United States
Environmental Protection
Agency

Environmental Monitoring and Support
Laboratory
Research Triangle Park NC 27711

EPA-600/4-79-044
July 1979

Research and Development

The Maximum Likelihood Approach to Probabilistic Modeling of Air Quality Data

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into nine series. These nine broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The nine series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies
6. Scientific and Technical Assessment Reports (STAR)
7. Interagency Energy-Environment Research and Development
8. "Special" Reports
9. Miscellaneous Reports

This report has been assigned to the ENVIRONMENTAL MONITORING series. This series describes research conducted to develop new or improved methods and instrumentation for the identification and quantification of environmental pollutants at the lowest conceivably significant concentrations. It also includes studies to determine the ambient concentrations of pollutants in the environment and/or the variance of pollutants as a function of time or meteorological factors.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

THE MAXIMUM LIKELIHOOD APPROACH TO PROBABILISTIC MODELING OF AIR QUALITY DATA

by

Terence Fitz-Simons and David M. Holland
Environmental Monitoring and Research Laboratory
Environmental Protection Agency
Research Triangle Park, N. C. 27711

ENVIRONMENTAL MONITORING AND SUPPORT LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, N. C. 27711

DISCLAIMER

This report has been reviewed by the Environmental Monitoring and Support Laboratory, U.S. Environmental Protection Agency, and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

FOREWORD

Measurement and monitoring research efforts are designed to anticipate potential environmental problems, to support regulatory actions by developing an in-depth understanding of the nature and processes that impact health and the ecology, to provide innovative means of monitoring compliance with regulations and to evaluate the effectiveness of health and environmental protection efforts through the monitoring of long-term trends. The Environmental Monitoring and Support Laboratory, Research Triangle Park, North Carolina, is responsible for development of: environmental monitoring technology and systems; agency-wide quality assurance programs for air pollution measurement systems; and technical support to the Agency's operating function including the Office of Air, Noise and Radiation, the Office of Toxic Substances and the Office of Enforcement.

In order for the laboratory to effectively analyze the data generated by these activities, statistical assumptions must be made regarding the underlying distribution of these data. This report presents a tool to aid in evaluating these assumptions. Any changes in the material presented in this report will be presented in future reports.

Thomas R. Hauser
Director
Environmental Monitoring and
Support Laboratory

ABSTRACT

Software developed by the authors using maximum likelihood estimation to fit six probabilistic models is presented. The software is designed as a tool for the air pollution researcher to determine what assumptions are valid in the statistical analysis of air pollution data for the purposes of standard setting, roll-back calculations, estimation of maximum concentrations, threshold approximations, and handling missing observations. The program fits user's data to the normal distribution, the 3-parameter lognormal distribution, the 3-parameter Weibull distribution, the 3-parameter gamma distribution, the Johnson S_B distribution (a 4-parameter lognormal distribution), and the 4-parameter beta distribution. The parameters are estimated using standard closed solutions to maximizing equations, Gaussian elimination to solve non-linear maximizing equations where possible, and a golden section search for all other parameters. Graphical output contains a histogram of the data superimposed by the fitted density for each model. Six goodness-of-fit criteria are supplied and ranked by the program to aid in the selection of the most appropriate choice among the six models. These criteria are absolute deviations (AD statistic), weighted absolute deviations (WAD statistic), Kolmogorov-Smirnov statistic, Cramer-von Mises-Smirnov statistic, the log-likelihood function, and the observed significance level of the Chi-square goodness-of-fit test. The results of applying the program to several subsets of the Los Angeles Catalyst Study data base are presented.

CONTENTS

Foreword iii
Abstract iv
Figures vi

1. Introduction 1
2. Methodology 2
3. Comparing Distributions 4
4. Application of MAXFIT 7
5. Future Developments 25
6. Summary 27

References 28
Appendix A 30
Appendix B 33

FIGURES

Number		Page
1	MAXFIT Output: Fit of normal distribution for data in Example 1	10
2	MAXFIT Output: Fit of 3-parameter lognormal distribution for data in Example 1	11
3	MAXFIT Output: Fit of Gamma distribution for data in Example 1	12
4	MAXFIT Output: Fit of Weibull distribution for data in Example 1	13
5	MAXFIT Output: Fit of Johnson S_b distribution for data in Example 1	14
6	MAXFIT Output: Fit of Beta distribution for data in Example 1	15
7	MAXFIT Output: Fit of the normal distribution for data in Example 2	18
8	MAXFIT Output: Fit of the 3-parameter lognormal distribution for data in Example 2	19
9	MAXFIT Output: Fit of the Gamma distribution for data in Example 2	20
10	MAXFIT Output: Fit of the Weibull distribution for data in Example 2	21
11	MAXFT Output: Fit of the Johnson S_B distribution for data in Example 2	22
12	MAXFIT Output: Fit of the Beta distribution for data in Example 2	23

TABLES

Number		Page
1	Available Distributions and Identifying Numbers	8
2	MAXFIT Output: Comparison Statistics for Data in Example 1	16
3	MAXFIT Output: Comparison Statistics for Data in Example 2	24

SECTION 1

INTRODUCTION

There is a growing interest in the application of probability models to air quality data in areas such as:

- standard setting,
- emission roll-back calculations,
- estimation of maximum concentrations,
- threshold approximations, and
- handling missing observations.

Selection of a valid probability model to describe air quality data is a difficult problem involving many factors. Larsen (1,2,3,4,5) proposed the two-parameter and three-parameter univariate log-normal models for describing air quality data collected in urban areas. Mage and Ott (6,7) introduced censorship to the three-parameter lognormal probability model whenever the third parameter was negative. These authors concluded that this model was a superior, general purpose model suitable for a large variety of environmental phenomena. Standard logarithmic probability paper was used to fit their proposed probability model. Curran and Frank (8) used semi-log graph paper to plot 1-F(x) vs. pollutant concentration as a technique for examining the behavior of the upper tails of air quality data. Curran and Frank pointed out that fitting techniques based upon two selected percentiles may be sensitive to the choice of an underlying distribution, and that the stability of these estimates for varying choices of percentiles needs investigation.

Computer software providing efficient estimation of parameters and evaluation of goodness-of-fit for several probability modes for ambient air quality data has not been available. A computer program, MAXFIT, is presented here for the purpose of identifying a suitable probability model to describe ambient air quality data using maximum likelihood techniques.

SECTION 2

METHODOLOGY

The approach employed is maximum likelihood estimation. This is a simple method, yielding estimates for almost every model imaginable, and its estimators have several statistically desirable properties. The various distribution models are compared by goodness-of-fit statistics providing an objective basis for choosing one of the six models for a particular application.

The main idea behind maximum likelihood estimation may be stated very simply. Given a random sample from a population distributed according to a density $f(x)$ with parameters α, β, and λ ($f[x;\alpha,\beta,\lambda]$), the maximum likelihood estimates ($\hat{\alpha}$,$\hat{\beta}$, and $\hat{\lambda}$) are found such that the likelihood that the random sample came from $f(x;\hat{\alpha},\hat{\beta},\hat{\lambda})$ is greater than or equal to the likelihood that the sample came from $f(x;\tilde{\alpha},\tilde{\beta},\tilde{\lambda})$; where $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\lambda}$ are any other estimates of α, β and λ.

Under certain assumptions discussed in Rao (9), maximum likelihood estimators have statistically desirable characteristics. First, they are consistent. Consistent estimators become "better" as the sample size increases. If biased, the bias decreases as the sample size gets larger. The variance of the estimators lessens as the sample size increases. Second, they converge to the true value of the parameters with a probability equal to one. If the entire population were included, the estimators would equal the parameters with a probability equal to one. Third, as n (sample size) approaches infinity, sampling distribution of the estimators approaches normal distribution.

The method of fitting using maximum likelihood is described below:

- The likelihood function is defined as the joint density of the random sample.

- The logarithm of the likelihood function is determined. This facilitates differentiation techniques and since the logarithm of X increases whenever X does, the likelihood and the log-likelihood reach a maximum at the same point.

- Differentiation techniques are employed to maximize the log-likelihood. If these techniques fail to yield equations which can be solved algebraically, the correct values of the parameters must be found by an iterative search. This usually requires a computer.

FORTRAN Software utilizing these concepts was developed based on a program written by Schreuder, et al (10) making use of several IMSL (International Mathematical and Statistical Library, 1975) subroutines. While modifying this program to run on the EPA UNIVAC 1110 situated at Research Triangle Park, North Carolina, several new capabilities were incorporated. These include the estimation of the lower boundary parameter in the 3-parameter lognormal (Johnson S_L), gamma, Weibull, the 4-parameter beta, and lognormal (Johnson S_B) distributions. Boundary parameters are estimated by a golden section search (11). The golden section search is a fairly efficient technique, accomplishing a 38 percent reduction in the interval of uncertainty for each iteration and requiring one evaluation of the function per iteration. Graphic output was also included to provide a visual display of the fit.

A list of the distributions that were fitted and the estimates used to fit them is shown in Appendix A. All summations are from 1 to the number of observations in the data set (n).

SECTION 3

COMPARING DISTRIBUTIONS

The program MAXFIT supplies the user with six goodness-of-fit statistics as a basis for comparing fitted distributions. The statistics supplied are: absolute deviations, weighted absolute deviations, the observed significance level of the Chi-square (χ^2) test, Kolmogorov-Smirnov statistic, Cramer-von Mises-Smirnov statistic, and the maximum value of the log-likelihood function. All these statistics are computed upon grouped data. However, the log-likelihood function is also computed using ungrouped data when the original data are available.

It should be noted that model validity tests, are designed to reject the null hypothesis that the data follow a certain distribution as few times as possible when the hypothesis is true. Therefore, these statistics are designed to indicate with some certainty when data do not follow the distribution stated in the null hypothesis. However, the literature discloses no other method to reveal how closely the data fit a distribution. The six various statistics are defined and discussed briefly below.

ABSOLUTE DEVIATIONS (AD)

This statistic puts less emphasis on large deviations between expected and observed frequencies than does the classic χ^2 statistic.

$$AD = \sum_{i}^{N} |P_{oi} - P_{ei}| \cdot n$$

where P_{oi} = observed proportion of data falling in the ith interval

P_{ei} = expected proportion of data falling in the ith interval

n = total number of observations in the data set

N = total number of intervals.

WEIGHTED ABSOLUTE DEVIATIONS (WAD)

The WAD statistic puts less emphasis upon the fit in the tails of the distribution.

$$WAD = \sum_{i}^{N} |P_{oi} - P_{ei}| \cdot n \cdot P_{ei}$$

CHI-SQUARE STATISTIC (χ^2)

The χ^2 statistic is very popular and widely used; however, it has poor power characteristics and is considered unreliable unless intervals have expected frequencies equal to or greater than five.

$$\chi^2 = \frac{\sum_i^N (P_{oi}-P_{ei})^2 \cdot n}{P_{ei}}$$

KOLMOGOROV-SMIRNOV STATISTIC (d)

The Kolmogorov-Smirnov statistic is an empirical distribution function (EDF) statistic. EDF is the observed cumulative distribution function. The statistic is the maximum distance between the observed and the expected cumulative frequency distribution.

$$d = \max|S_{n_i} - F_{n_i}|$$

where S_{n_i} = observed cumulative probability at the ith observed value

F_{n_i} = expected cumulative probability at the ith observed value

CRAMER-VON MISES-SMIRNOV STATISTIC (W^2)

The Cramer-von Mises-Smirnov statistic is the sum of squares of the distance between observed cumulative distribution and the expected cumulative distribution function weighted by the expected probabilities. This test puts less emphasis upon the tails of the distribution when measuring the fit.

$$W^2 = \sum_i^N (S_{n_i} - F_{n_i})^2 \cdot P_{ei} \cdot n$$

LIKELIHOOD

The likelihood is also presented as a goodness-of-fit statistic. If we are committed to fitting, using the maximum likelihood criteria, it is consistent to use the larger likelihood functions as a criterion for choosing between two or more distributions having the same number of parameters. Such a comparison is also the basis of the Neyman-Pearson theorem (9) dealing with the construction of a powerful test of a simple hypothesis involving two distributions.

The formulas of the log-likelihood functions of the various distributions are shown as follows where the superscript ^ represents the MLE estimates of the parameters and the summations are from 1 to n.

- Normal distribution:

$$\frac{-n}{2}\left(1+\log[2\pi]+\log[\hat{\sigma}^2]\right)$$

- Johnson S_L distribution:

$$\frac{-n}{2}\left(1+\log[2\pi]+\log[\hat{\sigma}^2]+2\hat{\mu}\right)$$

- Johnson S_B distribution:

$$-n(1/2-\log[\{\hat{\xi}-\hat{\lambda}\}/\hat{\sigma}]+\log[2\pi]/2+\Sigma\log[\hat{\xi}-x_i]/n+\Sigma\log[x_i-\hat{\lambda}]/n)$$

- Gamma distribution:

$$-n(\hat{\alpha}\log[\hat{\beta}]+\log[\Gamma\{\hat{\alpha}\}]-[\hat{\alpha}-1]\Sigma\log[x_i-\hat{\lambda}]/n+\frac{\Sigma[x_i-\hat{\lambda}]}{n\hat{\beta}})$$

- Weibull distribution:

$$-n(1-[\hat{\alpha}-1]\frac{\Sigma\log[x_i-\hat{\lambda}]}{n}+[\hat{\alpha}-1]\log[\hat{\beta}]-\log[\hat{\alpha}/\hat{\beta}])$$

- Beta distribution:

$$n(\log[\Gamma\{\hat{\alpha}+\hat{\beta}\}/\{\Gamma(\hat{\alpha})\Gamma(\hat{\beta})\}]-[\hat{\alpha}+\hat{\beta}-1]\log[\hat{\beta}-\hat{\lambda}]$$
$$+[\hat{\alpha}-1]\frac{\Sigma\log[x_i-\hat{\lambda}]}{n}+[\hat{\beta}-1]\frac{\Sigma\log[\hat{\xi}-x_i]}{n})$$

SECTION 4

APPLICATION OF MAXFIT

INPUTS

Subroutine USER must be supplied by the user to pass certain control variables to MAXFIT. An example of this subroutine is given in Appendix B.

The subroutine argument list must be as follows:

SUBROUTINE USER (X,P,XB,NGRP,TF,N,NC,BEGIN,STEP,TOL,TITLE,XMIN,XMAX, NDIST).

The variables that must be provided by the user are defined to be:

X = array of length m $\leq$2000 of class midpoints if the data are grouped; into classes or individual observations if the data are not grouped.

P = an array of length n $\leq$100 of class probabilities for data grouped into classes. This array is not defined for ungrouped data.

XB = an array of length p = n+1, consisting of the lower limits of each class with the last element being the upper limit of the last class. It is only needed if NGRP=1. Two additional classes will be created by MAXFIT to cover the entire range of the distribution being fitted.

NGRP = 1 if data are grouped, = 0 if data are individual observations. This variable indicates whether the statistics calculated by MAXFIT are to be based on individual observations or on class frequencies.

NDIST = an array of length 6, containing the identifying numbers for the desired distributions.

NC = number of classes if class limits, XB, are specified by the user. These limits must be specified if NGRP=1. If NC=0, MAXFIT calculates class limits based on value of BEGIN and STEP. The last class calculated by MAXFIT contains the maximum observed data point.

BEGIN = The lower boundary of the first class for which predictions are desired. Default value is slightly less than the observed data minimum value.

STEP = The class width. Default value is the (data range)/12.

TOL = primary cutoff point used in the searching subroutines. The Optional, default value is .0001.

TITLE = an array of length 10 which contains a 60-character title in elements of length 6. Default value is blanks.

N = number of classes for which the user wants predictions if NGRP=1, or the number of observations if NGRP=0.

If the user specifies the class limits, he gets predicted frequencies for the range of classes he specifies and the variable N should indicate the number of classes covered by that range. If the user does not specify classes, the program creates class limits based on the values of BEGIN and STEP. The variable TOL is used as the measure of precision desired in the maximum likelihood estimation processes where interaction is required. The available distributions and their identifying numbers are in Table 1.

TABLE 1. AVAILABLE DISTRIBUTIONS AND IDENTIFYING NUMBERS

Distribution Name	Distribution Number
Normal	1
Lognormal	2
Gamma	3
Weibull	4
S_B	5
Beta	6

The program provides the following output for each data set:

1. Observed minimum and maximum data value, number of observations, mean, variance, and standard deviation, index of skewness ($\sqrt{\beta_1}$), skewness squared (β_1) and index of kurtosis (β_2).

2. For each distribution fitted:

   - Estimated parameters for the distribution (notations and descriptions of the distributions are given in Appendix A).

   - By classes of the variable, the observed and predicted probabilities, the difference between them (residual), the cumulative observed and predicted probabilities, and the observed and predicted frequencies and their difference (residual frequency).

   - A CALCOMP plot with the density of the fitted distribution superimposed upon the histogram of the data.

3. A comparison of the distributions and their ranking from best (1) to worst (6) in terms of absolute deviations, weighted absolute deviations, Chi-square, Kolmogorov-Smirnov, Cramer-von Mises-Smirnov, and log likelihood statistics.

These statistics are ranked not as goodness-of-fit statistics, but as absolute measurements of fit. They may be used to calculate valid goodness-of-fit statistics when degrees of freedom are taken into account. The value of the observed significance level is printed out for the Chi-square goodness-of-fit test defined as the probability that a Chi-square random variable will be greater than the observed value. The observed value is calculated regrouping classes to insure that the predicted frequency is greater than 5. If N be the number of classes used to compute the statistic, the degrees of freedom for each model are as follows:

normal	N-3
3-parameter lognormal	N-4
3-parameter Gamma	N-4
3-parameter Weibull	N-4
Johnson S_B	N-5
4-parameter Beta	N-5.

EXAMPLES OF OUTPUT

Description of Data Sets

A short-term study was conducted by the Environmental Monitoring and Support Laboratory (EMSL) at the Los Angeles Catalyst Study (LACS) site from 7/20/78 to 8/30/78. Their purpose was to quantify the influence of NO_x emissions from automobiles on ozone (O_3) and nitrogen dioxide (NO_2) measurements made in close proximity to the freeway (12). The data base was stratified into low, medium and high O_3 categories by the background O_3 level at the upwind site.

Oxides of nitrogen (NO_x) and NO_2 measurements taken at the upwind LACS site 3 were used to fit the probability models. The ranges of the two data sets are equivalent due to the low nitrogen oxide (NO) measurements at the upwind site. However, the results of fitting probability models to these two data sets are very different.

The first example is NO_2 measurements recorded at site 3. The 92 observations range from 0.007 to 0.059 ppm. The output of MAXFIT is shown in Figures 1-6 and Table 2. The Weibull distribution appears to be the "best" model for this data. It has the highest log-likelihood value, the highest observed significance level of the Chi-square test and does fairly well when compared to other models by the remaining fitting criteria. The Weibull is also a useful model since data can be transformed to an exponential distribution and analyses may be performed assuming an exponential distribution.

The second example is NO_x measurements recorded at site 3. Again, the 92 observations range from 0.007 to 0.059 ppm. The output of MAXFIT for this

LACS SITE 3 NO2 HIGH 03 LEVEL

OBS.MIN.X = .007000
OBS.MAX.X = .059000
NO. OF OBS. = 92.0000

MEAN = .0246
VARIANCE = .0001
STANDARD DEV. = .0112

INDEX OF SKEWNESS = .7315
INDEX OF KURTOSIS = 3.3555
SKEWNESS SQUARED = .5351

NORMAL DISTRIBUTION
MEAN = .0246
VARIANCE = .0001
STANDARD DEVIATION = .0112

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
– INFINITY –	.007	.000000	.056922	–.056922	.000000	.056922	.000	5.237	–5.237
.007 –	.011	.086957	.059563	.027394	.086957	.116485	8.000	5.480	2.520
.011–	.016	.119565	.094116	.025449	.206522	.210601	11.000	8.659	2.341
.016–	.020	.195652	.128129	.067523	.402174	.338730	18.000	11.788	6.212
.020–	.024	.173913	.150290	.023623	.576087	.489020	16.000	13.827	2.173
.024–	.029	.054348	.151885	–.097538	.630435	.640905	5.000	13.973	– 8.973
.029–	.033	.130435	.132252	–.001817	.760870	.773158	12.000	12.167	– .167
.033–	.037	.130435	.099218	.031217	.891304	.872375	12.000	9.128	2.872
.037–	.042	.043478	.064132	–.020654	.934783	.936507	4.000	5.900	– 1.900
.042–	.046	.021739	.035715	–.013976	.956522	.972222	2.000	3.286	– 1.286
.046–	.050	.010870	.017136	–.006266	.967391	.989358	1.000	1.577	– .577
.050–	.055	.010870	.007083	.003786	.978261	.996441	1.000	.652	.348
.055–	.059	.010870	.002523	.008347	.989130	.998964	1.000	.232	.768
.059–	.063	.010870	.000774	.010096	1.000000	.999738	1.000	.071	.929
.063–INFINITY		.000000	.000262	.000262	1.000000	1.000000	.000	.024	–.024

LACS SITE 3 NO2 HIGH OZ LEVEL
NORMAL DISTRIBUTION

MU = 0.0246
SIGMA^2 = 0.0001

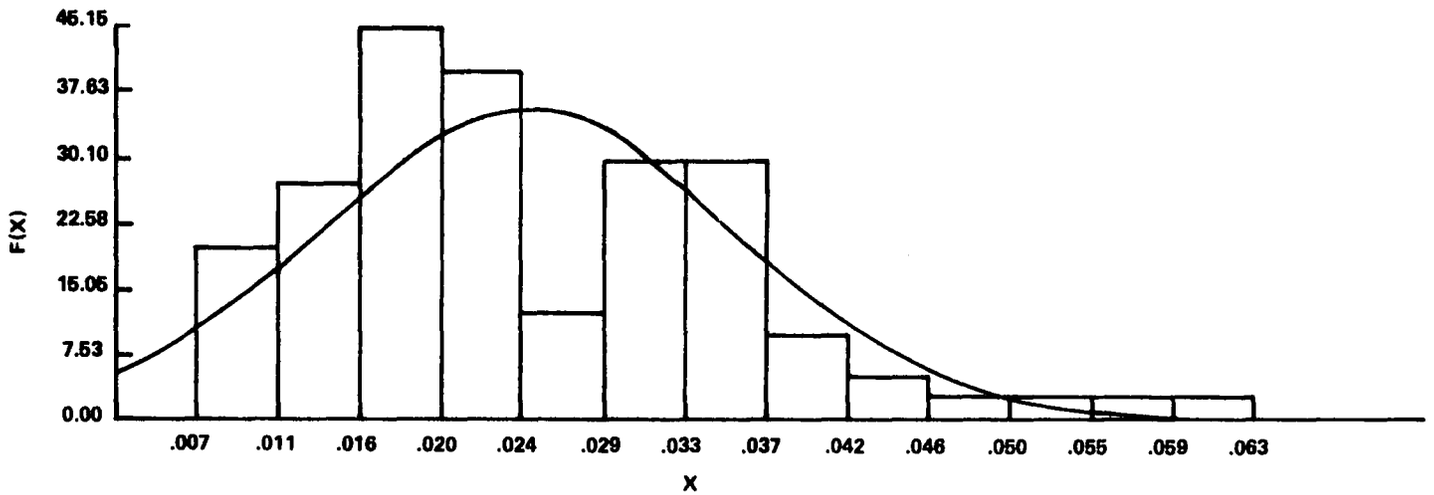

FIGURE 1 MAXFIT OUTPUT: FIT OF NORMAL DISTRIBUTION FOR DATA IN EXAMPLE 2

LACS SITE 3 NO2 HIGH 03 LEVEL

LOG NORMAL DISTRIBUTION

MEAN = −3.5020
VARIANCE = .1213
STANDARD DEVIATION = .3483
SEARCHED LOWER BOUND = −.007367

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
−.007−	.007	.000000	.016618	−.016618	.000000	.016618	.000	1.529	−1.529
.007−	.011	.086957	.068441	.018516	.086957	.085059	8.000	6.297	1.703
.011−	.016	.119565	.134676	−.015111	.206522	.219735	11.000	12.390	−1.390
.016−	.020	.195652	.170804	.024848	.402174	.390539	18.000	15.714	2.286
.020−	.024	.173913	.166827	.007086	.576087	.557366	16.000	15.348	.652
.024−	.029	.054348	.138416	−.084068	.630435	.695782	5.000	12.734	−7.734
.029−	.033	.130435	.103334	.027101	.760870	.799116	12.000	9.507	2.493
.033−	.037	.130435	.071926	.058509	.891304	.871042	12.000	6.617	5.383
.037−	.042	.043478	.047761	−.004283	.934783	.918803	4.000	4.394	−.394
.042−	.046	.021739	.030723	−.008984	.956522	.949526	2.000	2.827	−.827
.046−	.050	.010870	.019347	−.008478	.967391	.968874	1.000	1.780	−.780
.050−	.055	.010870	.012015	−.001145	.978261	.980889	1.000	1.105	−.105
.055−	.059	.010870	.007397	.003472	.989130	.988286	1.000	.681	.319
.059−	.063	.010870	.004532	.006338	1.000000	.992818	1.000	.417	.583
.063−INFINITY		.000000	.007182	−.007182	1.000000	1.000000	.000	.661	−.661

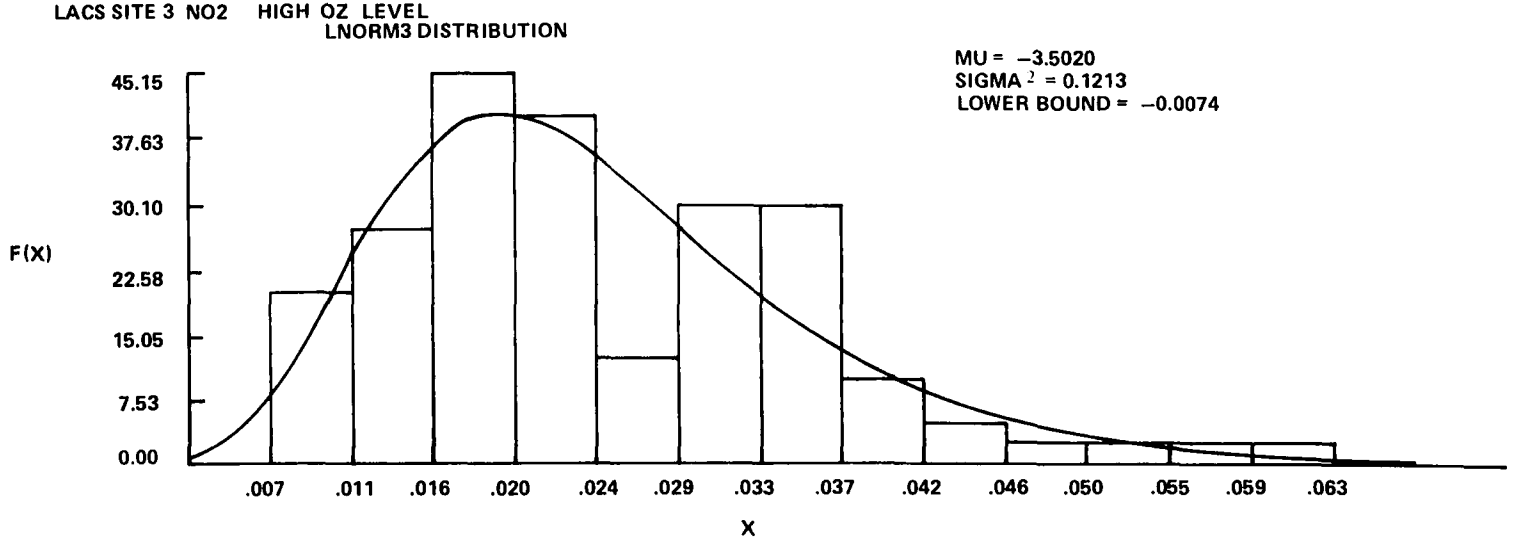

FIGURE 2 MAXFIT OUTPUT: FIT OF 3–PARAMETER LOG NORMAL DISTRIBUTION FOR DATA IN EXAMPLE 1

LACS SITE 3 NO2 HIGH 03 LEVEL

GAMMA DISTRIBUTION

ALPHA= 3.1854
BETA = .0066
SEARCHED LOWER BOUND = .003606

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.004–	.007	.000000	.010529	–.010529	.000000	.010529	.000	.969	–.969
.007–	.011	.086957	.079766	.007190	.086957	.090295	8.000	7.338	.662
.011–	.016	.119565	.145624	– .026059	.206522	.235919	11.000	13.397	–2.397
.016	.020	.195652	.168444	.027208	.402174	.404363	18.000	15.497	2.503
.020–	.024	.173913	.157326	.016588	.576087	.561688	16.000	14.474	1.526
.024–	.029	.054348	.129842	– .075494	.630435	.691530	5.000	11.945	–6.945
.029–	.033	.130435	.098862	.031573	.760870	.790393	12.000	9.095	2.905
.033–	.037	.130435	.071127	.059308	.891304	.861520	12.000	6.544	5.456
.037–	.042	.043478	.049068	– .005590	.934783	.910588	4.000	4.514	– .514
.042–	.046	.021739	.032771	– .011032	.956522	.943359	2.000	3.015	–1.015
.046–	.050	.010870	.021331	– .010461	.967391	.964690	1.000	1.962	– .962
.050–	.055	.010870	.013597	– .002727	.978261	.978286	1.000	1.251	– .251
.055–	.059	.010870	.008518	.002352	.989130	.986804	1.000	.784	.216
.059–	.063	.010870	.005258	.005611	1.000000	.992062	1.000	.484	.516
.063–	INFINITY	.000000	.007938	–.007938	1.000000	1.000000	.000	.730	–.730

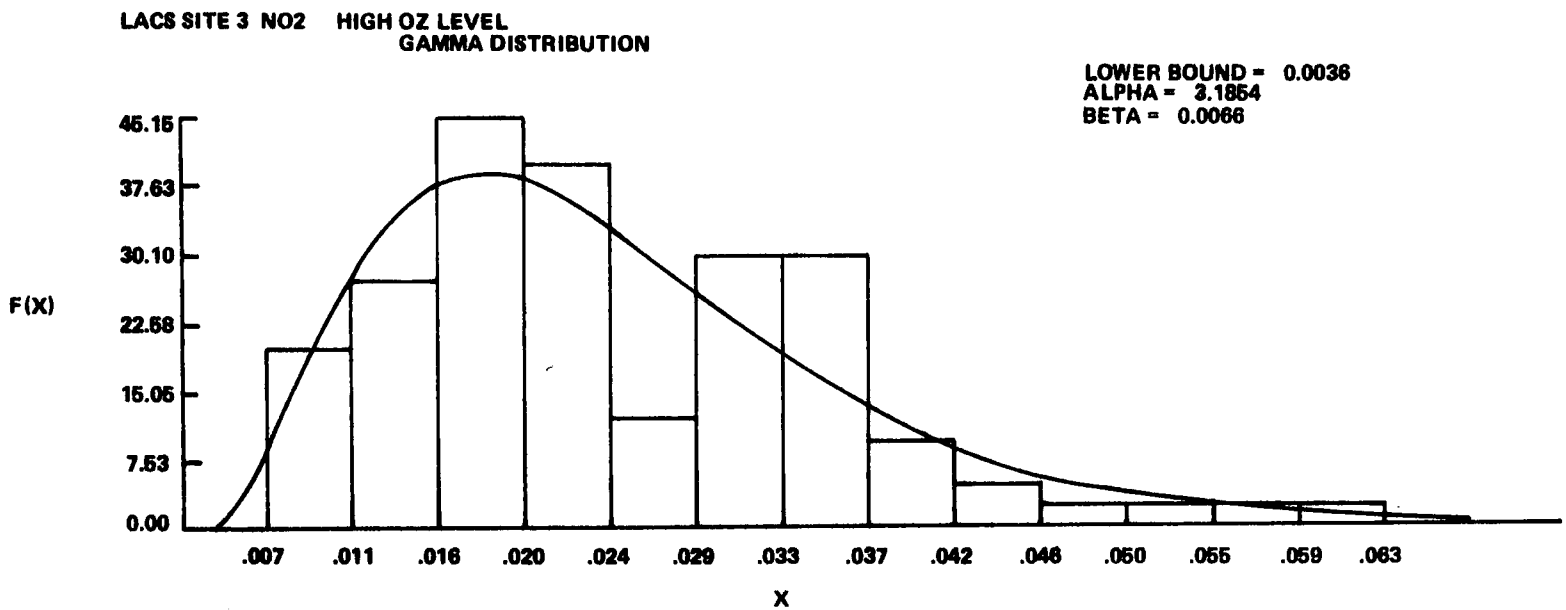

FIGURE 3 MAXFIT OUTPUT: FIT OF GAMMA DISTRIBUTION FOR DATA IN EXAMPLE 1

LACS SITE 3 NO2 HIGH 03 LEVEL

WEIBULL DISTRIBUTION

BETA = .0212
C = 1.7325
SEARCHED LOWER BOUND = .005694

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.006–	.007	.000000	.007848	–.007848	.000000	.007848	.000	.722	–.722
.007–	.011	.086957	.087676	–.000719	.086957	.095524	8.000	8.066	– .066
.011–	.016	.119565	.141050	–.021485	.206522	.236574	11.000	12.977	– 1.977
.016–	.020	.195652	.159711	.035941	.402174	.396285	18.000	14.693	3.307
.020–	.024	.173913	.153650	.020263	.576087	.549935	16.000	14.136	1.864
.024–	.029	.054348	.132469	–.078121	.630435	.682405	5.000	12.187	– 7.187
.029–	.033	.130435	.104811	.025624	.760870	.787216	12.000	9.643	2.357
.033–	.037	.130435	.077111	.053324	.891304	.864326	12.000	7.094	4.906
.037–	.042	.043478	.053188	–.009710	.934783	.917515	4.000	4.893	– .893
.042–	.046	.021739	.034591	–.012852	.956522	.952105	2.000	3.182	– 1.182
.046–	.050	.010870	.021298	–.010429	.967391	.973404	1.000	1.959	– .959
.050–	.055	.010870	.012454	–.001585	.978261	.985858	1.000	1.146	– .146
.055–	.059	.010870	.006934	.003936	.989130	.992792	1.000	.638	.362
.059–	.063	.010870	.003683	.007186	1.000000	.996475	1.000	.339	.661
.063–INFINITY		.000000	.003525	–.003525	1.000000	1.000000	.000	.324	–.324

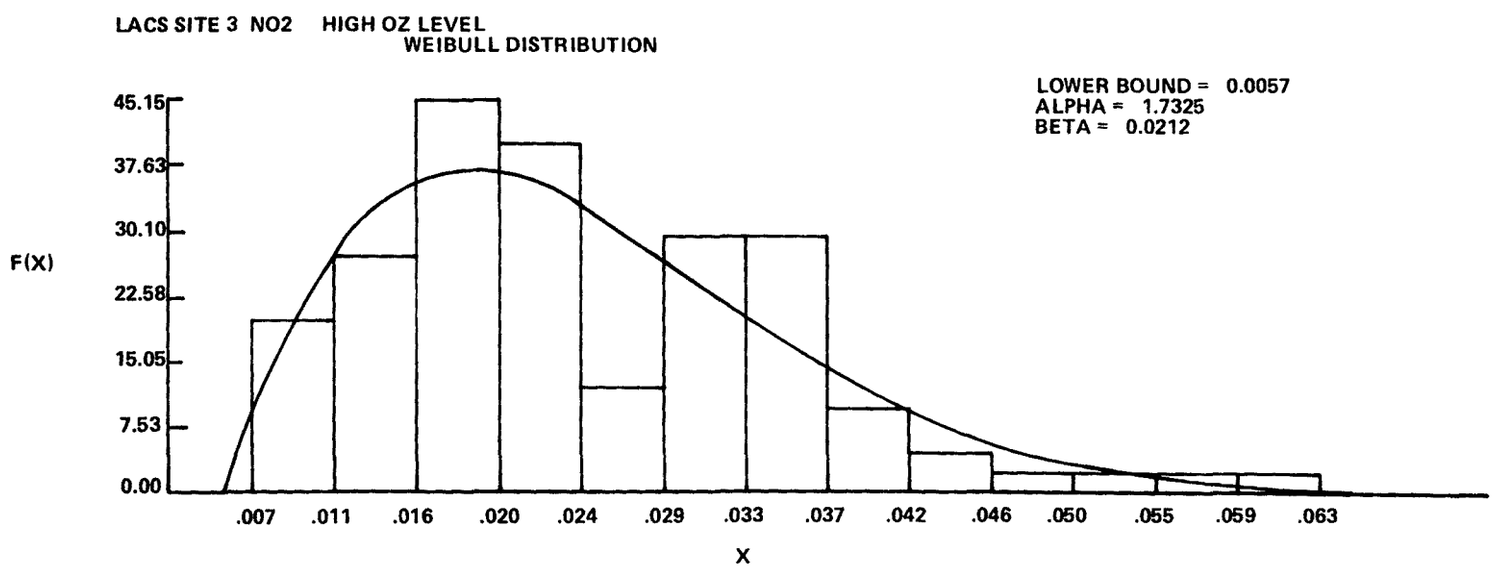

FIGURE 4 MAXFIT OUTPUT: FIT OF WEIBULL DISTRIBUTION FOR DATA IN EXAMPLE 1

LACS SITE 3 NO2 HIGH 03 LEVEL

SB DISTRIBUTION

MEAN = −1.0364
VARIANCE = .6181

MAX = .079396
MIN = .002743

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.003–	.007	.000000	.011040	−.011040	.000000	.011040	.000	1.016	−1.016
.007–	.011	.086957	.083043	.003913	.086957	.094084	8.000	7.640	.360
.011–	.016	.119565	.144029	−.024464	.206522	.238112	11.000	13.251	−2.251
.016–	.020	.195652	.161325	.034327	.402174	.399437	18.000	14.842	3.158
.020–	.024	.173913	.150945	.022968	.576087	.550382	16.000	13.887	2.113
.024–	.029	.054348	.128155	−.073807	.630435	.678537	5.000	11.790	−6.790
.029–	.033	.130435	.101948	.028486	.760870	.780485	12.000	9.379	2.621
.033–	.037	.130435	.076939	.053496	.891304	.857424	12.000	7.078	4.922
.037–	.042	.043478	.055223	−.011744	.934783	.912646	4.000	5.080	−1.080
.042–	.046	.021739	.037540	−.015800	.956522	.950186	2.000	3.454	−1.454
.046–	.050	.010870	.023917	−.013047	.967391	.974103	1.000	2.200	−1.200
.050–	.055	.010870	.014018	−.003148	.978261	.988120	1.000	1.290	−.290
.055–	.059	.010870	.007325	.003545	.989130	.995445	1.000	.674	.326
.059–	.063	.010870	.003231	.007638	1.000000	.998676	1.000	.297	.703
.063–	.079	.000000	.001324	−.001324	1.000000	1.000000	.000	.122	−.122

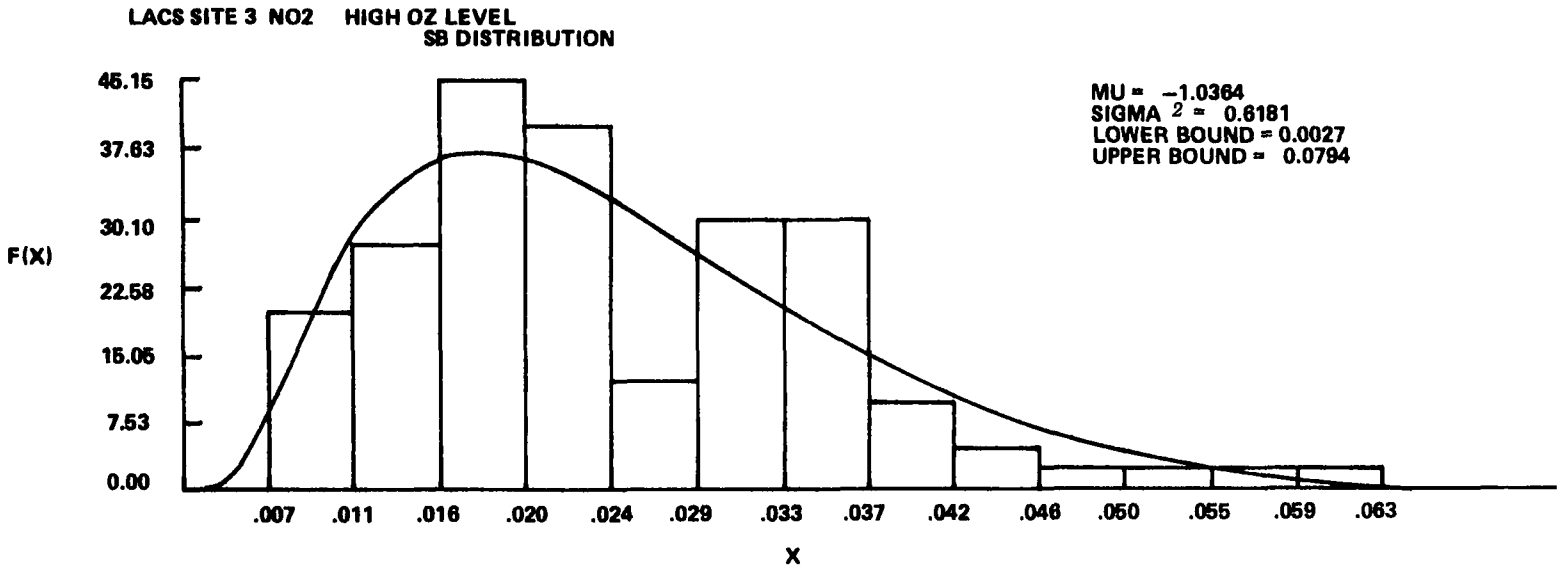

FIGURE 5 MAXFIT OUTPUT: FIT OF JOHNSON S_B DISTRIBUTION FOR DATA IN EXAMPLE 1

LACS SITE 3 NO2 HIGH 03 LEVEL

BETA DISTRIBUTION

ALPHA = 1.8673909
BETA = 5.5897199

MAX = .081378
MIN = .005647

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.006–	.007	.000000	.008148	–.008148	.000000	.008148	.000	.750	– .750
.007–	.011	.086957	.092974	– .006018	.086957	.101122	8.000	8.554	– .554
.011–	.016	.119565	.143431	– .023866	.206522	.244553	11.000	13.196	– 2.196
.016–	.020	.195652	.155967	.039685	.402174	.400520	18.000	14.349	3.651
.020–	.024	.173913	.147240	.026673	.576087	.547760	16.000	13.546	2.454
.024–	.029	.054348	.127375	– .073027	.630435	.675135	5.000	11.718	– 6.718
.029–	.033	.130435	.103017	.027417	.760870	.778152	12.000	9.478	2.522
.033–	.037	.130435	.078469	.051966	.891304	.856621	12.000	7.219	4.781
.037–	.042	.043478	.056330	– .012852	.934783	.912951	4.000	5.182	– 1.182
.042–	.046	.021739	.037954	– .016215	.956522	.950906	2.000	3.492	– 1.492
.046–	.050	.010870	.023793	– .012923	.967391	.974699	1.000	2.189	– 1.189
.050–	.055	.010870	.013676	– .002806	.978261	.988375	1.000	1.258	– .258
.055–	.059	.010870	.007042	.003827	.989130	.995417	1.000	.648	.352
.059–	.063	.010870	.003129	.007741	1.000000	.998546	1.000	.288	.712
.063–	.081	.000000	.001454	–.001454	1.000000	1.000000	.000	.134	–.134

LACS SITE 3 NO2 HIGH OZ LEVEL
BETA DISTRIBUTION

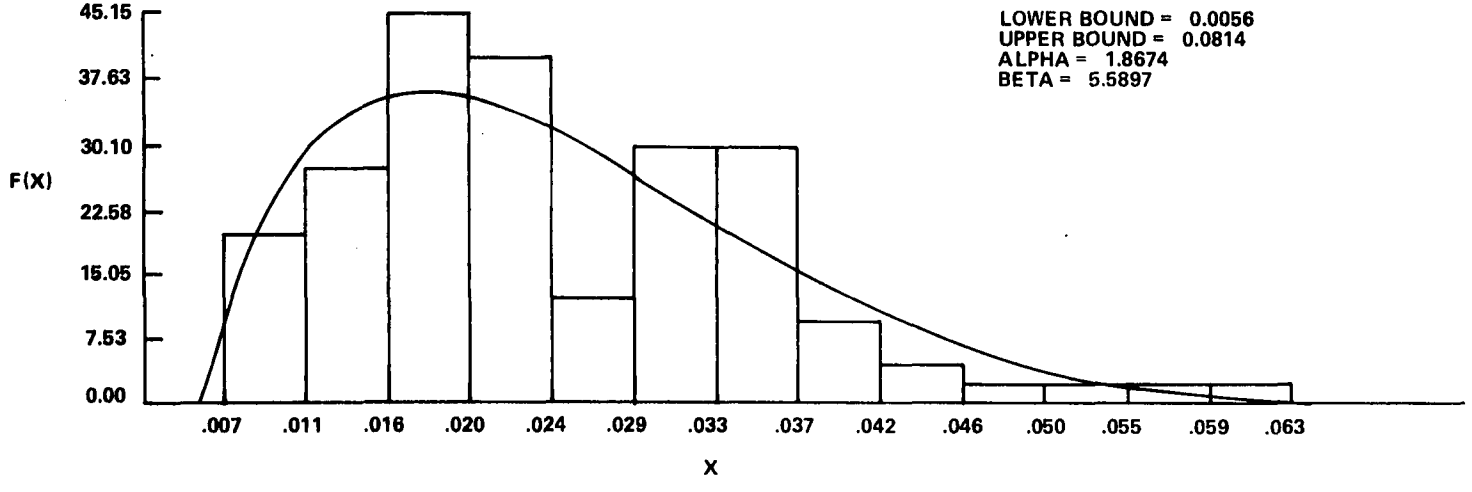

FIGURE 6 MAXFIT OUTPUT: FIT OF BETA DISTRIBUTION FOR DATA IN EXAMPLE 1

LACS SITE 3 NO 2 HIGH 03 LEVEL

	NORMAL	LOG NORMAL	GAMMA	WEIBULL	SB	BETA
ABSOLUTE DEVIATION	.36327934 + 002 (6)	.26840004 + 002 (1)	.27568650 + 002 (3)	.26914336 + 002 (2)	.28404869 + 002 (4)	.28944865 + 002 (5)
WEIGHTED ABS. DEVIATION	.36439217 + 001 (6)	.26141019 + 001 (1)	.27436755 + 001 (2)	.27989979 + 001 (3)	.28615729 + 001 (4)	.29539632 + 001 (5)
CHI–SQUARE (*)	.12294319 –001 (6)	.77316610 – 001 (2)	.49182300 – 001 (3)	.97144187 – 001 (1)	.36629757 – 001 (4)	.34430578 – 001 (5)
KOLMOGOROV–SMIRNOV	.87066924 – 001 (6)	.65347441 – 001 (5)	.61095604 – 001 (4)	.51969778 – 001 (3)	.48101939 – 001 (2)	.44699905 – 001 (1)
CRAMER–VON MISES–SMIRNOV	.18261896 + 000 (6)	.82409978 – 001 (5)	.76377309 – 001 (4)	.68976456 – 001 (2)	.64916543 – 001 (1)	.69541976 – 001 (3)
LOG LIKELIHOOD	.28306338 + 003 (6)	.28868319 + 003 (5)	.28923631 + 003 (4)	.29017593 + 003 (1)	.28974790 + 003 (3)	.29012675 + 003 (2)

* OBSERVED SIGNIFICANCE LEVEL

TABLE 2 MAXFIT OUTPUT: COMPARISON STATISTICS FOR DATA IN EXAMPLE 1

data set is shown in Figures 7-12 and Table 3. For this data set there is no obvious "best" model. Beta distribution has the highest log-likelihood value. However, gamma distribution has the lowest Kolmogorov-Smirnov statistic. The 3-parameter lognormal shows up best in the Chi-square test and weighted absolute deviations. Therefore, the 3-parameter lognormal distribution would probably serve as the "best" model in this example since it is a simple model to work with and scored reasonably well on several fit criteria.

LACS SITE 3 NOX HIGH 03 LEVEL

OBS.MIN.X = .007000
OBS.MAX.X = .059000
NO. OF OBS. = 92.0000

MEAN = .0252
VARIANCE = .0001
STANDARD DEV. = .0114

INDEX OF SKEWNESS = .7463
INDEX OF KURTOSIS = 3.2460
SKEWNESS SQUARED = .5570

NORMAL DISTRIBUTION

MEAN = .0252
VARIANCE = .0001
STANDARD DEVIATION = .0114

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
–INFINITY–	.007	.000000	.055187	– .055187	.000000	055187	.000	5.077	– 5.077
.007–	.011	.086957	.056838	.030118	.086957	.112025	8.000	5.229	2.771
.011–	.016	.076087	.089794	– .013707	.163043	.201819	7.000	8.261	– 1.261
.016–	.020	.239130	.122932	.116199	.402174	.324750	22.000	11.310	10.690
.020–	.024	.173913	.145847	.028066	.576087	.470597	16.000	13.418	2.582
.024–	.029	.043478	.149950	– .106472	.619565	.620547	4.000	13.795	– 9.795
.029–	.033	.130435	.133602	– .003167	.750000	.754149	12.000	12.291	– .291
.033–	.037	.097826	.103156	– .005330	.847826	.857305	9.000	9.490	– .490
.037–	.042	.086957	.069023	.017934	.934783	.926328	8.000	6.350	1.650
.042–	.046	.010870	.040022	– .029152	.945652	.966350	1.000	3.682	– 2.682
.046–	.050	.021739	.020110	.001629	.967391	.986460	2.000	1.850	.150
.050–	.055	.010870	.008756	.002114	.978261	.995216	1.000	.806	.194
.055–	.059	.000000	.003304	– .003304	.978261	.998519	.000	.304	– .304
.059–	.063	.021739	.001080	.020659	1.000000	.999599	2.000	.099	1.901
.063–INFINITY		.000000	.000401	– .000401	1.000000	1.000000	.000	.037	– .037

LACS SITE 3 NOX HIGH 03 LEVEL
NORMAL DISTRIBUTION

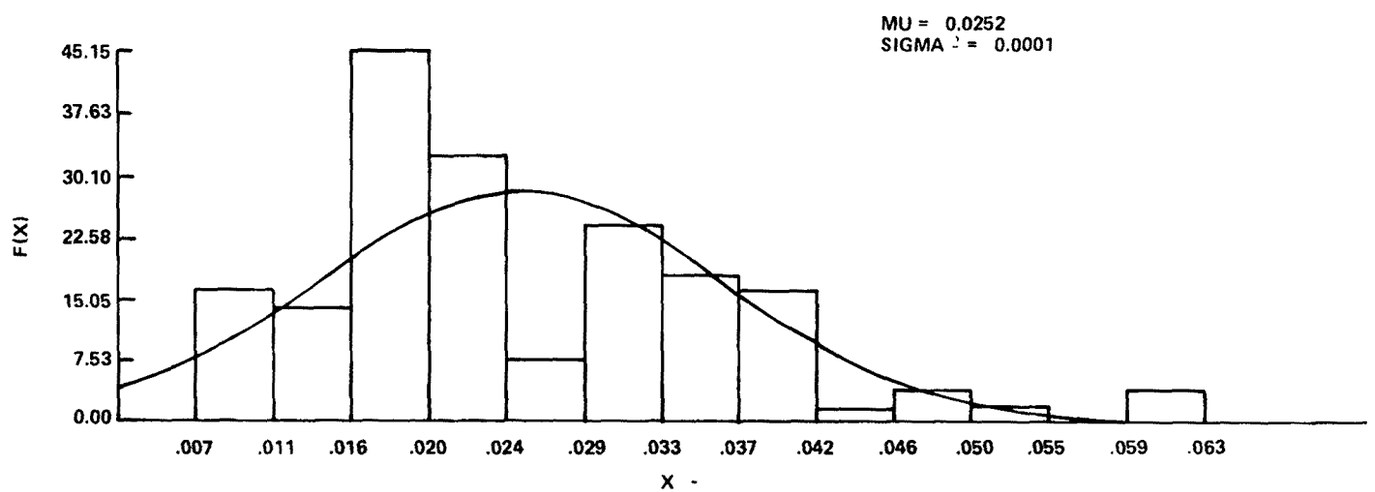

FIGURE 7 MAXFIT OUTPUT: FIT OF NORMAL DISTRIBUTION TO DATA IN EXAMPLE 2.

LACS SITE 3 NOX HIGH 03 LEVEL

LOG NORMAL DISTRIBUTION

MEAN = −3.5370
VARIANCE = .1337
STANDARD DEVIATION = .3657
SEARCHED LOWER BOUND = −.005922

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
−.006−	.007	.000000	.013136	−.013136	.000000	.013136	.000	1.209	−1.209
.007−	.011	.086957	.063109	.023847	.086957	.076246	8.000	5.806	2.194
.011−	.016	.076087	.130498	−.054411	.163043	.206744	7.000	12.006	−5.006
.016−	.020	.239130	.168756	.070374	.402174	.375500	22.000	15.526	6.474
.020−	.024	.173913	.166534	.007379	.576087	.542034	16.000	15.321	.679
.024−	.029	.043478	.139403	−.095925	.619565	.681437	4.000	12.825	−8.825
.029−	.033	.130435	.105139	.025296	.750000	.786576	12.000	9.673	2.327
.033−	.037	.097826	.074096	.023730	.847826	.860672	9.000	6.817	2.183
.037−	.042	.086957	.049934	.037023	.934783	.910606	8.000	4.594	3.406
.042−	.046	.010870	.032668	−.021799	.945652	.943274	1.000	3.005	−2.005
.046−	.050	.021739	.020962	.000777	.967391	.964237	2.000	1.929	.071
.050−	.055	.010870	.013287	−.002417	.978261	.977523	1.000	1.222	−.222
.055−	.059	.000000	.008360	−.008360	.978261	.985883	.000	.769	−.769
.059−	.063	.021739	.005241	.016498	1.000000	.991124	2.000	.482	1.518
.063−INFINITY		.000000	.008876	−.008876	1.000000	1.000000	.000	.817	−.817

LACS SITE 3 NOX HIGH 03 LEVEL
LNORM3 DISTRIBUTION

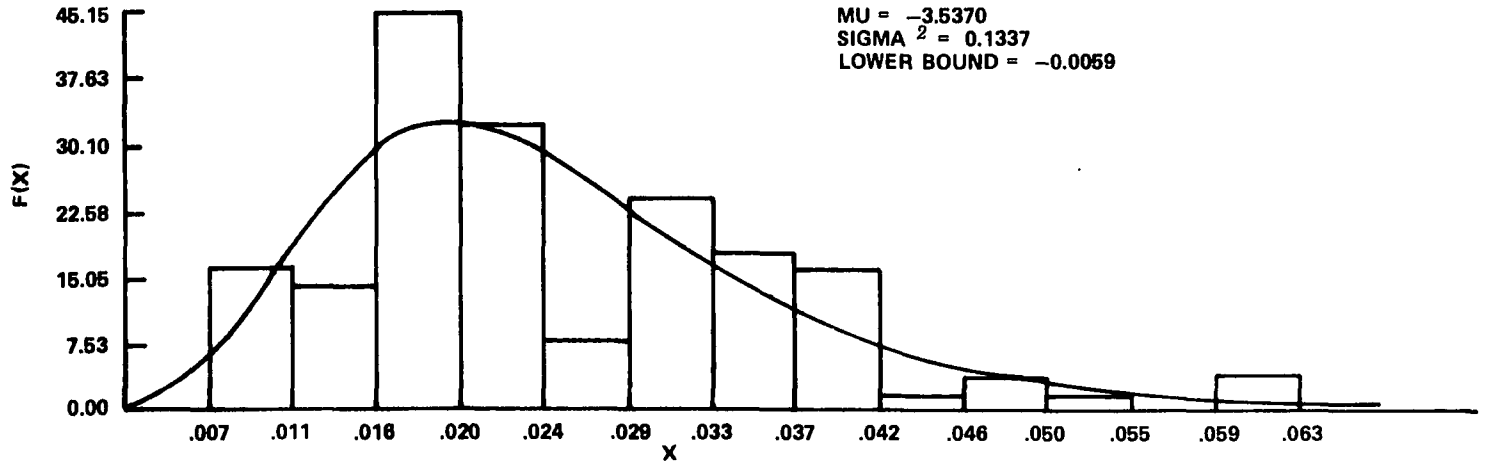

FIGURE 8 MAXFIT OUTPUT: FIT OF 3 PARAMETER LOG NORMAL DISTRIBUTION FOR DATA IN EXAMPLE 2.

LACS SITE 3 NOX HIGH 03 LEVEL

GAMMA DISTRIBUTION

ALPHA = 3.1819
BETA = .0067
SEARCHED LOWER BOUND = .003853

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.004–	.007	.000000	.008224	– .008224	.000000	.008224	.000	.757	– .757
.007–	.011	.086957	.072925	.014031	.086957	.081149	8.000	6.709	1.291
.011–	.016	.076087	.139776	– .063689	.163043	.220925	7.000	12.859	– 5.859
.016–	.020	.239130	.165745	.073385	.402174	.386670	22.000	15.249	6.751
.020–	.024	.173913	.157521	.016392	.576087	.544192	16.000	14.492	1.508
.024–	.029	.043478	.131863	– .088385	.619565	.676055	4.000	12.131	– 8.131
.029–	.033	.130435	.101669	.028766	.750000	.777724	12.000	9.354	2.646
.033–	.037	.097826	.074000	.023826	.847826	.851724	9.000	6.808	2.192
.037–	.042	.086957	.051615	.035342	.934783	.903339	8.000	4.749	3.251
.042–	.046	.010870	.034840	– .023970	.945652	.938179	1.000	3.205	– 2.205
.046–	.050	.021739	.022913	– .001174	.967391	.961092	2.000	2.108	– .108
.050–	.055	.010870	.014754	– .003884	.978261	.975846	1.000	1.357	– .357
.055–	.059	.000000	.009335	– .009335	.978261	.985181	.000	.859	– .859
.059–	.063	.021739	.005820	.015919	1.000000	.991002	2.000	.535	1.465
.063–INFINITY		.000000	.008998	– .008998	1.000000	1.000000	.000	.828	– .828

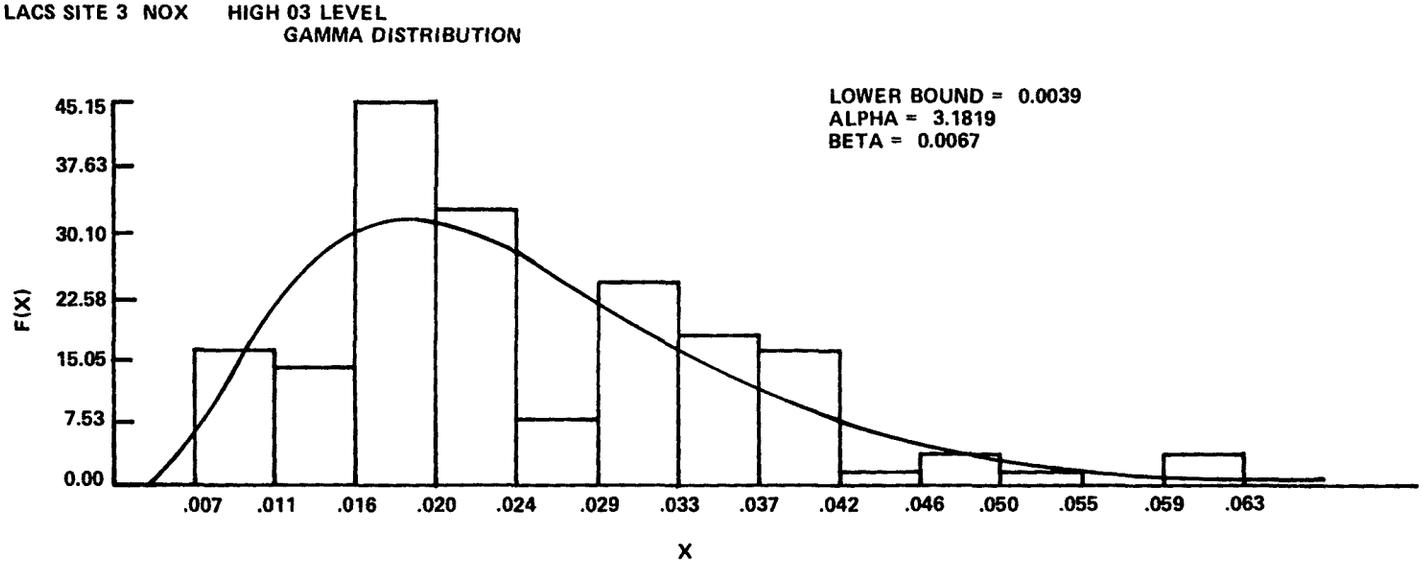

FIGURE 9 MAXFIT OUTPUT: FIT OF GAMMA DISTRIBUTION FOR DATA IN EXAMPLE 2.

LACS SITE 3 NOX HIGH 03 LEVEL

WEIBULL DISTRIBUTION

BETA = .0218
C = 1.7517
SEARCHED LOWER BOUND = .005742

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.006–	.007	.000000	.006655	–.006655	.000000	.006655	.000	.612	–.612
.007–	.011	.086957	.081266	.005691	.086957	.087921	8.000	7.476	.524
.011–	.016	.076087	.134664	–.058577	.163043	.222584	7.000	12.389	–5.389
.016–	.020	.239130	.155648	.083482	.402174	.378232	22.000	14.320	7.680
.020–	.024	.173913	.152516	.021397	.576087	.530748	16.000	14.031	1.969
.024–	.029	.043478	.133809	–.090331	.619565	.664557	4.000	12.310	–8.310
.029–	.033	.130435	.107686	.022749	.750000	.772243	12.000	9.907	2.093
.033–	.037	.097826	.080555	.017271	.847826	.852799	9.000	7.411	1.589
.037–	.042	.086957	.056479	.030477	.934783	.909278	8.000	5.196	2.804
.042–	.046	.010870	.037324	–.026454	.945652	.946602	1.000	3.434	–2.434
.046–	.050	.021739	.023344	–.001604	.967391	.969945	2.000	2.148	–.148
.050–	.055	.010870	.013861	–.002991	.978261	.983806	1.000	1.275	–.275
.055–	.059	.000000	.007832	–.007832	.978261	.991638	.000	.721	–.721
.059–	.063	.021739	.004221	.017519	1.000000	.995859	2.000	.388	1.612
.063–INFINITY		.000000	.004141	–.004141	1.000000	1.000000	.000	.381	–.381

LACS SITE 3 NOX HIGH 03 LEVEL
WEIBULL DISTRIBUTION

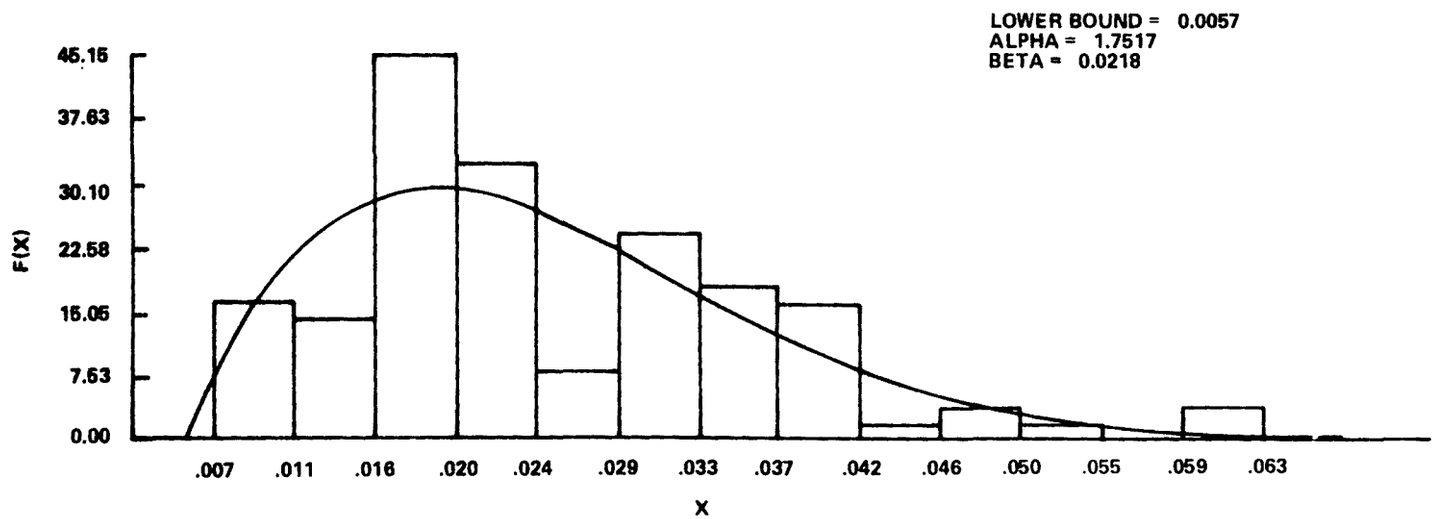

FIGURE 10 MAXFIT OUTPUT: FIT OF WEIBULL DISTRIBUTION FOR DATA IN EXAMPLE 2

LACS SITE 3 NOX HIGH 03 LEVEL

SB DISTRIBUTION

MEAN = −1.0339
VARIANCE = .6214

MAX = 080503
MIN = .002990

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.003–	.007	.000000	.008625	−.008625	.000000	.008625	.000	.794	− .794
.007–	.011	.086957	.076215	.010741	.086957	.084841	8.000	7.012	.988
.011–	.016	.076087	.138778	−.062691	.163043	.223619	7.000	12.768	− 5.768
.016–	.020	.239130	.158836	.080294	.402174	.382455	22.000	14.613	7.387
.020–	.024	.173913	.150719	.023194	.576087	.533174	16.000	13.866	2.134
.024–	.029	.043478	.129460	−.085982	.619565	.662634	4.000	11.910	− 7.910
.029–	.033	.130435	.104146	.026289	.750000	.766780	12.000	9.581	2.419
.033–	.037	.097826	.079535	.018291	.847826	.846315	9.000	7.317	1.683
.037–	.042	.086957	.057857	.029099	.934783	.904172	8.000	5.323	2.677
.042–	.046	.010870	.039966	−.029096	.945652	.944138	1.000	3.677	− 2.677
.046–	.050	.021739	.025978	−.004238	.967391	.970116	2.000	2.390	− .390
.050–	.055	.010870	.015631	−.004762	.978261	.985747	1.000	1.438	− .438
.055–	.059	.000000	.008472	−.008472	.978261	.994219	.000	.779	− .779
.059–	.063	.021739	.003946	.017793	1.000000	.998164	2.000	.363	1.637
.063–	.081	.000000	.001836	−.001836	1.000000	1.000000	.000	.169	− .169

LACS SITE 3 NOX HIGH 03 LEVEL
SB DISTRIBUTION

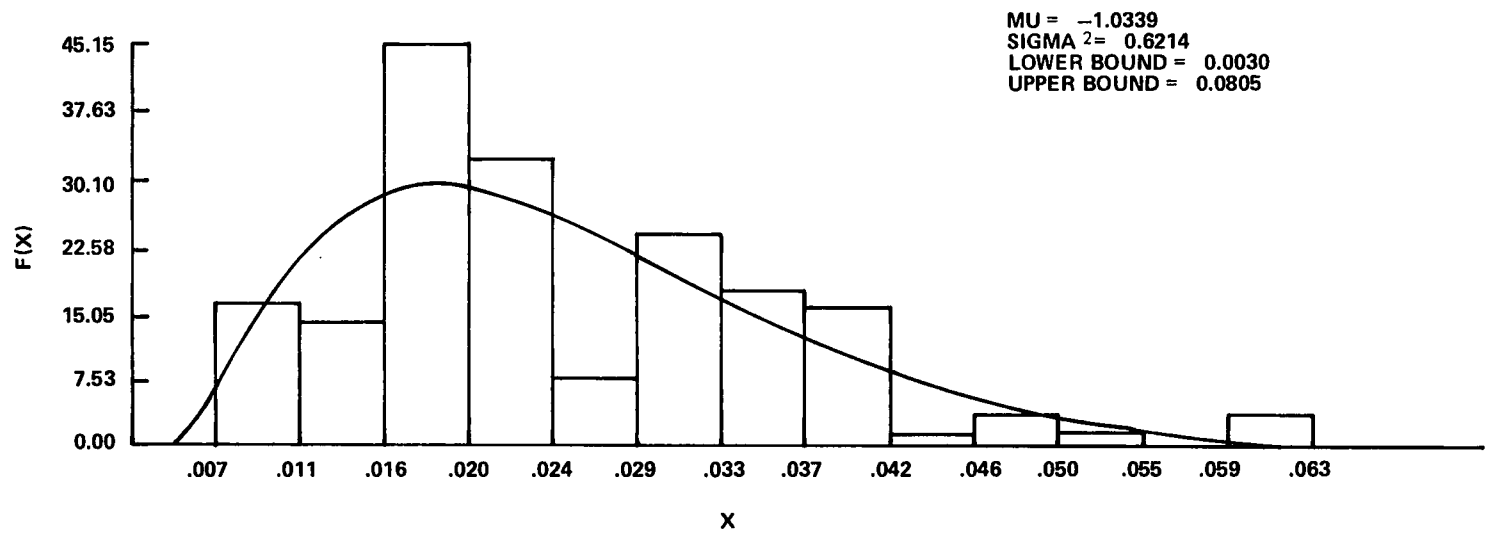

FIGURE 11 MAXFIT OUTPUT: FIT OF JOHNSON S_B DISTRIBUTION FOR DATA IN EXAMPLE 2

LACS SITE 3 NOX HIGH 03 LEVEL

BETA DISTRIBUTION

ALPHA = 1.9339035
BETA = 5.7444356

MAX = .083143
MIN = .005647

X1	X2	OBSERVED	PREDICTED	RESIDUAL	CUMULATIVE OBSERVED	CUMULATIVE PREDICTED	OBSERVED FREQUENCY	PREDICTED FREQUENCY	RESIDUAL FREQUENCY
.006–	.007	.000000	.006710	–.006710	.000000	.006710	.000	.617	–.617
.007–	.011	.086957	.084586	.002371	.086957	.091296	8.000	7.782	.218
.011–	.016	.076087	.137076	–.060989	.163043	.228372	7.000	12.611	–5.611
.016–	.020	.239130	.153161	.085970	.402174	.381532	22.000	14.091	7.909
.020–	.024	.173913	.147323	.026590	.576087	.528856	16.000	13.554	2.446
.024–	.029	.043478	.129321	–.085842	.619565	.658176	4.000	11.897	–7.897
.029–	.033	.130435	.105908	.024527	.750000	.764084	12.000	9.743	2.257
.033–	.037	.097826	.081618	.016208	.847826	.845702	9.000	7.509	1.491
.037–	.042	.086957	.059291	.027666	.934783	.904992	8.000	5.455	2.545
.042–	.046	.010870	.040477	–.029607	.945652	.945469	1.000	3.724	–2.724
.046–	.050	.021739	.025777	–.004038	.967391	.971246	2.000	2.371	–.371
.050–	.055	.010870	.015120	–.004251	.978261	.986366	1.000	1.391	–.391
.055–	.059	.000000	.008008	–.008008	.978261	.994375	..000	.737	–.737
.059–	.063	.021739	.003710	.018030	1.000000	.998084	2.000	.341	1.659
.063–	.083	.000000	.001916	–.001916	1.000000	1.000000	.000	.176	–.176

LACS SITE 3 NOX HIGH 03 LEVEL
BETA DISTRIBUTION

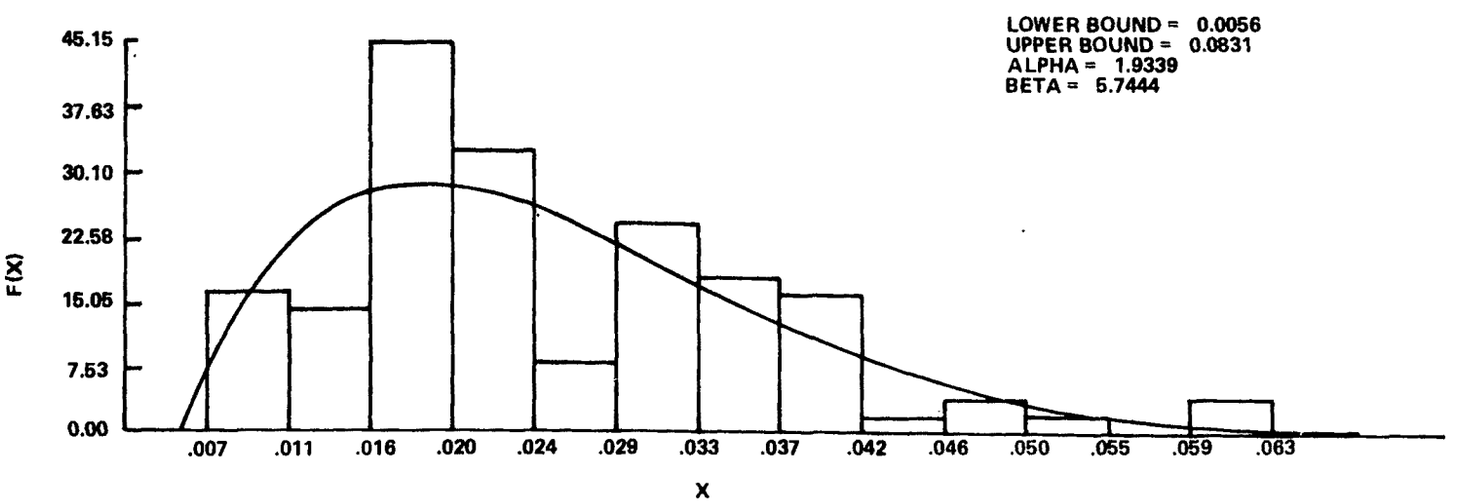

FIGURE 12 MAXFIT OUTPUT: FIT OF BETA DISTRIBUTION FOR DATA IN EXAMPLE 2

LACS SITE 3 NOX HIGH 03 LEVEL

	NORMAL	LOG NORMAL	GAMMA	WEIBULL	SB	BETA
ABSOLUTE DEVIATION	.39876353 + 002 (6)	.37706056 + 002 (3)	.38209408 + 002 (5)	.36539779 + 002 (1)	.37849178 + 002 (4)	.37050406 + 002 (2)
WEIGHTED ABS. DEVIATION	.40290285 + 001 (3)	.39116101 + 001 (1)	.40558162 + 001 (5)	.40038714 + 001 (2)	.40795364 + 001 (6)	.40345331 + 001 (4)
CHI–SQUARE (*)	.80831527 – 003 (6)	.27928927 – 001 (1)	.23569159 – 001 (2)	.91304910 – 002 (5)	.11633393 – 001 (3)	.92297671 – 002 (4)
KOLMOGOROV–SMIRNOV	.10548994 + 000 (6)	.61872242 – 001 (4)	.57881671 – 001 (1)	.59540748 – 001 (2)	.60575323 – 001 (3)	.65328118 – 001 (5)
CRAMER–VON MISES–SMIRNOV	.25241871 + 000 (6)	.11858179 + 000 (5)	.11282250 + 000 (2)	.11458265 + 000 (3)	.10825762 + 000 (1)	.11509788 + 000 (4)
LOG LIKELIHOOD	.28121652 + 003 (6)	.28741674 + 003 (5)	.28792753 + 003 (4)	.28859950 + 003 (2)	.28846644 + 003 (3)	.28860871 + 003 (1)

*OBSERVED SIGNIFICANCE LEVEL

TABLE 3 MAXFIT OUTPUT: COMPARISON STATISTICS FOR DATA IN EXAMPLE 2

SECTION 5

FUTURE DEVELOPMENTS

Obviously MAXFIT is not the answer to all air quality data distribution problems. Continuing efforts are planned to improve the program in several areas.

IMPROVEMENT OF THE SOFTWARE'S GROUPED DATA HANDLING CAPACITY

At present, software considers grouped data to be repeated observations at class mid-points. The simplistic method causes estimates to become unstable when a large proportion of the data falls in one class. This problem can be avoided by maximizing the probability that certain sample values fall in certain classes instead of maximizing a likelihood function of grouped data.

TRUNCATION AND CENSORING

If the underlying variate X cannot be observed in part or parts of its range, the distribution of X is usually termed truncated. In addition, observations which cannot be recorded below a level R, the variate X is truncated on the left. Similarly, the distribution of X can be truncated on the right and doubly truncated. Observations from any type of truncated distribution are in actuality drawn from an incomplete distribution. A truncated variate does not differ from other variates, but is treated in a special way because its distribution is generated by an underlying untruncated variable.

There are times when an experimenter is forced to put little faith in sample values when they occur above or below certain values. In these cases, such an experimenter might not wish to use the information contained in the data values in estimating parameters. This situation occurs when a value is below the minimum detectable limits of an instrument or measurement method. As in truncation, we may have censoring on the left when a certain minimum response is necessary to make a valid measurement, censoring on the right, and double censoring. Censoring can be more completely defined by distinguishing Type I and Type II censoring. Type I censoring is said to occur when the number of censored observations is a random variable, whereas, in Type II censoring the number of censored observations is fixed. Censoring is a property of the sample whereas truncation is a property of the distribution (14,15).

In the truncated and censored cases, parameters of frequency functions can be estimated by maximum likelihood estimation. For a continuous variate, with frequency function $f(x;\theta)$, the following likelihood function can be maximized when $f(x;\theta)$ is doubly truncated at known points a and b, with a<b.

$$L_T = \frac{\prod_{i=1}^{n} f(x_i;\theta)}{\left[\int_a^b f(x;\theta)\right]^n}$$

If we consider the same variate, doubly censored at fixed points a and b, with n_1 and n_2 sample observations unobserved for Type I censoring, the likelihood function when n_1 and n_2 are random variables is:

$$L_I \propto \left(\int_{-\infty}^{a} f(x;\theta)dx\right)^{n_1} \prod_{i=1}^{n-n_1-n_2} f(x_i;\theta) \left(\int_b^{\infty} f(x;\theta)dx\right)^{n_2}$$

If censoring is of Type II with n_1 and n_2 fixed, the likelihood function is:

$$L_{II} \propto \left(\int_{-\infty}^{X_{n_1+1}} f(x;\theta)dx\right)^{n_1} \prod_{i=n_1}^{n-n_2} f(x_i;\theta) \left(\int_{X_{n-n_1}}^{\infty} f(x;\theta)dx\right)^{n_2}$$

It would be informative to investigate the statistical properties of maximum likelihood estimates from truncation and censoring.

A GRAPH OF THE DATA MOMENTS ON THE β_1-β_2 PLANE

A useful measure of the flexibility of a distribution is the skewness coefficient ($\sqrt{\beta_1}$), and the kurtosis coefficient ($\sqrt{\beta_2}$), where $\sqrt{\beta_1} = U_3/U_2^{3/2}$ and $\beta_2 = U_4/U_2$ and $U_K = \int_{-\infty}^{\infty}[x-E(x)]^k f(x)dx$.

A graph of the β_1-β_2 space can visually show the range of skewness and kurtosis covered by various statistical distributions (16). The primary use of the β_1-β_2 space is to suggest distributions which might fit a set of data based on sample estimates of β_1 and β_2.

GOODNESS-OF-FIT STATISTICS

At present, there are several simplifications in the calculations of the goodness-of-fit statistics. The Kolmogorov-Smirnov and the Cramer-von Mises-Smirnov statistics are computed on the basis of grouped data. Since this introduces bias, it would be desirable to calculate these statistics upon ungrouped data. We intend to update MAXFIT to eliminate these simplifications in the near future.

SECTION 6

SUMMARY

MAXFIT is a vehicle for applying a statistically valid method of fitting distributions. The software makes use of the maximum likelihood method of estimation to fit the normal distribution, the 3-parameter lognormal distribution, the 3-parameter gamma distribution, the 3-parameter Weibull distribution, the Johnson S_B distribution, and the 4-parameter beta distribution. With the increased availability of high speed computers, we feel that maximum likelihood is a better method to fit probability models than outdated, inefficient graphical techniques.

There are many managers within EPA who must make statistical assumptions regarding the distribution of air pollution data. These assumptions may affect agency policy. Such assumptions affect standard setting, emission roll-back calculations, estimation of maximum concentrations, threshold approximations and the handling of missing observations. It is hoped MAXFIT could be used as a tool to aid in making such assumptions.

In fitting air quality data, we have found large variations in the shape of the distributions suggested by the data. We consider fitting one distribution to all air quality data to be inadvisable. Even the two highly similar data sets presented as examples in this report led to the selection of different models as "best." With this software, several models can be fit and the goodness-of-fit of each can be compared in several ways. Thus, a rational decision can be made as to which model would be adequate for a given data base and purpose.

REFERENCES

1. Larsen, R. I. A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards. Publication No. AP-89, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, 1971.

2. Larsen, R. I. An Air Quality Data Analysis System for Interrelating Effects, Standards, and Needed Source Reductions. J. Air Poll. Control Assoc. 23:993, 1973.

3. Larsen, R. I. An Air Quality Data Analysis System for Interrelating Effects, Standards, and Needed Source Reductions. Part 2, J. Air Poll. Control Assoc. 24:551, 1974.

4. Larsen, R. I. An Air Quality Data Analysis System for Interrelating Effects, Standards, and Needed Source Reductions. Part 3, J. Air. Poll. Control Assoc. 26:325, 1976.

5. Larsen, R. I. An Air Quality Data Analysis System for Interrelating Effects, Standards, and Needed Source Reductions. Part 4, J. Air Poll. Control Assoc. 27:454, 1977.

6. Mage, D. T., and W. R. Ott. Refinements of the Lognormal Probability Model for Analysis of Aerometric Data. J. Air Poll. Control Assoc. 28(8):796-798, 1978.

7. Ott, W. R., and D. T. Mage. A General Purpose Univariate Probability Model for Environmental Data Analysis. Comput. & Ops. Res. 3:209-216, 1976.

8. Curran, T. C., and N. H. Frank. Assessing the Validity of the Lognormal Model When Predicting Maximum Air Pollution Concentrations. Annual Meeting of the Air Pollution Control Association, Boston, Massachusetts, 1975.

9. Rao, C. R. Linear Statistical Inference and Its Applications. 2nd Ed., John Wiley and Sons, New York City, New York, 1973.

10. Schreuder, H. T., W. L. Hafley, E. W. Whitehorne, and B. T. Dare, Maximum Likelihood Estimation for Selected Distributions (MLESD). Tech. Report No. 61, School of Forest Resources, North Carolina State University, Raleigh, North Carolina, 1978.

11. Box, M. J., D. Davies, and W. H. Swann. Non-linear Optimization Techniques. Oliver & Boyd, Edinburgh, Scotland, 1969.

12. Rodes, C. R., and D. M. Holland. NO_2/O_3 Sampler Siting Study. (To be published as an EPA Environmental Monitoring Series Report).

13. IMSL Subroutine Library, Vol. 1 & 2, International Mathematical and Statistical Library, Inc., Houston, Texas, 1975.

14. Hald, A. Maximum Likelihood Estimation of the Parameters of a Normal Distribution which is Truncated at a Known Point. Skandinavisk Aktuarietidskrift, Vol. 32:119, 1949.

15. Kendall, M. G., and A. Stuart. The Advanced Theory of Statistics. Vol. 2, 2nd Ed. Hafner Publishing Co., New York City, New York, 1967.

16. Johnson, N. L., and S. Kotz. Continuous Univariate Distributions - 1. Houghton Mifflin Co., Boston, Massachusetts, 1970.

APPENDIX A. DISTRIBUTION DESCRIPTIONS

Normal Distribution:

$-\infty < x < \infty$
$-\infty < \mu < \infty$
$0 < \sigma^2 < \infty$

Density: $f(x) = EXP[-(x - \mu)^2/2\sigma^2]/\sqrt{2\pi\sigma^2}$

Estimates:

grouped data:* $\hat{\mu} = \dfrac{\sum x_i \cdot n_i}{n}$ $\quad \hat{\sigma}^2 = \dfrac{\sum (x_i - \hat{\mu})^2 \cdot n_i}{n}$

ungrouped data: $\hat{\mu} = \dfrac{\sum x_i}{n}$ $\quad \hat{\sigma}^2 = \dfrac{(x_i - \mu)^2}{n}$

Johnson S_L (3-Parameter Lognormal Distribution):

$\lambda < x < \infty$
$-\infty < \lambda < \infty$
$-\infty < \mu < \infty$
$0 < \sigma^2 < \infty$

Density: $f(x) = EXP\{-[\log(x - \lambda) - \mu]^2/2\sigma^2\}/[\sqrt{2\pi\sigma^2} \cdot (x - \lambda)]$

Estimates:

grouped data: $\hat{\lambda}$ is searched for, $\hat{\mu} = \dfrac{\sum \log(x_i - \hat{\lambda}) \cdot n_i}{n}$

$$\hat{\sigma}^2 = \frac{\sum [\log(x_i - \hat{\lambda})^2 - \hat{\mu}]^2 n_i}{n}$$

ungrouped data: $\hat{\lambda}$ is searched for, $\hat{\mu} = \dfrac{\sum \log(x_i - \hat{\lambda})}{n}$

$$\hat{\sigma}^2 = \frac{[\log(x_i - \hat{\lambda}) - \hat{\mu}]^2}{n}$$

Johnson S_B Distribution (a 4-Parameter Lognormal Distribution):

Density: $f(x) = EXP\{-[\log \frac{x-\lambda}{\xi - x} - \mu]^2/2\sigma^2\} \cdot (\xi - \lambda)/[\sqrt{2\pi\sigma^2} \cdot (X - \lambda) \cdot (\xi - X)]$

$\lambda < X < \xi$
$-\infty < \lambda < \xi < \infty$
$-\infty < \mu < \infty$
$0 < \sigma^2 < \infty$

(continued)

*In the case of grouped data, n_i = the number of observations falling in the ith class, x_i = the midpoint of the ith class, and $n = n_1$.

APPENDIX A (continued)

Estimates:

grouped data: $\hat{\lambda}$ and $\hat{\xi}$ are searched for, $\hat{\mu} = \dfrac{\sum \log\left(\dfrac{x_i - \hat{\lambda}}{\hat{\xi} - x_i}\right) \cdot n_i}{n}$

$$\hat{\sigma}^2 = \frac{\sum [\log\left(\dfrac{x_i - \hat{\lambda}}{\hat{\xi} - x_i}\right) - \hat{\mu}]^2 \cdot n_i}{n}$$

ungrouped data: $\hat{\lambda}$ and $\hat{\xi}$ are searched for, $\hat{\mu} = \dfrac{\sum \log\left(\dfrac{x_i - \hat{\lambda}}{\hat{\xi} - x_i}\right)}{n}$

$$\hat{\sigma}^2 = \frac{\sum [\log\left(\dfrac{x_i - \hat{\lambda}}{\hat{\xi} - x_i}\right) - \hat{\mu}]^2}{n}$$

Gamma Distribution:

Density: $f(x) = \dfrac{(x - \lambda)^{\alpha-1} \cdot \mathrm{EXP}[-(x - \lambda)/\beta]}{\beta^{\alpha}\Gamma(\alpha)^*}$ $\quad \lambda < x < \infty$, $-\infty < \lambda < \infty$, $0 < \alpha < \infty$, $0 < \beta < \infty$

Estimates: λ, α, and β are searched for in both the grouped and ungrouped data sets

Weibull Distribution:

Density: $f(x) = \dfrac{\alpha}{\beta} \cdot \left(\dfrac{x-\lambda}{\beta}\right)^{\alpha-1} \cdot \mathrm{EXP}[-\left(\dfrac{x-\lambda}{\beta}\right)^{\alpha}]$ $\quad \lambda < x < \infty$, $-\infty < \lambda < \infty$, $0 < \alpha < \infty$, $0 < \beta < \infty$

Estimates:

grouped data: $\hat{\lambda}$ is searched for,

$\hat{\alpha}$ must satisfy $\dfrac{\sum (x_i - \hat{\lambda})^{\hat{\alpha}} \log(x_i - \hat{\lambda}) \cdot n_i}{\sum (x_i - \hat{\lambda})^{\hat{\alpha}} \cdot n_i} - \dfrac{\sum \log(x_i - \hat{\lambda}) \cdot n_i}{n} - \hat{\alpha} = 0$

$$\hat{\beta} = \left(\frac{\sum (x_i - \hat{\lambda})^{\hat{\alpha}} \cdot n_i}{n}\right)^{\frac{1}{\hat{\alpha}}}$$

ungrouped data: $\hat{\lambda}$ is searched for,

$\hat{\alpha}$ must satisfy $\dfrac{\sum (x_i - \hat{\lambda})^{\hat{\alpha}} \log(x_i - \hat{\lambda})}{\sum (x_i - \hat{\lambda})^{\hat{\alpha}}} - \dfrac{\sum \log(x_i - \hat{\lambda})}{n} - \hat{\alpha} = 0$

(continued)

* $\Gamma(\alpha) = \int_0^{\infty} X^{\alpha-1} e^{-x} dx.$

APPENDIX A (continued)

Beta Distribution:

Density: $f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\ \Gamma(\beta)} \quad (\xi-\lambda)^{1-\alpha-\beta}(x-\lambda)^{\alpha-1}(\xi-x)^{\beta-1}$

$\lambda<x<\xi$
$-\infty<\lambda<\xi<\infty$
$0<\alpha<\infty$
$0<\beta<\infty$

Estimates:

ungrouped: $\hat{\alpha}$, $\hat{\beta}$, $\hat{\lambda}$, and $\hat{\xi}$ are searched for.

grouped data: $\hat{\alpha}$, $\hat{\beta}$, $\hat{\lambda}$, and $\hat{\xi}$ are searched for.

APPENDIX B

```
      SUBROUTINE USER(X,P,XB,NGRP,TF,N,NC,BEGIN,STEP,TOL,
     ◆ TITLE,XMIN,XMAX,NDIST)
      IMPLICIT DOUBLE PRECISION(A-H,O-Y)
      REAL TITLE(1)
      DATA IFL/'H'/
      DIMENSION X(1),XB(1),NDIST(1),P(1)
      I=0
      XMAX=-9.D6
      XMIN=9.D6
      TOL=.0001D0
      TITLE(1)='LACS S'
      TITLE(2)='ITE 3 '
      TITLE(3)=' NOX  '
      TITLE(4)=' HIGH '
      TITLE(5)='O3 LEV'
      TITLE(6)='EL    '
2     READ (5,10,END=3) IOZL,TC,AVGSP,ISITE,RNOX
10    FORMAT(A1,T17,F7.0,F6.2,I3,T48,F5.3)
      IF (ISITE.NE.3.OR.IOZL.NE.IFL) GO TO 5
      IF (RNOX.LE.1.D-6) GO TO 5
      I=I+1
      X(I)=RNOX
      XI=RNOX
      IF(XI.LT.XMIN) XMIN=XI
      IF(XI.GT.XMAX) XMAX=XI
5     GO TO 2
3     DO 4 J=1,6
4     NDIST(J)=1
      NGRP=0
      N=I
      TF=N
      NC=0
      BEGIN=XMIN-1.D-5
      STEP=(XMAX-XMIN)/12.D0
      RETURN
      END
```

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA 600/4-79-044	2.	3. RECIPIENT'S ACCESSION NO.
4. TITLE AND SUBTITLE THE MAXIMUM LIKELIHOOD APPROACH TO PROBABILISTIC MODELING OF AIR QUALITY DATA		5. REPORT DATE July 1979
		6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S) Terence Fitz-Simons, David M. Holland		8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS		10. PROGRAM ELEMENT NO. 1AD 883
		11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS Environmental Monitoring and Support Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, N.C. 27711		13. TYPE OF REPORT AND PERIOD COVERED Final
		14. SPONSORING AGENCY CODE EPA-600/08
15. SUPPLEMENTARY NOTES		

16. ABSTRACT

Software using maximum likelihood estimation to fit six probabilistic models is presented. The software is designed as a tool for the air pollution researcher to determine what assumptions are valid in the statistical analysis of air pollution data for the purposes of standard setting, roll-back calculations, estimation of maximum concentrations, threshold approximations, and handling missing observations. The program fits user's data to the normal distribution, the 3-parameter lognormal distribution, the 3-parameter Weibull distribution, the 3-parameter gamma distribution, the Johnson S_B distribution (a 4-parameter lognormal distribution), and the 4-parameter beta distribution. The parameters are estimated using standard closed solutions to maximizing equations, and a golden section search for all other parameters. Graphical output contains a histogram of the data superimposed by the fitted density for each model. Six goodness-of-fit criteria are supplied and ranked by the program to aid in the selection of the most appropriate choice among the six models. These criteria are absolute deviations (AD statistic), weighted absolute deviations (WAD statistic), Kolmogorov-Smirnov statistic, Cramer-von Mises-Smirnov statistic, the log-likelihood function, and the observed significance level of the Chi-square goodness-of-fit test. The results of applying the program to several subsets of the Los Angeles Catalyst Study data base are presented.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Maximum Likelihood estimations, Lognormal, gamma, Weibull, beta, Johnson S_B, Johnson S_L, Distribution	Statistics Statistical Modeling	43F 68A

18. DISTRIBUTION STATEMENT Release to Public	19. SECURITY CLASS *(This Report)* Unclassified	21. NO. OF PAGES 33
	20. SECURITY CLASS *(This page)* Unclassified	22. PRICE -

EPA Form 2220-1 (9-73)