SAM: A Computer Program for Statistical Analysis and Modelling

J. Bai[12], A. J. Jakeman[2], M. McAleer[3] and J. A. Taylor[2]

1. Australian Bureau of Statistics, Canberra.
2. Centre for Resource and Environmental Studies, Australian National University
3. Department of Economics, University of Western Australia

Abstract A computer program, SAM, has been developed to undertake the sophisticated task of statistical analysis and modelling for PCs, UNIX stations and mainframes. This paper illustrates the various procedures and methodologies adopted within the computer package. SAM is a useful tool for practitioners in selecting an appropriate probability distribution from a set of alternatives to represent statistical, environmental, socio-economic, and engineering data. It also estimates the parameters of the distribution, predicts the desired percentile values, and calculates the minimum errors associated with the prediction. The package also provides extreme value, reliability and life-testing analyses that are essential in many scientific fields.

1. Introduction

While many statistical software packages are currently available, there is a demand for a more specialized computer program that can undertake complex tasks in relation to statistical data collection and analysis. SAM is a program which is specifically designed to be useful for those who have a special interest in using sophisticated statistical techniques to study data sets arising from the statistical, environmental, socio-economic, scientific and engineering fields.

In many applications, statistical data can usefully be summarised using probability distributions as an aid in assessment and in the decision making process. For example, the most useful environmental indicators in many cases are the mean, variability, extreme events, and the probability of exceeding specified values. Such indicators form the basis of many, if not most, environmental standards. All these indicators of environmental quality can be derived from correctly identified probability distributions and their parameter values. Reliability and life-testing analyses use such distributional forms to determine the probability of failure of life for particular scientific and engineering objects from progressive censoring.

SAM is based on extensive research into the applications of reliability and quality control, air and water quality, hydrology, meteorological prediction and forecasting, and road traffic and safety data. It has also expanded its applicability to a broad range of socio-economic, scientific and engineering phenomena. This program combines powerful identification tools with maximum likelihood parameter estimation in an easy-to-use package. The current version of the program is limited only by the applicability of the distributional models included in the program for the data in question. Focus is placed on the most popular and widely applicable probability distributions, such as the gamma, Weibull, lognormal, normal and exponential models. For the gamma, Weibull and lognormal distributions, estimation of the two- and three-parameter distributions, and the ability to select from among these distributions, have been included in the program.

Compared with many other computer programs used for statistical analysis and modelling, this program offers greater flexibility and applicability, with advanced state-of-the-art techniques and extensive proven application to a wide range of real data sets.

2. Program Overview

The program incorporates many useful design features based upon the authors' own experience in analysing a wide range of environmental, socio-economic, scientific and engineering data sets. These design features make the program powerful and easy to use, as is shown in the following.

User Interface

This program adopts an easy-to-use interface, based on menus that make the program both learner friendly and user friendly.

On-Line Help

An on-line help facility is available, with examples which provide user-selected levels of help for the new and experienced user. Program-generated error messages provide detailed descriptions of problems and the most likely remedies.

Data Processing

A flexible command structure allows both interactive and batch processing of data. The analysis of individual data

sets can be achieved rapidly using the program in an interactive mode. This feature is particularly useful when conducting an exploratory analysis. However, where a standard set of outputs is required or multiple data sets need to be analyzed, batch files can be used as inputs to the processor.

Data Input

Straightforward formats for data input speed up the analysis. A range of simple unformatted data files can be accepted by default. This program can also recognize the data format from most popularly-used databases.

Presentation of Results

The processor will produce a set of default or standard outputs. User-definable outputs can be easily added. A flexible plotting utility is available to aid in the analysis of the input data and the presentation of results. These include plots of autocorrelation, time series, regression and error analysis to aid in the assessment of data.

Modelling Application

A number of widely applicable probability distributions have been included. In the current version of the program, these are the normal, exponential, and two- and three-parameter gamma, Weibull and lognormal distributions.

Parameter Estimation

A key processor designed to perform the maximum likelihood estimation of parameters for the probability distribution models has been designed to be applicable to the full range of potential parameter values for each distribution. The parameter estimation processor has been tested on a wide range of actual environmental data sets.

Model Identification

A special rule-based expert system aids in the interpretation of the results. A comprehensive range of well-tested state-of-the-art identification and selection criteria have been incorporated into the program.

Monte Carlo Simulation

The program also includes simulation tools to enable sampling from the probability distributions listed above. Such simulation tools allow quantitative investigation of the ability of the identification and parameter estimation procedures to predict desired properties or derived quantities of probability distributions. For example, this facility allows the user to investigate the ability of the maximum likelihood approach to estimate a percentile of a particular distribution.

Another important simulation tool is the application of the bootstrap method to data sets in order to determine the variability in a desired user-specified quantity, such as a percentile or parameter value.

3. Statistical Features

3.1 Model Specification

Exponential and Normal Distributions

For the exponential and normal distributions, the reader is referred to the extensive literature available on these distributions in Kendall and Stuart [1979].

Gamma, Weibull and Lognormal Distributions

Probability density functions for the three-parameter gamma, Weibull and lognormal distributions for a random sample are given as follows:

Gamma:

$$f(x)=\frac{1}{\beta\Gamma(\alpha)}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x-\gamma}{\beta}\right)\right] \qquad (1)$$

Weibull:

$$f(x)=\frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}\exp\left[-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right] \qquad (2)$$

Lognormal:

$$f(x)=\frac{1}{\alpha\sqrt{2\pi}}(x-\gamma)^{-1}\exp\left\{-\frac{\left[\log(x-\gamma)-\beta\right]^2}{2\alpha^2}\right\} \qquad (3)$$

In equations (1), (2) and (3), α represents the shape parameter, β the scale parameter, γ the location parameter, and Γ the gamma function. The two-parameter versions of the density functions of the gamma, Weibull and lognormal distributions are the same as in (1), (2) and (3), with $\gamma=0$ in each case. In the above equations, $\alpha>0$, $\beta>0$ and γ is less than the minimum observed sample value.

The maximised value of the likelihood function is an essential statistic employed in many criteria used to discriminate among alternative models. For a sample x_1, x_2, ..., x_n of n independently and identically distributed random observations, the log-likelihood functions for the three-parameter gamma, Weibull and lognormal distributions are given as follows:

Gamma:

$$\log L=-n\alpha\log\beta-n\log\Gamma(\alpha)+(\alpha-1)\sum_{i=1}^{n}\log(x_i-\gamma)-\sum_{i=1}^{n}\left(\frac{x_i-\gamma}{\beta}\right) \qquad (4)$$

Weibull:

$$\log L = n\log\alpha - n\alpha\log\beta + (\alpha-1)\sum_{i=1}^{n}\log(x_i-\gamma) - \sum_{i=1}^{n}\left(\frac{x_i-\gamma}{\beta}\right)^{\alpha} \quad (5)$$

Lognormal:

$$\log L = -\frac{n}{2}\log(2\pi\alpha^2) - \sum_{i=1}^{n}\log(x_i-\gamma) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}\left[\log(x_i-\gamma)-\beta\right]^2 \quad (6)$$

The parameters of the three log-likelihood functions are estimated by maximum likelihood methods. Since the general maximum likelihood procedure for the three-parameter gamma and Weibull distributions will frequently fail to converge when the (unknown) shape parameter is less than or equal to unity, a computationally efficient approach that circumvents this problem is used (for further details, see Bai et al.1991).

Extreme Value Distributions

For N sets of samples, with each sample consisting of x_{1j}, x_{2j}, ..., x_{nj} of n independently and identically distributed random observations, X_1, X_2, ..., X_n is obtained from

$$X_j = \max\left(x_{1j}, x_{2j}, \ldots, x_{nj}\right), \quad j = 1,2,\cdots,N \quad (7)$$

The cumulative function of the underlying extreme value distribution of X_j is of the form $\left[F(x)\right]^n$. The type I version of the asymptotic extreme value distribution is generally given by its cumulative function as

$$\Pr\left(X_j < X\right) = \exp\left(-e^{-(X-\theta)/\upsilon}\right) \quad (8)$$

which approaches the true extreme value distribution as the sample size tends to infinity. θ and υ are the parameters related to the initial distribution that can be any form of distributional function described previously (for further details, see Gumbel [1958]).

Order Statistics

For an ordered sample x_1, x_2, ..., x_n of n independently and identically distributed random observations, the joint probability density function of the order statistics is given by

$$f(x_1,\cdots,x_n) = n!\prod_{i=1}^{n} f(x_i) \quad (9)$$

where $-\infty < x_1 < \cdots < x_n < \infty$ and $f(x_i)$ is the underlying density function. For the type 2 censoring with discontinued sampling, after obtaining the first r ordered observations, the joint density function of the order statistics is given by

$$f(x_{1,\cdots,r}) = \frac{n!}{(n-r)!}\left[1-F(x_r)\right]^{n-r}\prod_{i=1}^{r} f(x_i) \quad (10)$$

The type 2 progressive censoring is the major interest for reliability and life-testing analysis.

3.2 Statistical Criteria and Tests

Akaike's Information Criterion (AIC) and Schwarz's Information Criterion (SIC)

Let x_1, x_2, ..., x_n represent a random sample of n observations. Interest here lies in discriminating among nested and non-nested two- and three-parameter probability distributions in which the null hypothesis of interest is $H_0: \gamma = 0$ against the alternative $H_1: \gamma \neq 0$. Denoting the maximized values of the two- and three-parameter variants of a particular log-likelihood function as $\log L_0$ and $\log L_1$, respectively, the AIC and SIC may be expressed, respectively, as:

Choose the $\begin{Bmatrix}2\\3\end{Bmatrix}$ parameter distribution if

$$\textit{AIC}: \quad \log L_0 - 2\begin{Bmatrix}>\\<\end{Bmatrix}\log L_1 - 3 \quad (11)$$

$$\textit{SIC}: \quad \log L_0 - \log n\begin{Bmatrix}>\\<\end{Bmatrix}\log L_1 - 3\log n/2 \quad (12)$$

Generalised Information Criterion (GIC)

The generalised information criterion (GIC) procedure has been proposed for the discrimination of distributional structures among a set of alternatives by Bai et al.[1991]. The GIC is based on the equivalence between some well-known information criteria and hypothesis tests, and attempts to determine the false distributions based on sample information. Large differences between the maximised values of log-likelihood functions will lead to rejection of the distribution with the lower value. Discriminated distributions are separated into two categories, the superior and badly fitting categories. The distributions in the superior category perform within an acceptable tolerance level and there are no significant differences among their performances. The GIC procedure may provide several alternatives rather than one particular distribution for a particular set of data. In the event of there being several sets of data, the distribution with the highest probability of acceptance in the superior category will be chosen.

Let x_1, x_2, ..., x_n be n independently and identically distributed random observations. Denote $\log L_j$ as the maximised log-likelihood value of distribution j $(j = 1,2,\cdots,m)$, with the ordering given as

$$\log L_1 > \log L_2 > \cdots > \log L_m. \quad (13)$$

Then distribution j will be rejected in favour of distribution l if

$$GIC: \quad -2\log L_l + T_0 < -2\log L_j \tag{14}$$

where $T_0 > 0$. The value T_0 is the tolerance level required in order to reject distributions as being significantly different from each other, and is equivalent to the rejection region discussed in the previous section. For the nested case, T_0 can be expressed as

$$T_0 = c \tag{15}$$

where c is the critical value of the χ^2 distribution with one degree of freedom. In the non-nested case, T_0 is given by

$$T_0 = 2z^* \tag{16}$$

where $z^* > 0$, and the asymptotic upper bound on the significance level is given by the cumulative standard normal distribution function evaluated at $\Phi\left(-\sqrt{2z^*}\right)$.

The motivation behind the GIC procedure is straightforward. First, for a given sample, select the distribution with the highest maximised log-likelihood value among $\log L_j$ $(j = 1,2,\cdots,m)$. This distribution then belongs to the category of superior distributions and is also used as a standard for further inference. Second, reject the false alternatives among the remaining distributions in terms of the given tolerance. The distributions which perform within an acceptable tolerance level of the best fitting distribution are retained in the superior category, and the distribution with the highest probability of acceptance over different sets of data will be chosen from the superior category.

Loss Functions

In this study, loss functions recommended for assessing model performance have been chosen to establish the effect of discrimination criteria on the intended use of the model. These functions are the relative bias (*BIAS*) and the relative root mean square error (*RRMSE*), which are evaluated at the upper percentiles of the distributions. For an estimate q_i of a quantity of interest q, these loss functions are defined in terms of deviations from the observed value q in each data set. The definitions used for the loss functions are:

$$BIAS(q) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{q}_i - q}{q}\right) \tag{17}$$

$$RRMSE(q) = \left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\hat{q}_i - q}{q}\right)^2\right]^{0.5} \tag{18}$$

where N is the number of data sets. For present purposes, the quantity q denotes the upper percentiles of the sampling distributions.

Considering the loss functions above, two performance criteria defined in terms of the relative root mean square error (*RRMSE*) are also used. The first criterion is based on the mean of *RRMSE* in the upper (U) percentiles, and the second is based on the mean of *RRMSE* in the entire or full (F) percentiles of the distribution. For an estimate q_{ij} of a quantity of interest q, these performance criteria are defined in terms of deviations from q in the percentiles of interest, where i denotes the specific percentiles estimated and j corresponds to the specific data set fitted. The definitions of the upper percentile error and full percentile error are as follows:

$$UPE = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{1+(1-p)n_j}\sum_{i=pn_j}^{n_j}\left(\frac{\hat{q}_{ij} - q}{q}\right)^2\right]^{0.5} \tag{19}$$

$$FPE = \frac{1}{N}\sum_{j=1}^{N}\left[\frac{1}{n_j}\sum_{i=1}^{n_j}\left(\frac{\hat{q}_{ij} - q}{q}\right)^2\right]^{0.5} \tag{20}$$

where p is the location of the p-quantile, which is chosen at the 98 per cent level for *UPE* and 100 per cent for *FPE*. N is the number of data sets, and n_j is the sample size for data set j. For present purposes, q denotes the percentile quantities related to the upper and full percentile errors.

Hypothesis Tests

The performances of two well-known procedures for testing goodness of fit are also considered, namely the chi-square (CHI) test and Kolmogorov-Smirnov (KS) test. Classifying the n observations into κ categories, the chi-square statistic is of the form:

$$CHI = \sum_{i=1}^{\kappa}\frac{(f_i - np_i)^2}{np_i} \tag{21}$$

which has an asymptotic χ^2 distribution with $(\kappa - \ell - 1)$ degrees of freedom under H_0. The p_i are hypothetical probabilities, the f_i are empirical frequencies, and ℓ is the number of parameters estimated for each distribution. For fitting the real data in Section 4 below, $\kappa = 10$ and $\ell = 2$ or $\ell = 3$. The KS test, which is defined in terms of the maximum absolute difference between the sample

distribution function $S_n(x)$ and the hypothetical distribution function $F_0(x)$, is given by

$$D_n = \sup_x |S_n(x) - F_0(x)| \quad (22)$$

Large observed values of the D_n statistic lead to rejection of the hypothesis $F_0(x)$ (for further details, see Kendall and Stuart [1979]).

4. Application Examples: Fitting Water Quality Data

The water quality data used here comprise weekly observations on stream water chemical determinants from the Birkenes catchment in Norway. There are fourteen years of reasonably complete records available for five chemical species: sulphate, nitrate, sodium, calcium and hydrogen ions. The parameter distributions applied to characterise these data are the two- and three-parameter gamma (G2 and G3), lognormal (L2 and L3) and Weibull (W2 and W3). These distributions are generally capable of representing skewed environmental data (for further information, see Jakeman et al. [1991]).

4.1 Examination of Stationarity and Autocorrelation

To start with the analysis of the data, it is important to examine an underlying assumption when fitting probability distributions to observations, such as stationarity and low autocorrelation for observations. To obtain an indication of the adequacy of this assumption, three simple analyses were undertaken. First, the variability of the mean between the summer and winter seasons was examined for each annual period. Large differences for most years would imply a non-stationary seasonality for the variable, and hence the need to build a causal model with seasonal rather than annual predictive capabilities. Second, regressions against flow were constructed for the different variables for each year. This would indicate if short-term concentrations were flow related, and hence if they might be better modelled taking this variable into account. Third, autocorrelation functions were computed for each variable and for each year. A summary of the outcome of these exercises, using the median result, is given in Table 1.

4.2 Estimation of distributional properties

Tables 2 and 3 report the results of fitting by maximum likelihood the six distributional forms to the 14 annual sets of weekly observations of sulphate and sodium concentrations, respectively. In the case of sulphate, the L2 distribution fits the observed mean concentration with an accuracy of 0.9 per cent, on average, and fits the observed maximum of each annual data set with an average accuracy of 6.5 per cent. Only the W2 distribution yields unacceptable errors in fitting these quantities.

It is interesting to see that for nitrate, the two-parameter distributions fit the data better than the three-parameter distributions in terms of the observed mean and maximum error. Among the two-parameter distributions, W2 is the superior distribution, while L2 is unacceptable and G2 is only marginally better than L2. The average error in fitting the observed mean and maximum is close to 10 per cent.

For weekly sodium concentrations, only the W2 distribution yields poor fitting errors. The three-parameter distributions provide no substantial improvements over the more parsimonious alternatives. The G2 and L2 distributions fit the observed mean very closely and better than the others, while the observed maximum is fitted to an accuracy of less than 5 per cent, on average.

4.3 Discrimination among distributions

Tables 2 to 4 also show the results of applying the GIC, AIC and SIC statistics, and the chi-square (CHI) and K-S tests. All tests use the 98 per cent confidence level. Consider the sulphate results. The GIC selects L2 and W3 as not being significantly inferior to alternatives in 12 of the 14 cases. The other two information criteria (AIC and SIC) are not as definitive. The chi-square test, and especially the K-S test, are more definitive, preferring G2 and L2, followed by W3 and G3. L2 is certainly the most preferred or equally preferred by all criteria, except by AIC.

Comparing the results for nitrate concentrations, all the information criteria prefer the G3 distribution. The chi-square test prefers G2, while the K-S test finds all two-parameter distributions to be acceptable. The explanation for the different results can arise because the information criteria are based on the log-likelihood ratio while hypothesis tests use sampling frequencies. The two-parameter distributions may fit the observations better, but based on the log-likelihood ratio, G3 is superior to the others.

Results of applying the three information criteria and two asymptotic tests for sodium concentrations are quite consistent. The G2 distribution is most preferred, or is at least equally preferred, to L2 by four of the five discrimination methods, the chi-square test accepting G2 for 11 of the 14 cases, and L2 on 12 occasions. It also fitted the observed means most closely and, on average, fitted the observed maximum with the least error. However, the L2 distribution appears almost as acceptable as G2.

Figure. 1 illustrates the range of the fitting performance attained by the L2 distribution when estimation is by maximum likelihood on annual sets of weekly sulphate concentrations. The left-hand plot shows the worst fit to the observed maximum concentration obtained over the 14

data sets (in 1986), while the right-hand plot shows the best fit to this observation (in 1987). Figures. 2 and 3 repeat this illustration of the range of error in fitting the observed maximum for nitrate and sodium. In this case, the preferred G2 distribution is used, as was discussed previously.

5. Further Developments

As with most computer software programs, this statistical program is undergoing continual development and revision. At the current stage, the probability distributions included in this program are the normal, exponential, two- and three-parameter gamma, Weibull and lognormal distributions. In later versions, other alternatives, such as the logistic, Cauchy and Pareto distributions, will be considered and included. To improve the applicability of this program, it is planned to include multivariable distribution analysis and joint distribution analysis. Future versions of this program will also facilitate discrete sampling analysis, which has also been used extensively.

6. Conclusion

A computer program has been developed to undertake statistical analysis and modelling, which offers the advantages of being easy to use, flexible and incorporates state-of-the-art algorithms and numerical techniques. The program has been successfully applied to a large range of data sets, as illustrated in the application.

References

Kendall, M.G. and A. Stuart, Advanced Theory of Statistics, Vol. 2(4), 1979.

Bai, J., A..J. Jakeman and M. McAleer, A new approach to maximum likelihood estimation of the three-parameter gamma and Weibull distributions, Australian Journal of Statistics, 33, 397-410, 1991.

Bai, J., A.J. Jakeman and M. McAleer, Model selection for environmental data, Environmetrics, 1, 211-254.1991.

Bai, J., A.J. Jakeman and M. McAleer, Estimation and discrimination of alternative air pollution models, Ecological Modeling, 64, 89-124, 1992.

Bai, J., A.J. Jakeman and M. McAleer, On the use of extreme value distributions for predicting the upper percentiles of environmental data, Mathematics and Computers in Simulation, 33, 483-488, 1992.

Gumbel, E.J., Statistics of Extremes, Columbia University Press, New York, 1958.

Galambos, J., The Asymptotic Theory of Extreme Order Statistics, John Wiley, New York, 1978.

Johnson, N.L., S. Kotz and N. Balakrishnan, Continuous Univariate Distributions , volume 1, John Wiley & Sons, New York, 1994.

Jakeman, A.J., P.G. Whitehead, A. Robson, J.A. Taylor and J. Bai, 'Investigation of a new approach to predict water quality extremes with a case study of chemical determinants in stream water, Water Science and Technology, 24 (6), 25-33, 1991.

TABLE 1 Summary of Investigation of IID Assumptions on Stream Acidity Variables at Birkenes

Species	Median value over 14 years		
	Relative difference in summer Vs winter mean	R^2 correlation with flow*	First lag autocorrelation
sulphate	0.08	0.19	0.66
nitrate	0.29	0.09	0.55
sodium	0.12	0.06	0.65
calcium	0.17	0.45	0.64
hydrogen	0.23	0.57	0.51

*correlation is result of regression of log species concentration on log flow, which is generally higher than that of concentration on flow.

TABLE 2 Discrimination Statistics in Fitting 14 Annual Sets of Samples of Weekly Sulphate Concentrations

Distributional	Errors in Observed Mean			Errors in Observed Maximum			Number of Acceptances				
Model	AVG	SDV	MAX	AVG	SDV	MAX	GIC	AIC	SIC	CHI	K-S
G3	0.017	0.017	0.070	0.079	0.050	0.183	10	1	1	8	10
G2	0.006	0.002	0.008	0.064	0.046	0.178	9	4	4	9	14
L3	0.015	0.010	0.041	0.074	0.039	0.140	7	1	1	7	8
L2*	0.009	0.003	0.014	0.065	0.048	0.176	12	3	5	9	14
W3	0.024	0.033	0.137	0.063	0.042	0.139	12	3	1	8	13
W2	0.361	0.475	0.999	X	X	X	2	2	2	5	7

X = large error
* = preferred distribution
AVG = average of rrmse over 14 data sets
SDV = standard deviation
MAX = maximum of rrmse values

GIC = Generalised Information Criterion
AIC = Akaike Information Criterion
SIC = Schwarz Information Criterion
CHI = Chi square test
K-S = Kolmogorov-Smirnov test

TABLE 3 Discrimination Statistics in Fitting 14 Annual Sets of Samples of Weekly Nitrate Concentrations

Distributional	Errors in Observed Mean			Errors in Observed Maximum			Number of Acceptances				
Model	AVG	SDV	MAX	AVG	SDV	MAX	GIC	AIC	SIC	CHI	K-S
G3	0.252	0.206	0.641	0.364	0.378	1.363	10	6	6	5	6
G2	0.130	0.066	0.263	0.116	0.089	0.312	8	4	4	12	14
L3	0.165	0.139	0.548	77.497	278.706	1082.385	0	0	0	1	1
L2	0.192	0.096	0.380	0.493	0.494	2.021	5	0	0	8	13
W3	0.205	0.187	0.594	0.696	1.094	3.041	9	0	0	8	9
W2'*	0.099	0.074	0.265	0.102	0.074	0.288	8	4	4	10	14

TABLE 4 Discrimination Statistics in Fitting 14 Annual Sets of Samples of Weekly Sodium Concentrations

Distributional	Errors in Observed Mean			Errors in Observed Maximum			Number of Acceptances				
Model	AVG	SDV	MAX	AVG	SDV	MAX	GIC	AIC	SIC	CHI	K-S
G3	0.013	0.011	0.038	0.056	0.063	0.245	8	1	1	7	8
G2*	0.006	0.004	0.013	0.047	0.067	0.284	12	6	7	11	14
L3	0.012	0.010	0.042	0.049	0.053	0.226	7	1	1	7	7
L2	0.008	0.006	0.020	0.050	0.068	0.284	12	3	5	12	14
W3	0.030	0.042	0.133	0.057	0.059	0.230	11	3	0	10	12
W2	0.576	0.489	1.000	X	X	X	3	0	0	4	5

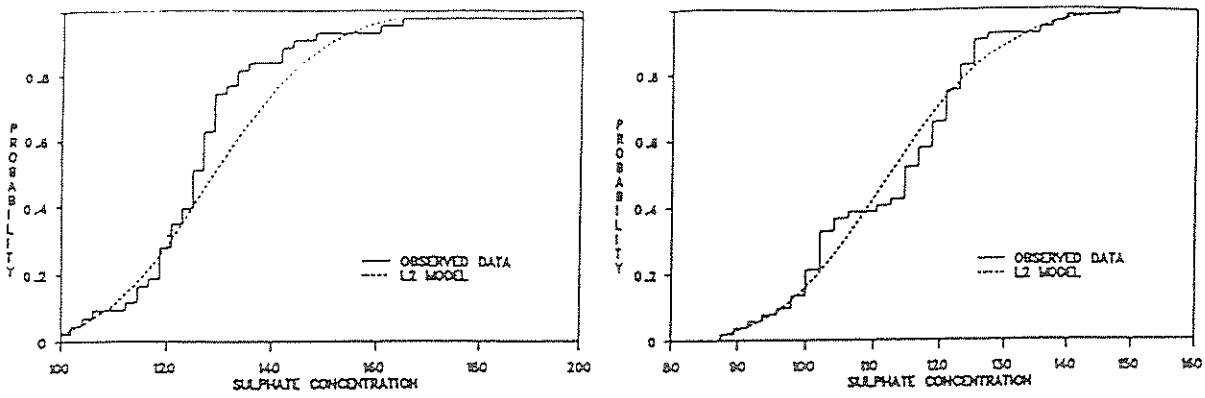

Fig. 1. Fit of the L2 distribution to weekly sulphate concentrations in 1986 and 1987.

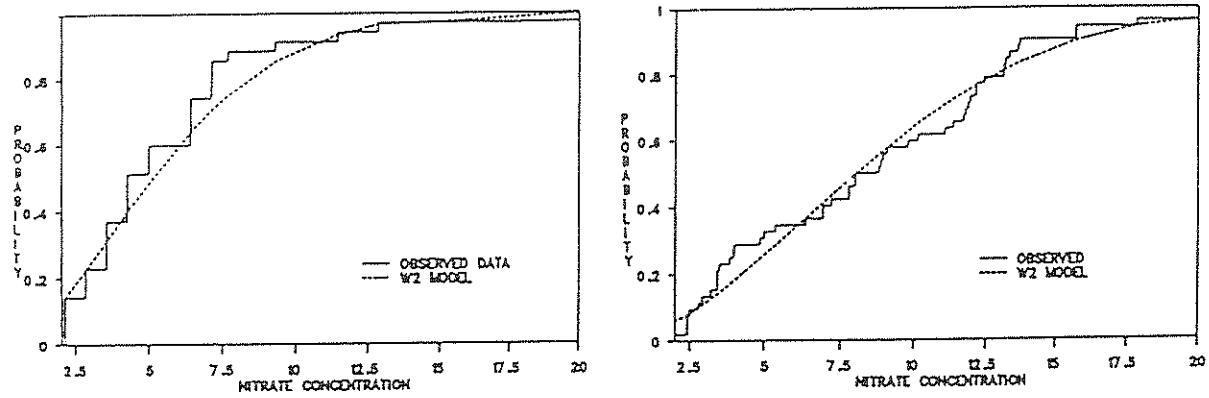

Fig. 2. Fit of the W2 distribution to weekly nitrate concentrations in 1980 and 1987.

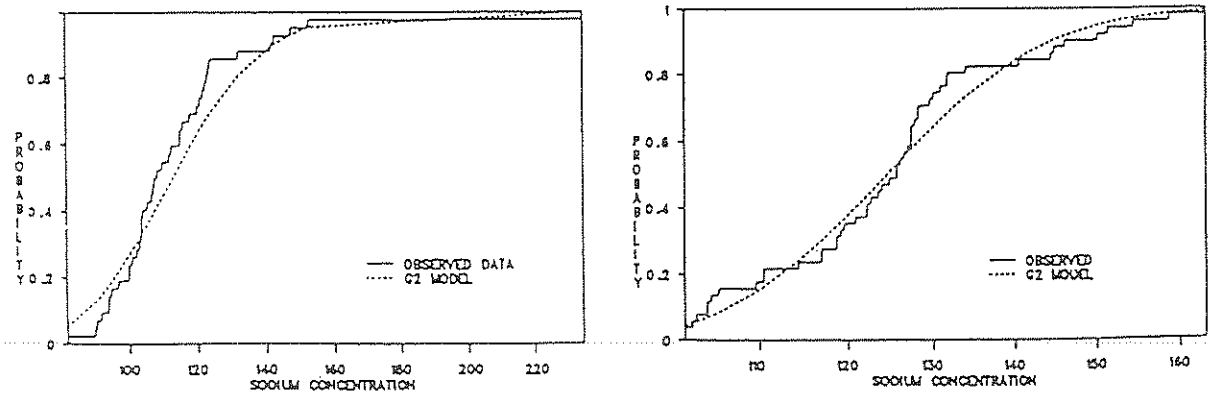

Fig. 3. Fit of the G2 distribution to weekly sodium concentrations in 1988 and 1983.