of this distribution is such that it can best be expressed in terms of an inverse distribution function. With this distribution, a variable, x, is given by:

$$x = -a(1 - U)^b + c(1 - U)^d + e \quad (1)$$

where a, b, c, d, and e are the five distribution parameters and U is the Uniform (0,1) variate. Unlike other distributions with more than three parameters, this distribution can be fitted without using the sample variance or any higher sampling moments (Houghton, 1978).

Singh (1979) demonstrated that the Wakeby distribution is, in actuality, a heterogeneous distribution suitable for representing a mixed variable phenomenon with two component variables: The variable x could alternately be expressed as follows:

$$x = x_1 + x_2 + e \quad (2)$$

$$x_1 = -a(1 - U)^b \quad (3)$$

$$x_2 = c(1 - U)^d \quad (4)$$

where x_1 and x_2 are the component variables, and e is a location parameter. The fitting procedure proposed by Houghton (1978) for this distribution effectively fits x_1 to the left tail and x_2 to the right tail. In addition to being suitable for representing a mixed variable phenomenon, such as saturated hydraulic conductivity or steady state infiltrability, variates conforming to the Wakeby distribution can easily be generated with a uniform random number generator. Any homogeneous distribution function

Table 1. Distributions in VTFIT

Distribution	Density Function*	Parameter Evaluation Method†
	Extreme Value Distributions	
Gumbel (maxima)	$f(x) = \frac{e^{-(x-\gamma)/\beta} e^{-e^{-(x-\gamma)/\beta}}}{\beta}$	B
Gumbel (minima)	$f(x) = \frac{e^{-(\gamma-x)/\beta} e^{-e^{-(\gamma-x)/\beta}}}{\beta}$	B
Frechet (maxima, 2P)	$f(x) = \frac{\alpha}{\beta}\left(\frac{\beta}{x}\right)^{\alpha+1} e^{-(\beta/x)^{\alpha}}$	A
Frechet (maxima, 3P)	$f(x) = \frac{\alpha}{\beta}\left(\frac{\beta}{x-\gamma}\right)^{\alpha+1} e^{-[\beta/(x-\gamma)]^{\alpha}}$	B,C
Frechet (minima)	$f(x) = \frac{\alpha}{\beta}\left(\frac{\beta}{\gamma-x}\right)^{\alpha+1} e^{-[\beta/(\gamma-x)]^{\alpha}}$	B,C
Weibull (maxima)	$f(x) = \frac{\alpha}{\beta}\left(\frac{\gamma-x}{\beta}\right)^{\alpha-1} e^{-[(\gamma-x)/\beta]^{\alpha}}$	B,C
Weibull (minima, 2P)	$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-[x/\beta]^{\alpha}}$	E
Weibull (minima, 3P)	$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} e^{-[(x-\gamma)/\beta]^{\alpha}}$	B,C
	Other Distributions	
Gaussian	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[(x-\mu)^2/(2\sigma^2)]}$	A
Log-Gaussian (2P)	$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-[(\ln x-\mu)^2/(2\sigma^2)]}$	A
Log-Gaussian (3P)	$f(x) = \frac{1}{(x-\gamma)\sqrt{2\pi\sigma^2}} e^{-\{[\ln(x-\gamma)-\mu]^2/(2\sigma^2)\}}$	B,C
Exponential (1P)	$f(x) = \beta e^{-\beta x}$	A
Exponential (2P)	$f(x) = \beta e^{-\beta(x-\gamma)}$	B,C
Beta	$f(x) = \frac{x^{\alpha_1-1}(1-x)^{\alpha_2-1}}{\text{Beta}(\alpha_1,\alpha_2)}$	B
Gamma (2P)	$f(x) = \frac{\beta^{-\alpha} x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)}$	D
Gamma (3P)	$f(x) = \frac{\beta^{-\alpha}(x-\gamma)^{\alpha-1} e^{(x-\gamma)/\beta}}{\Gamma(\alpha)}$	B,C
Log-Pearson Type III	$f(x) = \frac{e^{\gamma/\beta} x^{\beta-1}}{\lvert\beta\rvert\Gamma(\alpha)}\left(\frac{\ln x-\gamma}{\beta}\right)^{\alpha-1}$	D
Inverted Gamma	$f(x) = \frac{x^{-(\alpha+1)} e^{-\beta/x}}{\beta^{-\alpha}\Gamma(\alpha)}$	D

* **Parameter Notation**
α = Shape parameter.
β = Scale parameter.
γ = Location parameter.

† **Parameter Evaluation Method**
A = Parameters estimated from sample moments.
B = Parameters estimated by direct optimization of log-likelihood function.
C = Initial estimate of location parameter obtained following Kline and Bender (1990).
D = Parameters estimated from sample moments, then adjusted for moment bias.
E = Parameters evaluated by solving equations given by Law and Kelton (1991).

can be mapped unto a Uniform (0,1) distribution by integrating the distribution function (Shiparo and Brain, 1981). The Wakeby distribution should be able to approximate homogeneous distribution functions as well.

Table 2. Goodness-of-fit tests used in VTFIT and MIXFIT

Test	Test Statistic
Log-Likelihood	$\ln(L) = \sum_{i=1}^{n} \ln\left[\hat{f}(x_i)\right]$
Kolmogorov-Smirnov	$D = \max\left\{D^+, D^-\right\}$
Kuiper	$V = D^+ + D^-$
Cramer-von Mises	$W^2 = \sum_{i=1}^{n}\left[\hat{F}(x_i) - \frac{2i-1}{2n}\right]^2 + \frac{1}{12n}$
Anderson-Darling	$A^2 = -\frac{1}{n}\sum_{i=1}^{n}\left\{(2i-1)\ln\left[\hat{F}(X_{(i)})\left\{1-\hat{F}(X_{(n-i+1)})\right\}\right]\right\} - n$
Chi-Square*	$\chi^2 = \sum_{j=1}^{k}\frac{(N_j - np_j)^2}{np_j}$

* For this test the entire range of the fitted distribution has to be first divided into k adjacent intervals $[a_0,a_1)$, $[a_1,a_2)$, . . . , $[a_{k-1}, a_k)$

Notation:

$\hat{f}(\ )$ Fitted probability density function.

$\hat{F}(\ )$ Fitted cumulative distribution function.

D^+ $\max\left\{\frac{i}{n} - \hat{F}(X_{(i)})\right\}$ for i = 1 to n

D^- $\max\left\{\hat{F}(X_{(i)}) - \frac{i-1}{n}\right\}$ for i = 1 to n

$X_{(i)}$ x_i sorted in ascending order

N_i Number of x_is in the jth interval $[a_{j-1}, a_j)$

p_j $\int_{a_{j-1}}^{a_j} \hat{f}(x)\,dx$

Houghton (1978) demonstrated the ability of the Wakeby distribution to duplicate the log-Gaussian distribution, while Singh (1979) demonstrated that the Wakeby distribution can be used to duplicate a mixture of two Gaussian distributions.

A heterogeneous distribution function for mixtures may be expressed in the following manner:

$$x = x_1 + x_2 \tag{5}$$

where x_1 and x_2 are the components of the mixture, x. The Wakeby distribution, a member of this family of heterogeneous distributions, includes an explicitly defined location parameter. However, the location parameter may be implicit in either or both of the components. Alternatively, x may be expressed in terms of the distributions of the two components:

$$x = INV(F_1) + INV(F_2) \tag{6}$$

where F_1 and F_2 are the cumulative distribution functions for x_1 and x_2, respectively, and the function INV(F) represents the inversion of the cumulative distribution function, F. The distribution of x is a distribution of mixtures or sums, alluded to by Hawkins (1974). If it is assumed that the two component processes are independent, then the density function is given by:

$$f(x) = \int_{z=0}^{\infty} f_1(x-z) \cdot f_2(z)\,dz \tag{7}$$

where $f_1(\ )$ and $f_2(\ )$ are the density functions for the two component processes. By a suitable change of dummy variables (let $\tau = x - z$), equation 7 becomes:

$$f(x) = \int_{-\infty}^{x} f_1(\tau) \cdot f_2(x-\tau)\,d\tau \tag{8}$$

If both $f_1(x)$ and $f_2(x)$ are bounded at zero, then equation 8 is simply given by:

$$f(x) = \int_{0}^{x} f_1(\tau) \cdot f_2(x-\tau)\,d\tau \tag{9}$$

The right hand side of equation 9 is the so-called *convolution integral* which can be evaluated either by transformation to the frequency domain using the Laplace or Fourier transform methods, or by numerical schemes. In general,

$$f(x) = \int_{ll}^{x-ll} f_1(\tau) \cdot f_2(x-\tau)d\tau \qquad (10)$$

where ll is the minimum of the lower limit of either of the component distributions. Except in cases where ll is zero and the product of the two functions are tractable, f(x) has to be evaluated numerically for any x value. Although some distributions, most notably the Gaussian distribution, has a lower bound at negative infinity, this does not pose an insurmountable problem since approximate lower bounds can always be determined. For example, a Gaussian distribution with mean, μ, and standard deviation, σ, may be considered to have a lower limit of $\mu - 3.5\sigma$, where the ordinate of the density function is less than 0.001. Although similar limits may be set for other unbounded distributions, most of the distributions used in engineering are bounded by zero in their lower tails.

By definition, the cumulative distribution function is the integral of the density function and may be expressed as follows:

$$F(x) = \int_{x=-\infty}^{x} f(x)dx \qquad (11)$$

For the distribution of mixtures shown in equation 7 above,

$$F(x) = \int_{-\infty}^{x} \int_{z-ll}^{\infty} f_1(x-z) \cdot f_2(z)dzdx \qquad (12)$$

Since the component distribution functions are, at worst, piecewise continuous for all real numbers, the integrals can be interchanged. Thus,

$$F(x) = \int_{z-ll}^{\infty} f_2(z) \left[\int_{-\infty}^{x} f_1(x-z)\,dx \right] dz \qquad (13a)$$

$$= \int_{z-ll}^{\infty} f_2(z) \left[\int_{-\infty}^{x-z} f_1(x)\,dx \right] dz \qquad (13b)$$

$$= \int_{z-ll}^{\infty} \left[F_1(x-z) - F_1(-\infty) \right] f_2(z)dz \qquad (13c)$$

Since $F_1(x)$ identically vanishes at the lower limit,

$$F(x) = \int_{z-ll}^{\infty} F_1(x-z) \cdot f_2(z)dx \qquad (14)$$

$$F(x) = \int_{ll}^{x-ll} F_1(\tau) \cdot f_2(x-\tau)d\tau \qquad (15)$$

The right hand side of equation 15 is another convolution integral that can be evaluated, as mentioned earlier.

Equations 6, 10, and 15 constitute a triad that completely describes the distribution of the mixture of two populations. In the development of this triad, the only restriction placed on the component distributions in the mixture is that they be independent. These equations are used as building blocks in the development of a procedure to fit mixed-variables distributions to data.

The method of fitting the component distributions is dependent on the nature of the heterogeneous distribution. For mixed-variables distributions, the steps in fitting procedure are as follows.

1. The data are ordered in descending order and each data point partitioned into two components using an s-curve separation procedure developed for this activity. In the s-curve procedure, the *ith* value of the first component in the mixture is given by:

$$X_{li} = \begin{cases} 2^{k-1}\left(\dfrac{i}{N+1}\right)^k X(i); & i < \dfrac{N}{2} \\ \left[1 - 2^{k-1}\left(1 - \dfrac{i}{N+1}\right)^k\right] X(i); & i \geq \dfrac{N}{2} \end{cases} \qquad (16)$$

$$\text{for } k \geq \frac{-\log 2}{\log(N+1) - \log 2}$$

where $X_{(i)}$ is the *ith* value among the ordered sample, N is the sample size and k is the order of the s-curve. The s-curves for orders ranging from the lower limit to 10 are shown in figure 2. The use of the zero order s-curve (k = 0) results in the division of each observation in the dataset into two halves. In the limit as the order increases, all the observations with a rank less than N/2 are attributed to first named component, and the other observations attributed to the other component. Thus, the partitioning method used by Houghton (1978) to fit the parameters of the Wakeby distribution may be interpreted as an infinite order s-curve method. For practical purposes, a high k value (10 or more) would give the same effect. In negative order s-curves (k < 0), the greater portion of the observations with ranks less than N/2 are attributed to the second named component in the mixture. The assignment of a value for k is a subjective procedure that influences the initial estimates of the parameters of the fitted distribution, thereby influencing the time required to obtain final parameter estimates. k does not significantly influence the value of the final parameter estimates. A zero value is usually adequate except in cases where the values for one component of the mixture is in excess of an order of magnitude more than the other.
2. x_1 is fitted to a homogeneous distribution function. The distribution may be any of the 18 distributions specified in VTFIT. x_2 is obtained by subtracting x_1 from x, that is,

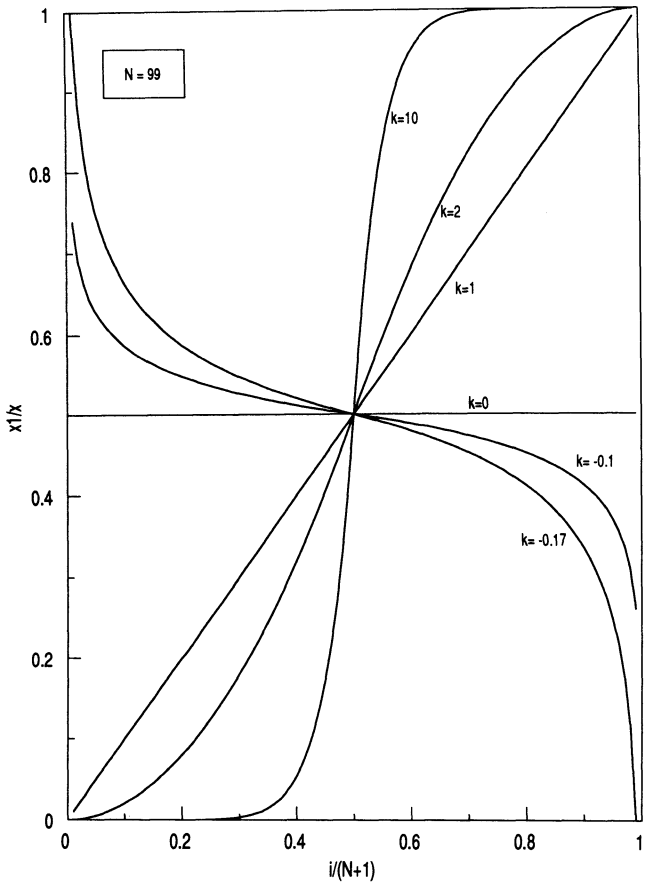


Figure 2–S-curves for apportioning data in fitting mixed-variables distributions.

$$X_{2i} = X_{(i)} - INV\left[F_1\left(\frac{N-i+1}{N+1}\right)\right] \qquad (17)$$

and fitted to a second distribution function. As with x_1, this may be any of the distributions in VTFIT, including the distribution fitted to x_1.

3. The values of the parameter estimates for the two distributions fitted in step 2 are adjusted until a specified distance measure, or object function, is optimized. This optimization is performed with a modified version of the Constrained Rosenbrock Procedure presented by Kuester and Mize (1973). This procedure is a multivariate constrained optimization method that allows for nonlinear constraints, and does not require specification of the derivative of the object function. The algorithm presented by Kuester and Mize (1973) was modified to allow for a choice of any one of four object functions and to allow interactive termination. The four object functions were obtained from a list of possible functions provided by Titterington et al. (1985). The likelihood function is included on this list but it was not included in the routine since it is unbounded for some heterogeneous distributions.
4. Visual plots of the data and the fitted distributions are displayed and the six goodness-of-fit statistics are evaluated. This evaluation requires the inversion of the cumulative distribution function and the determination of the density and the cumulative distribution functions. These three tasks are facilitated by equations 6, 10, and 15, respectively. In order to solve equations 10 and 15, algorithms were developed for numerical convolution.

Mixed-functions Distributions (MFDs)

Mixed-functions distributions are more commonly cited in the literature than mixed-variables distributions. The idea of fitting mixed-functions distributions was first suggested by Pearson (1894) in an effort to explain the reasons why a data set did not conform to a Gaussian distribution, even after the data was logarithmically transformed. He suggested the distribution that best fitted the data could be decomposed into a number of overlapping Gaussian distributions, and developed a moments-based routine to decompose a distribution into two such components. His routine involved the estimation of the coefficients of a ninth order polynomial expression. Cohen (1967) modified the method and reduced the solution to the evaluation of the first and second moments of each of the component distributions and of the proportion of each distribution in the mixture. Singh (1974) also developed a least squares method of fitting a mixture of two Gaussian distributions to monthly streamflows and annual floods. Medgyessy (1977) later showed, in a rigorous manner, that the decomposition could be extended to more than two parent distributions, and developed procedures for determining the optimum number of Gaussian distributions into which a distribution could be decomposed.

Potter (1958) put the decomposition procedure on a phenomenological basis when he suggested that the "doglegs" and reverse curvatures occurring in some flood frequency curves, resulted from the sampling of different populations that represented distinct processes. Stoddard and Watt (1970) added credence to Potter's theory when they successfully developed a distribution for extreme floods in Ontario by superimposing the distribution functions for summer rainfall floods and winter snowmelt floods. Upchurch (1970), as reported by Hawkins (1974), also used a graphical method to decompose the distribution for beach sediments in the Great Lakes into three component distributions, each representing a different depositional process.

The contributions by Potter (1958), Stoddard and Watt (1970), Upchurch (1970), Hawkins (1974), and Singh (1974, 1987) were all related to mixed functions distributions, using either Gaussian or log Gaussian component distributions. Medgyessy (1977) developed procedures for combining two or more exponential distributions while Rossi et al. (1984) developed a distribution consisting of the mixture of two extreme value distributions.

The typical method of parameter estimation for mixed functions distributions is to a) separate the data into two roughly equal parts; b) estimate the parameters for each half; and c) iteratively adjust the parameters until the difference between observed and fitted deviations for the distribution function is minimized. Hirshboeck (1987) developed a novel method for specifying mixed functions distributions for partial duration flood series in which she

divided the data into subpopulations based on the prevailing climatic conditions when each flow event occurred. She then fitted separate Gaussian distributions to each subpopulation and combined them based on the percentage of the total number of events in each subpopulation. Such a method is only suitable if, in addition to the data series itself, there is other information available, such as in this case where the type of storm producing each flow event was known.

The general equation for a distribution function, that consists of a mixture of two homogeneous distribution functions, is given by

$$f(x) = \alpha f_1(x) + (1 - \alpha) f_2(x) \quad (18)$$

where $f_1(x)$ and $f_2(x)$ are the component density functions, and α is a proportioning constant (Hawkins, 1974). The cumulative density function may be determined by integrating the expression for f(x). Thus,

$$F(x) = \int_{x=-\infty}^{x} [\alpha f_1(x) + (1 - \alpha) f_2(x)] dx \quad (19a)$$

$$= \int_{x=-\infty}^{x} \alpha f_1(x) dx + \int_{x=-\infty}^{x} (1 - \alpha) f_2(x) dx \quad (19b)$$

Since α and $(1 - \alpha)$ are constants, they may be taken outside the integral sign. Thus,

$$F(x) = \alpha F_1(x) + (1 - \alpha) F_2(x) \quad (20)$$

where $F_1(x)$ and $F_2(x)$ are the cumulative density functions corresponding to $f_1(x)$ and $f_2(x)$, respectively. x may be expressed in terms of the cumulative density functions:

$$x = INV(F) = INV(\alpha F_1 + F_2 - \alpha F_2) \quad (21)$$

and has to be numerically evaluated for most distribution functions.

Equations 18, 20, and 21 completely specify mixed functions distributions and are used to develop a fitting routine. For these distributions, the steps in fitting procedure are as follows.

1. The data are ranked in ascending order and separated into two groups either by randomly assigning the observations to the groups, or by putting the smaller observations in one group and the larger in another. In both cases this corresponds to initially setting α to a value of 0.5.
2. Both groups are then fitted to user-selected homogeneous distribution functions. The distributions may be any of the 18 distributions specified in VTFIT.
3. The fit is optimized as described in step 4 of the mixed-variables routine. The only difference is that the optimization of mixed functions distributions involves an extra parameter, namely α, the partitioning coefficient. α is constrained to vary between 0.0 and 1.0, inclusive of the end points.
4. Visual plots of the data and the fitted distributions are displayed and the goodness-of-fit statistics are evaluated. This evaluation requires the inversion of the cumulative distribution function; and the determination of the density and the cumulative distribution functions. These three tasks are facilitated by equations 21, 18, and 20, respectively. The specified object functions were the same as those used for the mixed variables distribution.

Sample Applications

Application to Flood Flow Analysis. MIXFIT was used to fit mixed Gaussian MFDs to maximum flow rates from four USGS gauging stations in Illinois, shown to be best represented by mixture distributions (Singh, 1974). In addition to evaluating the goodness-of-fit statistics for each distribution, MIXFIT was also used to estimate the 10-year and 100-year return-period peak-annual flow events. The 10-year and the 100-year return-period peak-annual flow events correspond to the 90th and 99th percentile value of the fitted distribution, respectively.

The goodness-of-fit statistics for the data from the four stations are shown in table 3. The table also shows the corresponding statistics obtained by fitting Log-Pearson type III distributions to the same data. For most distributions, including the heterogeneous distributions that can be fitted with MIXFIT, the distributions of the goodness-of-fit statistics are not known. However, corresponding goodness-of-fit statistics resulting from fitting different distributions to the same data can be compared if certain assumptions are made about the distributions of these test statistics.

The distribution of a goodness-of-fit statistic resulting from fitting a distribution with known parameters to a data set is independent of the distribution fitted to the data (Shiparo and Brain, 1981). If the distribution of a goodness-of-fit statistic resulting from fitting a distribution with **unknown** parameters is approximated by the distribution resulting from fitting a distribution with **known** parameters, then it is possible to compare the fit quality of different distributions fitted to the same data. Such a comparison, although not exact, can be used to select a best fitting distribution.

Table 3. Goodness-of-fit test statistics for selected Illinois streamflow data

Distribution	Kolmogorov-Smirnov	Kuiper	Cramer-von Mises	Anderson-Darling	90th percentile (ft^3/s)	99th percentile (ft^3/s)
South Branch Kishwaukee River near Fairdale						
Log-Pearson Type III	0.091	0.176	0.398	2.288	4,718	7,841
Gaussian Mixture	0.046	0.086	0.050	0.410	5,150	11,759
Henderson Creek near Oquawka						
Log-Pearson Type III	0.091	0.176	0.189	1.437	6,557	12,984
Gaussian Mixture	0.069	0.133	0.188	1.132	6,778	16,268
Salt Creek near Rowell						
Log-Pearson Type III	0.109	0.210	0.254	1.613	6,408	15,121
Gaussian Mixture	0.081	0.162	0.218	1.452	7,933	18,918
Skillet Fork at Wayne City						
Log-Pearson Type III	0.137	0.270	0.907	5.717	12,267	30,654
Gaussian Mixture	0.087	0.170	0.178	1.574	4,904	23,510

In this instance, the values of the goodness-of-fit statistics for the mixed-Gaussian MFDs, are smaller than their counterparts for the Log-Pearson type III distributions. Smaller values are, in a general sense, indicative of an improved fit. The most remarkable differences occurred with the Cramer-von Mises statistics for the Kishwaukee River near Fairdale, and Skillet Fork at Wayne City. In both instances the values of this statistic for the MFDs are less than a third of the corresponding values for the Log-Pearson type III distributions. This statistic is a measure of the mean deviation between the fitted cumulative density function and the empirical distribution function for the data. It can thus be interpreted that, for this data, the heterogeneous distributions provide an improvement in fit quality, as compared to the homogeneous distributions.

The fitted distributions and the empirical distribution function (EDF) for the Kishwaukee River near Fairdale are shown in figure 3. Both fitted distributions match the empirical distribution function closely, but the superiority of the mixed-Gaussian distribution is visually evident. The Log-Pearson type III distribution overpredicts the magnitude of floods between the 20th and the 70th percentiles, and underpredicts flood magnitudes between the 70th and the 95th percentile. In the upper tail, the MFD mirrors the behavior of the empirical distribution more closely than the other distribution.

The Anderson-Darling statistic is weighted to emphasize deviations in the tails of distributions. If this statistic is used for comparison, then the mixed-Gaussian distributions fit the tails of the data better than the Log-Pearson type III distributions. The behavior of a distribution at the extremes is of special importance in risk analysis. In the United States, the Interagency Advisory Committee on Water Data (1982) recommends that the Log-Pearson type III be applied to flood flow analysis. The value for the 100-year flood at Henderson Creek near Oquawka obtained by using this distribution is about 20% less than the value obtained by using the mixed-Gaussian distribution. The use of the 99th percentile (100-year event) of the Log-Pearson distribution in a design associated with flood mitigation might result in an unexpectedly high number of events exceeding the design flow.

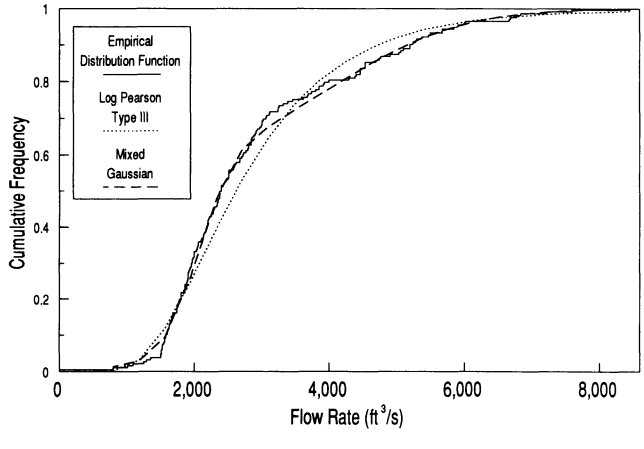


Figure 3–Comparison of empirical distribution function and cumulative density functions fitted to streamflow data from the South Branch Kishwaukee River near Fairdale, Ill.

Application to Steady-state Infiltrability Measurements. Loague and Gander (1990) presented a set of steady-state infiltrability measurements that reportedly conforms to a log Gaussian distribution. The data set consists of 157 measurements, taken on a 25 × 25 m grid, from a 10 ha experimental watershed at Chickasha, Oklahoma. The measurements were taken with a large-diameter single-ring infiltrometer as described by Loague (1990a). Previously, the data has been used in the geostatistical characterization of infiltration (Loague and Gander, 1990), in the reevaluation of a rainfall-runoff model (Loague, 1990b), and in the determination of the effectiveness of using soil texture as a surrogate for saturated hydraulic conductivity (Loague, 1992a). The measurements may be considered to be independent since Loague and Gander (1990) also determined the minimum sampling distance for independence to be 5.0 m. Independence is assumed in the development of all the goodness-of-fit tests (Shiparo and Brain, 1981). The goodness-of-fit statistics for the three best fitting distribution functions in VTFIT, and three of the best fitting distributions from each of the two classes of distributions in MIXFIT, are presented in table 4. As in the above example on flood flow, the 90th and 99th percentile values for each distribution are also presented. Figure 4 shows the empirical distribution function for the data, with one cumulative density function from each class of distributions superimposed on the same graph.

The difference in fit quality between the distributions is visually insignificant, even in the upper tails of the distributions. However, the goodness-of-fit statistics paint a different picture. Corresponding goodness-of-fit statistics, particularly the Cramer-von Mises and Anderson-Darling statistics, are much smaller for the heterogeneous distributions, which might be an indication that they fit the data better than the homogeneous distributions. In this example, the homogeneous distributions appear to underestimate percentile values in the upper tail of the data.

Mixed-variables vs. Mixed-functions Distributions

In the second example application above, not only did the heterogeneous distributions provide better fits than the homogeneous distributions, but both classes of heterogeneous distributions had comparable goodness-of-

Table 4. Goodness-of-fit test statistics for R-5 infiltrability data

Distribution	Kolmogorov-Smirnov	Kuiper	Cramer-von Mises	Anderson-Darling	90th percentile (10^{-5} m/s)	99th percentile (10^{-5} m/s)
Homogeneous Distributions						
Log Gaussian (3P)	0.054	0.102	0.070	0.477	3.08	5.70
Gamma (3P)	0.064	0.122	0.058	0.427	3.08	5.15
Weibull	0.068	0.129	0.081	0.490	3.13	5.12
Mixed Variables Distributions (MVDs)						
Gamma/ Weibull	0.048	0.082	0.047	0.355	3.21	5.28
Log Gaussian/ Gaussian	0.042	0.079	0.044	0.068	3.34	7.89
Weibull/ Exponential	0.050	0.098	0.065	0.072	3.70	6.69
Mixed Functions Distributions (MFDs)						
Log Gaussian (3P)/ Exponential	0.040	0.080	0.034	0.373	3.85	7.70
Log Gaussian/ Frechet	0.043	0.080	0.029	0.288	3.17	13.82
Log Gaussian/ Exponential	0.040	0.077	0.025	0.332	3.69	7.37

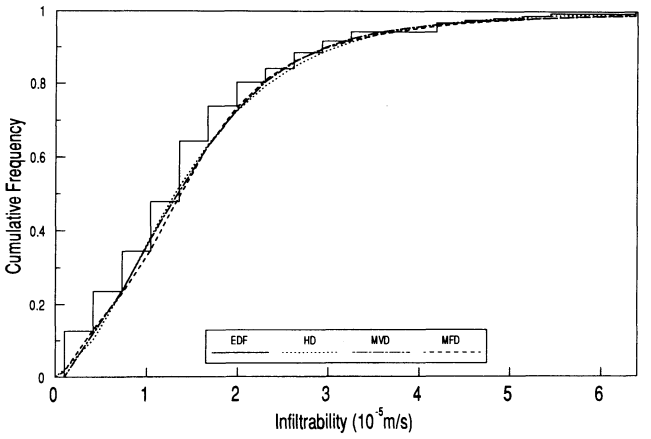


Figure 4–Comparison of empirical distribution function and cumulative density functions fitted to R-5 infiltrability data.

fit statistics. The question naturally arises as to which class of heterogeneous distributions are best suited to a specified process. There is a precedent for discussion of this nature in the literature. Houghton (1978) fitted a set of flood flow data to the Wakeby distribution, shown to be a mixed-variables distribution. Singh (1979) fitted the same data to a mixed-functions distribution, consisting of two log Gaussian distributions, and obtained a better fit. Both classes of heterogeneous distributions performed better than any homogeneous distribution. Singh (1979) attributed the improved fit of the mixed-functions distribution, over and against the mixed-variables distribution, to the fact that the underlying premise of the mixed-functions distributions was more in line with the processes that gave rise to the floods, that is, each flood mainly resulted either from one process or another, but not from both. It, therefore, appears that the justification for each class of distributions is phenomenologically based.

Summary and Conclusions

Procedures were developed for fitting heterogeneous distribution functions to data. These fitting procedures are novel, and are markedly different from the use of regression methods for fitting multiple parameter distributions. The fitting procedures do not utilize any moment above the second order, and allow for more than an order of magnitude increase in the number of heterogeneous distributions that can be fitted.

In most of the cases where the fitting of heterogeneous distributions resulted in better goodness-of-fit statistics, large change was observed in the Anderson-Darling goodness-of-fit statistic. This statistic is a measure of how well the given distribution fits the data in the tails. In these cases, therefore, applications that rely on the extreme values of a random variable would benefit from the use of heterogeneous distributions.

Hardware Requirements and Software Availability

The minimum system requirements for running MIXFIT include an IBM or IBM compatible computer with a graphics adapter and 256k of free memory. Additional recommended requirements are a hard disk with at least 1 MB of free memory for program installation, an 80286 or faster processor, an 80 × 87 math co-processor, and a color monitor.

Copies of the software may be obtained by contacting either author at the Biological Systems Engineering Department, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

References

Arora, K. and V. P. Singh. 1989. Comparative evaluation of the estimators of the Log-Pearson type (LP) 3 distribution. *J. of Hydrol.* 105(1):19-37.

Cohen, A. C. 1967. Estimation in mixtures of two normal distributions. *Technometrics* 9(1):15-28.

Cooke, R. A. and S. Mostaghimi. 1992. A microcomputer based routine for obtaining mean watershed precipitation from point values. *Computers and Geosciences* 18(7):823-837.

Cooke, R. A., S. Mostaghimi and F. Woeste. 1993. VTFIT: A microcomputer-based routine for fitting probability distribution functions to data. *Applied Engineering in Agriculture* 9(4):401-408.

Hald, A. 1952. *Statistical Theory with Engineering Applications.* New York: John Wiley.

Hawkins, R. H. 1974. A note on mixed distributions in hydrology. In *Proc. of the Symp. on Statistical Hydrol.*, 336-344. U. S. Dep. Agric. Misc. Pub. 1275.

Hirschboeck, K. K. 1987. Hydroclimatically-defined mixed distributions in partial duration flood series. In *Hydrologic Frequency Modeling*, ed. V. P. Singh, 199-212. Dordrecht, Holland: D. Reidel Publishing Company.

Houghton, J. C. 1978. Birth of a parent: The Wakeby distribution for modeling flood flows. *Water Resources Research* 14(6):1105-1109.

Interagency Advisory Committee on Water Data. 1982. Guidelines for determining flood flow frequency. Bulletin #17B. Reston, Va.: Hydrology Subcommittee, OWDC, U. S. Geological Survey.

Jiang, S. and D. Kececioglu. 1992. Maximum likelihood estimates, from censored data, for mixed-Weibull distributions. *IEEE Transactions on Reliability* 41(2):248-255.

Kite, G. W. 1975. Confidence limits for design events. *Water Resources Research* 11(1):48-53.

Kline, D. E. and D. A. Bender. 1990. Maximum likelihood estimation for shifted Weibull and lognormal distributions. *Transactions of the ASAE* 33(1):330-335.

Kuester, J. L. and J. H. Mize. 1973. *Optimization Techniques with Fortran.* New York: McGraw-Hill.

Loague, K. 1992. Using soil texture to estimate saturated hydraulic conductivity and the impact of rainfall-runoff simulations. *Water Resources Bulletin* 28(4):687-693.

———. 1990a. Simple design for simultaneous steady-state infiltration experiments with ring infiltrometers. *Water Resources Bulletin* 26(5):935-938.

———. 1990b. R-5 revisited 2. Reevaluation of a quasi-physically based rainfall-runoff model with supplemental information. *Water Resources Research* 26(3):973-987.

Loague, K. and G. A. Gander. 1990. R-5 revisited 1. Spatial variability of infiltration on a small rangeland catchment. *Water Resources Research* 26(5):957-971.

Medgyessy, P. 1977. *Decomposition of Superpositions of Distribution Functions.* New York: John Wiley.

Pearson, K. 1894. Contributions to the mathematical theory of evolution. *Royal Soc. London, Phil. Trans.* 185:71-110

Policello, G. E. II. 1981. Conditional maximum likelihood in Gaussian mixtures. In *Statistical Distributions in Scientific Work*, Vol. 5, eds. C. Taillie, G. P. Patil and B. A. Baldessari, 111-125. Dordrecht, Holland: D. Reidel Publishing Company.

Potter, W. D. 1958. Upper and lower frequency curves for peak rates of runoff. *Am. Geophys. Union Trans.* 39(1):100-105.

Rao, D. V. 1980. Log Pearson type 3 distribution: Method of mixed moments. *J. of the Hydraulics Division, ASCE* 106(3):999-1019.

Rojiani, K. B., K. A. Tarbell, V. O. Shanholtz and F. E. Woeste. 1985. Simulation of correlated rainfall amount and duration data. *Transactions of the ASAE* 28(1):129-132.

Rosenbrock, H. H. and C. Storey. 1966. *Computational Techniques for Chemical Engineers.* New York: Pergamon Press.

Rossi, F., M. Fiorentino and P. Versace. 1984. Two-component extreme value distribution for flood frequency analysis. *Water Resources Research* 20(7):847-856.

Shiparo, S. S. and C. W. Brain. 1981. A review of distributional testing procedures and development of a censored sample distributional test. In *Statistical Distributions in Scientific Work*, Vol. 5, eds. C. Taillie, G. P. Patil and B. A. Baldessari, 1-24. Dordrecht, Holland: D. Reidel Publishing Company.

Singh, K. P. 1987. Development of a versatile flood frequency methodology and its application to flood series from different countries. In *Hydrologic Frequency Modeling*, ed. V. P. Singh, 183-198. Dordrecht, Holland: D. Reidel Publishing Company.

————. 1979. Comment on "Birth of a parent: The Wakeby distribution for modeling flood flows" by John C. Houghton. *Water Resources Research* 15(5):1285-1287.

————. 1974. Two-distribution methods for fitting mixed distributions in hydrology. In *Proc. of the Symp. on Statistical Hydrol.*, 336-344. USDA Misc. Pub. 1275.

Stoddart, R. B. L. and W. E. Watt. 1970. Flood frequency prediction for intermediate drainage basins in southern Ontario. C. E. Res. Rpt. 66. Kingston, Ontario: Dept. Civil Engineers, Queens University.

Titterington, D. M., A. F. M. Smith and U. E. Makov. 1985. *Statistical Analysis of Finite Mixture Distributions.* New York: John Wiley & Sons.

Upchurch, S. B. 1970. Mixed population sediment in nearshore environment. Paper presented at Great Lakes Research Conference, Buffalo, N.Y.