On the application of mixtures of two log-normal distributions to the analysis of water quality data

P.T. ADAMSON and M.J. DIXON*
Directorate of Water Affairs, Private Bag X313, Pretoria 0001.

Abstract

This paper is primarily concerned with the application of a direct search technique to optimise the parameters of a mixture of two log-normal distributions in a hydrological field where data are distributed according to a mixing of component distributions. Such data plot on probability paper as a zig-zag curve and although sampled as a single phenomenon two or more separate populations are evident. In the analysis of river water quality it is reasonable to expect that samples taken in the dry season will be distributed according to a different mean and variance to those taken in the wet season. The procedures examined herein allow the decomposition of the data into its component distributions and the consequent estimation of the seasonal parameters associated with river water quality. A number of examples and results are shown and the computer programs used, including those routines for plotting the data, are available along with test data and results.

Introduction

Many sets of earth-science data plot on probability paper as zig-zag curves (Tanner, 1958) and are commonly modelled using the Pearson Type I or Type IV distribution. Such data samples are generally drawn from a mixture of two or more component distributions and the result is designated a "compound" or "mixed" distribution. These models are pertinent to many practical situations in sedimentology, geomorphology, meteorology and hydrology. For example, the distribution of particle sizes in a sediment sample will depend on the clastic properties of the source rocks so that a number of distributions of grain size will be evident from a single sieve analysis. Potter (1958) and Ashkanazy and Weeks (1975) have observed a mixture of random variables in the distribution of floods generated by various types of synoptic situation and snowmelt. Even the components of the hyposometric curve of the earth have been analysed using a compound distribution (Tanner, 1962).

Compound normal distributions have accounted for most of the work in the dissection of mixed distribution phenomena. Other mixtures of continuous distributions have been widely studied in the statistical literature but of these only a compound extreme value model has been applied to earth-science data, in fact to non-homogenous flood data. (Canfield, *et al*, 1980). Mixtures of discrete distributions such as the compound Poisson could for example be fruitful in the analysis of the arrivals of extremes where the extremes themelves, such as storm depths, are distributed according to a mixture of components.

Hawkins (1974) has distinguished between mixed distribution and mixed variable phenomena. The former fall into the "either-or" category in the sense that the data sample originates from a number of discrete sources. For example, floods may be generated by rainfall events or snowmelt, extreme rainfalls may be initiated by warm fronts or cold fronts or convection. The mixed variable phenomenon is rather more complex and involves a measurement of components already in a combined state. In other words each member of the sample reflects a mixture of sources as opposed the total sample reflecting a mixture of sources. The carbonate content of streamflow, for example, reflects the balance between ground and surface runoff and the contribution of the carbonate content of each.

The statistical implications of the distinction are that different moment estimators should be used in the decomposition of the mixtures. Although water quality should generally be viewed as a mixed variable phenomenon it is felt that the mixed distribution approach offers much promise in the efficient estimation of percentiles, in the separation of seasonal components and in the physical interpretation of the annual cycle of parameter values. Additionally, the method of fitting proposed herein largely circumvents the statistical ordinances which a more theoretical and less workable approach might imply.

Existence of Mixed Distributions in Water Quality Data

Streamflow quality when expressed in terms of concentrations will, over time, reflect variations and trends in physical and man-made processes which to a greater or lesser degree will manifest themselves in the distribution of the sample. Thus, whilst sampled as a single phenomenon a number of distinct populations will be evident to which definite associations can be ascribed. For example, dilution effects in the flood season will imply that the data are distributed according to a different mean and variance than in the dry season when concentrations of certain ions may be higher as a consequence of the greater relative contributions of ground-water discharge. Irrigation practices in certain seasons may be reflected in a distinct population type, with return flows associated with leaching resulting in seasonally high values of salts. Effluent discharge may be seasonal and spillage or flushing from dams may be used to control and dilute pollutants. Such practices and catchment processes will elicit an effect upon the total sample of streamflow quality which will be heterogenous and can be thought of as drawn from a mixture of a number of relatively simple distributions. Although it may be straightforward to propose on physical grounds the number of separate distributions in the mixture, statistical considerations limit such mixtures to two or three, beyond which parameter estimation becomes less and less feasible without massive sample sizes. Even moment estimators for a two distribution model are dubious for sample sizes of less than 1 000 without prior knowledge of the parameters. (Hawkins, 1974).

Mixed distribution phenomena are most easily identified when plotted on log-probability paper using some empirical plotting position such as Weibull (rank/$(N+1)$). "Dog-legs" and

*Present address: c/o Murray & Roberts/Concor, P O Box 445, Plettenberg Bay 6600.

zig-zags are evident with generally straight lines between them. Figure 1 shows three such plots for chloride, TDS and hardness samples, two populations being evident for the TDS sample and three for those of chloride and hardness. Obviously, one cannot fit a single model to such data without recourse to the more esoteric members of the Pearson or Beta family of distributions with their associated fiendish estimators. Besides, the fact that the mixture of models has physical significance, supports the approach of attempting to decompose the data into its component distributions, thus providing some insight into the causative physical processes.

Plots and fits of a mixture of two log-normal models to a considerable number of water quality samples drawn from rivers and dams throughout South Africa have suggested such a model to be perfectly adequate in characterizing the probabalistic structure of water quality data as well as providing a possible means of interpreting the parameters and their seasonal variation.

Theory: The Mixed Log-Normal Model

The distribution of a mixture of two log-Normal distributions may be written as:

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

where

$$F_1(x) = \frac{1}{\sigma_1 (2\pi)^{1/2}} \int_{-\infty}^{x} exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\} dx \quad \ldots\ldots\ldots\ldots \quad (2)$$

and

$$F_2(x) = \frac{1}{\sigma_2 (2\pi)^{1/2}} \int_{-\infty}^{x} exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\} dx \quad \ldots\ldots\ldots\ldots \quad (3)$$

with x the log transformed data, α a proportionality factor and μ_1, μ_2, σ_1, σ_2 the parameters or the means and standard deviations of the component distributions. Furthermore, $\alpha + (1-\alpha) = 1$ and α is non-negative. The first non-central and the first five central moments of (1) are given by Cohen (1967) and equations linking the two distribution parameters with the mixed distribution parameters are (Singh, 1979):

$$\mu = \alpha\mu_1 + (1-\alpha)\,\mu_2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (4)$$

$$\sigma^2 = \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \alpha(1-\alpha)(\mu_2-\mu_1)^2 \quad \ldots\ldots\ldots\ldots \quad (5)$$

$$\gamma = [3\alpha(1-\alpha)(\mu_1-\mu_2)(\sigma_1^2-\sigma_2^2) + \alpha(1-\alpha)(1-2\alpha)(\mu_1-\mu_2)^3]/\sigma^3 \quad \ldots\ldots\ldots\ldots \quad (6)$$

where μ, σ and γ are the mean standard deviation and skewness of (1).

Generally, since five parameters need to be estimated from the sample, the first five sample moments are required. In water quality analysis and for hydrological data as a whole sample requirements for the confident estimation of such high order moments are rarely, if ever, available. Thus the nonic polynomial solution derived by Pearson (1894) and circumvented by Cohen (1967) for more practical applications are not applicable in the present study unless some a-priori assumptions are made about the parameters. Graphical analogue techniques used by Tung (1966) in the analysis of chromatograms are far too slow and complex for routine hydrological analysis, nor are they particularly accurate.

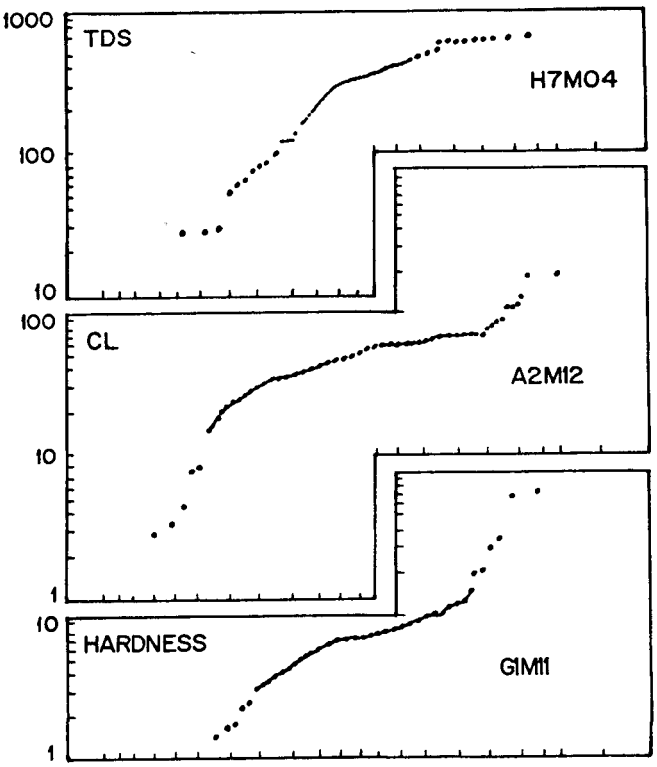

Figure 1

Examples of the mixed distribution of water quality criteria from natural streams: H7M04 – Huis River at Barrydale (TDS, mg/l); A2M12 – Crocodile River at Kalkheuwel (Chloride, mg/l); G1M11 – Watervals River at Watervalsberg (mg/l $CaCO_3$).

Singh (1968) and Singh and Sinclair (1972) proposed that the distribution of monthly flows and of annual flood peak maxima may be composed of a mixture of two log-normal distributions. "Optimal" values of the five parameters were obtained by minimising $\Sigma(\Delta Z)^2$ using equations (4) and (5) as constraints, where ΔZ is the observed Z minus the fitted Z and Z is the standard normal deviate. This is the scheme followed here where observed Z is computed via the Weibull plotting position and fitted Z from equation (1) given a set of parameter estimates. These are then 'optimised' in terms of $\Sigma(\Delta Z)^2$ using a direct search algorithm.

Explicitly we seek a constrained minimisation of the sum –

$$\Sigma(\Delta Z)^2 = \Sigma[Z(P=\frac{m}{N+1}) - Z(\alpha . F_1(x) - (1-\alpha) . F_2(x))]^2 \ldots \quad (7)$$

given initial estimates of α, μ_1, μ_2, σ_1 and σ_2. The possibility that the optimisation procedure converges to a local minimum rather than an absolute one was investigated by varying the initial parameter estimates between a range of values. For example, α was initialised at *0,01* and *0,99* with no effect on the final "optimal" solution. However, such routines are prone to convergence to local minima and the present application of a direct search technique should always be complimented by a plot of the data and the computed probability model to ensure that the solution is a realistic one. Such numerical optimisation techniques could also be used to compute the maximum likelihood estimators of the parameters, either by the maximisation of the likelihood, subject to the constraints of equations (4) and (5) or one could maximise the likelihood without constraints and obtain estimates of all five parameters simultaneously. The stability of such maximum likelihood estimators, required sample sizes and

comparison of the results with those derived in this report constitute a purely statistical piece of research and such results will be reported separately.

Verification of Model and Fitting Procedure

To illustrate the practical application of the mixed model of water quality criteria some two hundred samples of hardness, TDS content, chloride and sodium ion concentrations were investigated and plots made to assess performance and fit. Streamflow and reservoir data were considered with sample sizes ranging from 100 to 1 000. Agreement between observed and expected probability was found to be generally most satisfactory within the range of percentiles of hydrological interest. Below the 10% and above the 90% non-exceedence probabilities, which may be considered extremes, the results should be treated with caution, although a plot of the results soon reveals any flaws in the fit. The fitting procedure can, however, be easily weighted to provide a "best-fit" over any particular range of the data sample.

Figure 2 shows the distributions and associated frequency densities computed for the samples given in Figure 1. The log-normal distributions that have been decomposed from the original mixture are shown as F_1 and F_2. It is apparent that plots such as these can provide a considerable insight into the structure of such samples. The TDS data show a bimodal frequency density with the higher concentrations asymptotically approaching the log-normal distribution F_2 *(6,009; 0,335)* and the lower concentrations the log-normal distribution F_1 *(4,565; 0,629)*. The seasonal association of the two distributions can easily be established by further inspection of the data.

The distribution of the hardness sample is rather more complex with the high and low values asymptotically distributed as F_2 *(1,859; 1,024)*. The mid-range values have very low variance and are log-Normally distributed as F_1 *(1,954; 0,224)*. Again, the physical or seasonal association of the distributions could easily be investigated.

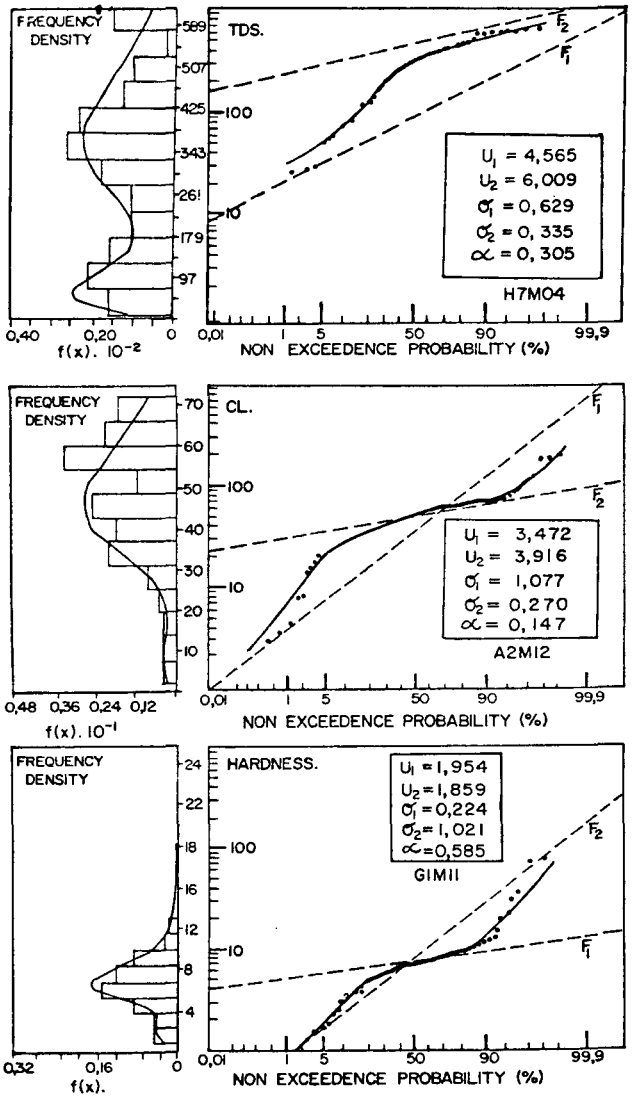

Figure 2
Fit of mixture of two log-normal distributions to the data of Figure 1 with associated decomposed log-normal distributions (F_1, F_2).

Robustness of the Mixed Log-Normal Model

Essentially the mixture of two log-normal distributions is a five parameter model for which, if classical moments estimators were used for parameter estimation would not provide a particularly stable result given the requirement of high order sample moment estimates. The fitting procedure recommended here, however, merely requires the estimation of the mean and variance directly from the sample, which then provide the constraints (equations (4) and (5)) for the optimization procedure. The virtue of this lies in the apparent robustness of the model given various sample sizes. Figure 3 shows the effect on percentile estimates of progressive sample size reduction. The total number of TDS samples available is 1 000 but these are distributed rather erratically throughout the history of the water quality station (c.f. Figure 4). Sampling progressed from random to weekly to regular daily analyses. The imposition of an increasing daily gap between selected analyses apparently affects the percentile estimates to a reasonably small degree, the difference between no gap and a 10 day gap being only about 8% for the estimate of the 50% percentile. Comparable results have been achieved for a number of stations and quality criteria.

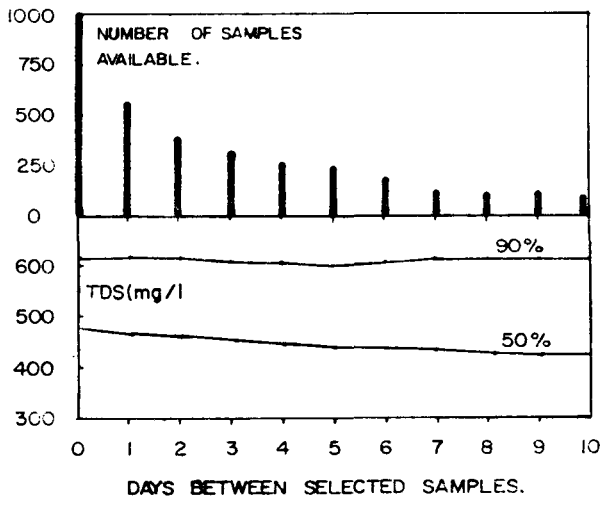

Figure 3
C2M22. Vaal River At Baalkfontein. Effect on estimates of 50% and 90% percentiles for progressive sample size reduction.

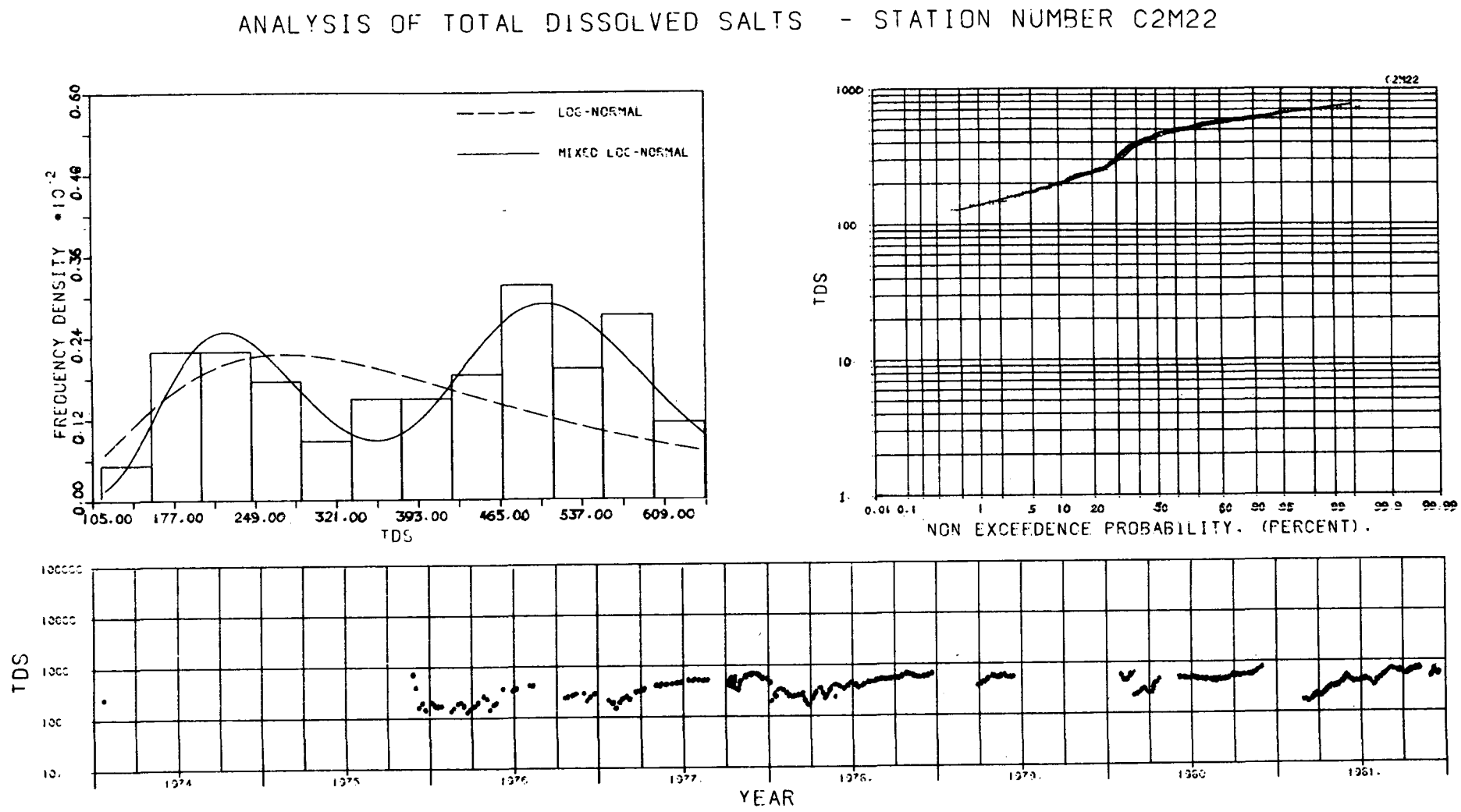

Figure 4(a)
***Sample plotter routine output for TDS data** (C2M22. Vaal River at Baalkfontein)*

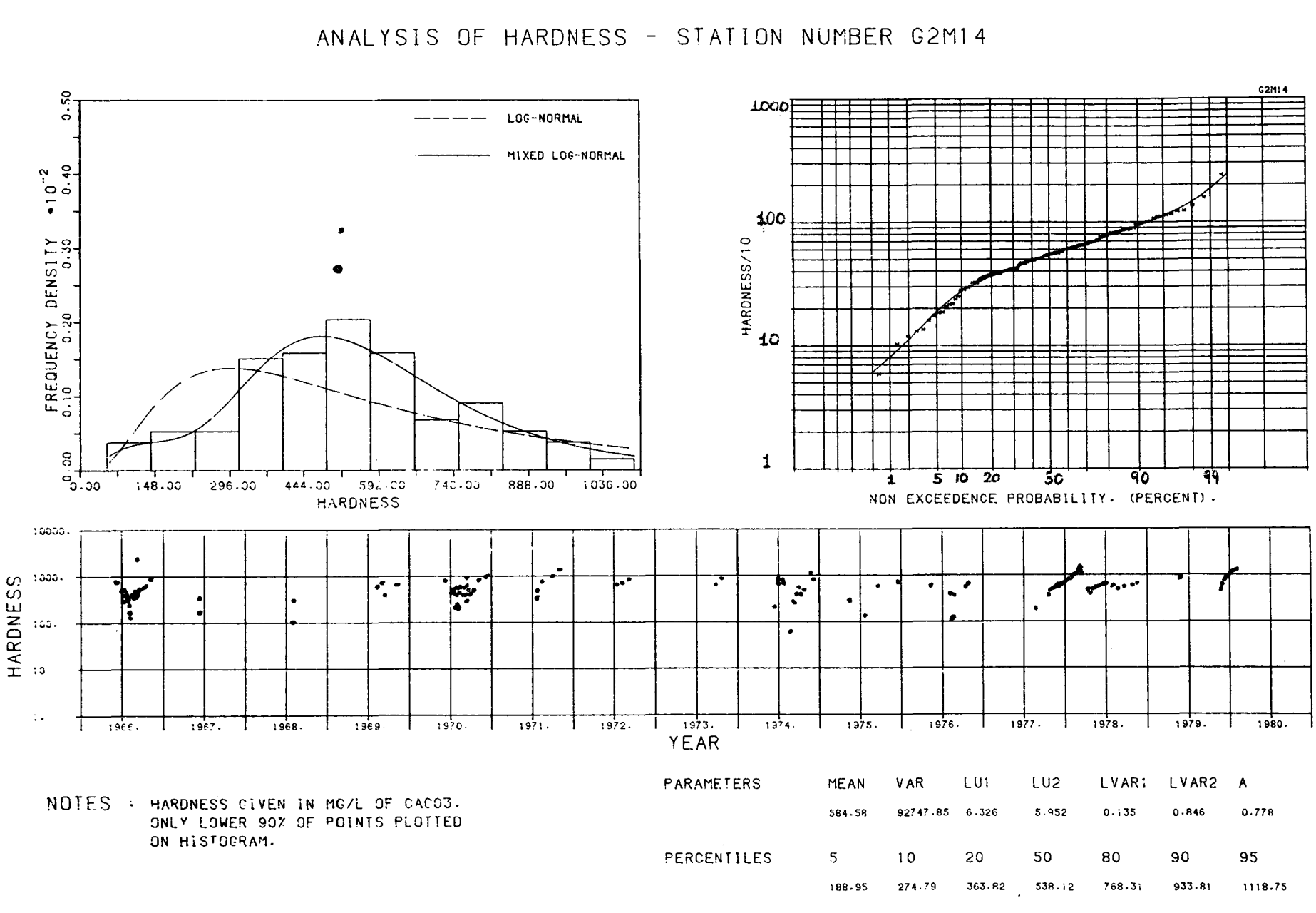

Figure 4(b)
Sample plotter routine output for hardness data (G2M14. Diep River at Visserhoek)

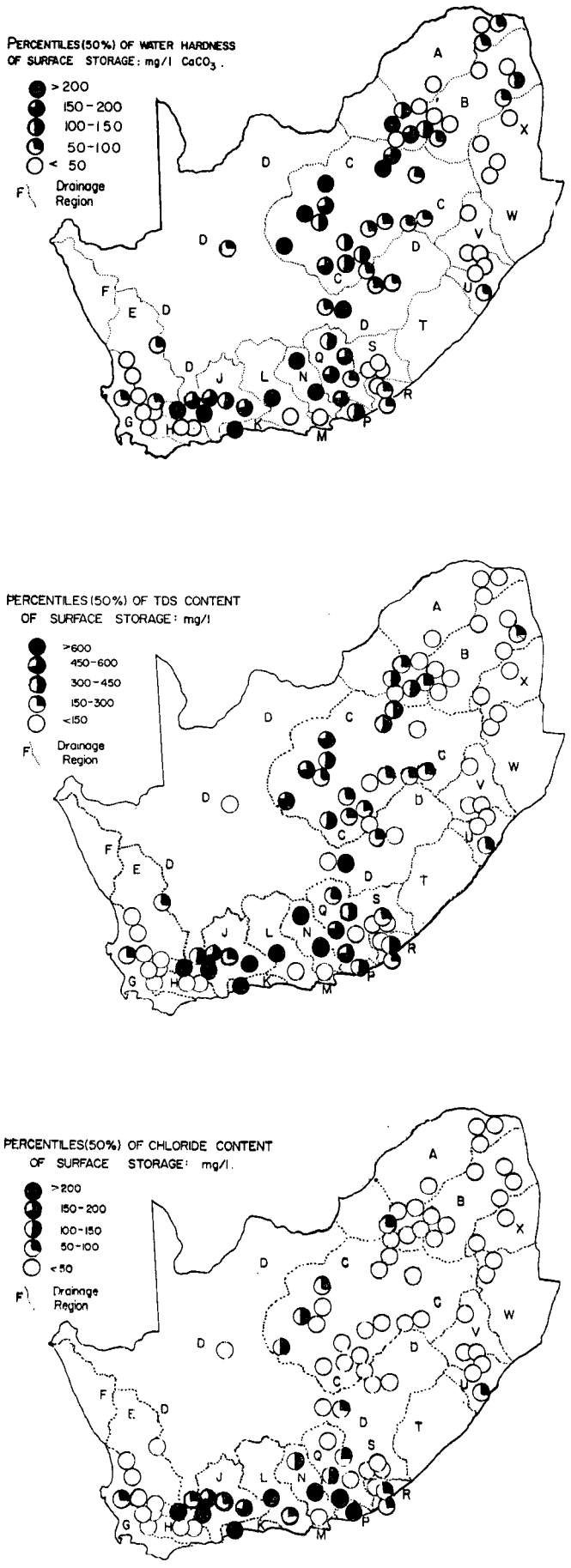

Figure 5
Application of the outputs from program system to assess the geography of the quality of South African surface water storage.

Computer Program and Plotter Routines

A system of programs and plotter subroutines has been designed to fit the mixture of log-normal distributions to water quality data and plot the results and sampling history. Two illustrative examples are shown in Figure 4. The water quality data bank available to the Directorate of Water Affairs is accessed directly and can provide data at any number of selected monitoring stations. Files of required quality criteria are then automatically created. For example, TDS is regressed upon electrical conductivity to extend a file where the latter is available when no TDS analysis was actually performed. The results of this operation are listed along with the plotted output. A direct search algorithm "optimises" the fit of the mixed distribution after sample mean and variance have been estimated and finally the input files to the plot routines are created.

The plot sequence is divided into two parts, namely that which gives the empirical and theoretical probabilities and those which provide the frequency density, sampling history, moments, percentiles and additional comments. All graphs and chronologies are automatically scaled and the fit of a single log-normal model to the sample is shown with the frequency density for comparative purposes. All programs are in FORTRAN IV and the plot routines designed for a CALCOMP flat-bed plotter. Print-outs with sample input and output are available upon request.

Processing times obviously vary with sample size. For 500 analyses the direct search algorithm for parameter estimation takes 1 min 40 s for 500 iterations. Total plot-time for the result is less than 5 min.

Potential Applications

An immediate and obvious application of the model is in the estimation and mapping of percentiles of quality criteria for dams and river monitoring stations. An example is shown in Figure 5 for South African surface storage where sample sizes ranged from 50 upwards and the gap between selected analyses was 23 days, or the mean gap between dated analyses. Dams, however, tend to be rather more conservative than rivers in terms of the range and variance of conservative quality criteria and any reasonable change in the gap between selected analyses elicits little effect on the overall result.

For the illustrative study of the geographical variability of catchment TDS and chloride concentrations shown in Figure 6, a 3 day gap was used.

The physical significance of the mixed model has been suggested earlier and plots such as those shown in Figure 2 certainly provide insight into the distribution and physical association of the data with various seasons, irrigation practices or dam operation. There is no doubt that such plots should be an essential part of preliminary data analysis and may provide indications of subsequent fruitful directions of investigation.

A potential and possibly important application of such a mixed model lies in the interpretation of the annual cycle of parameter estimates. Singh (1968) certainly provides convincing evidence that physical reasoning may be attached to the variations over the year of μ_1, μ_2, σ_1 and σ_2 in the analysis of monthly streamflows where the seasonal variability of the parameters is associated with the relative contributions of base and surface runoff to the total monthly volume. Attempts to interpret the seasonality of water quality in a similar way did not, however, meet with such success that any confident physical reasoning

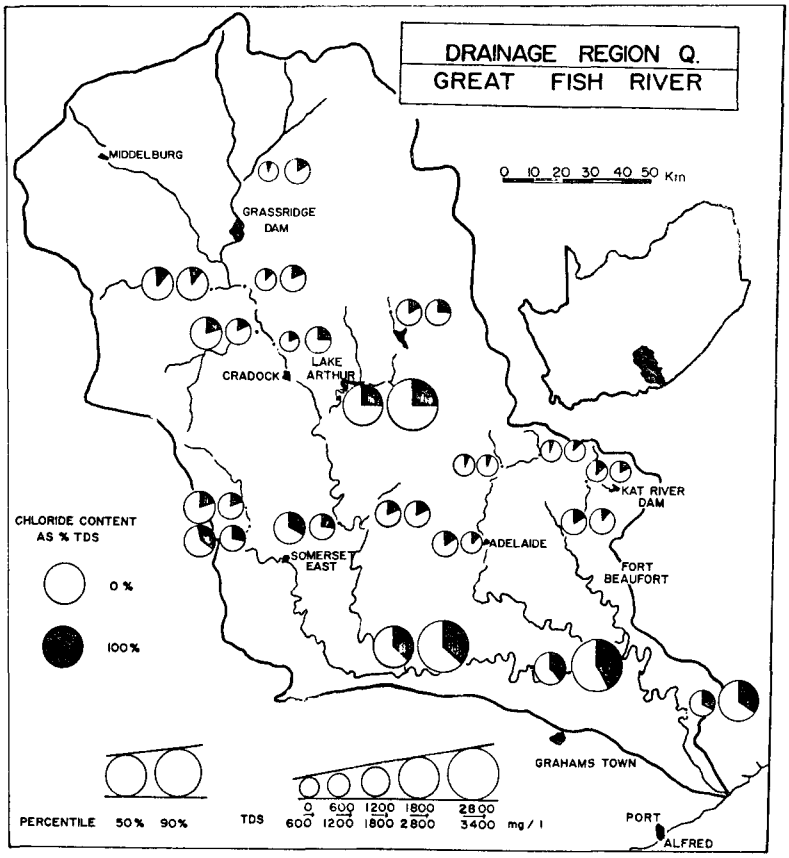

Figure 6
Application of the outputs from the program system to the mapping of catchment TDS and chloride content.

could be attempted. The data records as yet available in South Africa are too short to permit such monthly analyses. An outline study, however, is shown in Figure 7 for station C2M22 which provided the largest data base currently available (1 000 analyses). The monthly variation of TDS concentrations is shown to peak in December followed by a sudden fall and then a steady rise through the dry season. The 90% percentile is relatively unsteady as a consequence of extreme concentrations which would be expected to occur during very low flows. Such an annual cycle illustrates the effects of controlled flushing and natural spill from Vaal Dam causing a sudden dilution of TDS concentrations during January, February and March. The slow rise reflects decreasing releases and natural flows and consequent concentration effects.

The parameter estimates for the 12 models, one for each month, show some interesting results but sampling variance is also evident there being only 6 years of data available and an average of 80 samples attached to each model. The means noticeably play little or no role in controling the diversity of the distribution function from month to month, this being effected by the variances and proportionality factor. There is evidence of a serial relationship and seasonality in the magnitudes of σ_1, σ_2 and α. They are apparently not random. The cycle of σ_1 may be associated with a higher variance in the upper quantiles of concentrations during the low flow season with the reverse being true for σ_2. The behaviour of α generally indicates the relative importance of F_1 and therefore σ_1 during the low flow season.

Although this example is far from ideal it has served to illustrate the potential for the physical interpretation of parameter estimates and certainly once larger data bases become available such studies could prove rewarding.

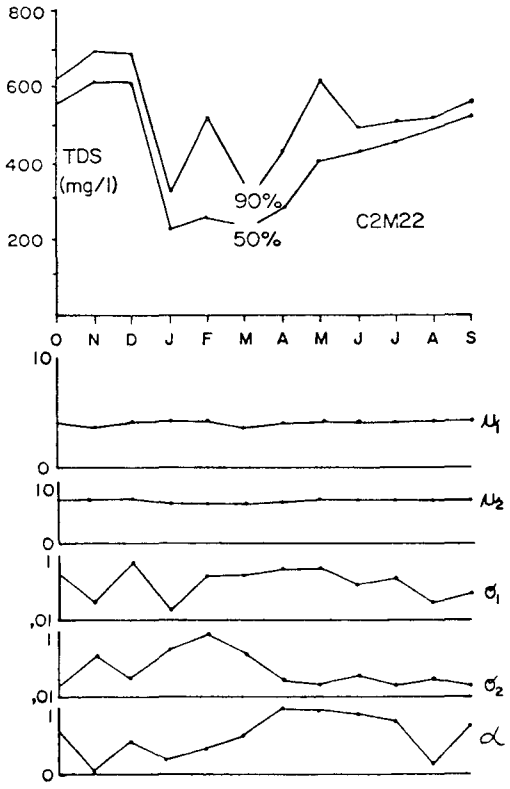

Figure 7
C2M22. Vaal River at Baalkfontein. Annual cycle of TDS percentiles and associated parameters of the mixed log-normal model.

Conclusions

The analysis of data suspected to be drawn from a mixture of component distributions is shown to be a feasible and relatively straightforward proposition. The estimation procedures recommended for a mixture of two log-normal models are workable, robust and appropriate to the analysis of water quality data. The model efficiently reflects the probabalistic structure of the concentrations of various ions and salts and is shown to have potential in the physical interpretation of the mixture and its parameters. Combined with automatic fitting and plotting algorithms such a scheme is felt to provide a useful addition to the presentation of data drawn from burgeoning data banks.

Acknowledgement

This paper is published with the permission of the Manager: Scientific Services, Directorate of Water Affairs, Department of Environment Affairs and Fisheries.

References

ASHKANASY, N.M. and WEEKS, W.D. (1975) Flood frequency distribution in a catchment subject to two rainfall producing mechanisms. Hydrology Symposium, Armidale, NSW, Australia. May 18-21 Institute of Engineers, Sydney, pp. 153-167.

CANFIELD, R.V.; OLSEN, D.R.; HAWKINS, R.H. and CHEN, T.L. (1980) Use of extreme value theory in estimating flood peaks from mixed populations. Hydraulics and Hydrology Series, UWRL/H-80/01. College of Engineering. Utah State University, Logan, Utah.

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

HAWKINS, R.H. (1974) A note on mixed distributions in Hydrology. Proceedings of Symposium on Statistical Hydrology. USDA. Ag. Res. Service, Misc. Pub. No 1275. Washington D.C., pp. 336-344.

PEARSON, K. (1894) Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society* **185** 71-110.

POTTER, W.D. (1958) Upper and lower frequency curves for peak rates of runoff. *Transactions American Geophysical Union*. **39** 100-105.

SINGH, K.P. (1968) Hydrologic distributions resulting from mixed populations. Proceeding IASH. Symposium. The use of Analog and Digital Computers in Hydrology. Vol. II. Tucson Arizona. IAHS publication No 81. p. 671-681.

SINGH, K.P. (1979) Comment on Birth of a Parent: The Wakeby distribution for modelling flood flows. By J.C. Houghton. *Water Resources Research* **15** (5) 1285-1288.

SINGH, K.P. and SINCLAIR, R.A. (1972) Two-distribution method for flood frequency analysis. *Journal. Hydraulics Division. American Society of Civil Engineers* **98** (HY1) 29-44.

TANNER, W.F. (1958) The zig-zag nature of Type I and Type IV curves. *Journal of Sedimentary Petrology* **28** 372-375.

TANNER, W.F. (1962) Components of the hypsometric curve of the earth. *Journal of Geophysical Research* **67** (7) 2841-2843.

TUNG, L.H. (1966) Method of calculating molecular weight distribution function from Gel Permeation Chromatograms. *Journal of Applied Polymer Science* **10** 375-385.