EPA-600/5-75-015
September 1975

# A QUANTITATIVE METHOD FOR EFFLUENT COMPLIANCE MONITORING RESOURCE ALLOCATION

By

Arthur I. Cohen
Yaakov Bar-Shalom
Wendy Winkler
G. Paul Grimsrud

Contract No. 68-01-2232
Program Element No. 1HC619

Project Officer

Donald H. Lewis
Office of Air, Land, and Water Use
Environmental Protection Agency
Washington, D.C. 20460

Prepared for
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
WASHINGTON, D.C. 20460

ia'

Research reports of the Office of Research and Development, Environmental Protection Agency, have been grouped into five series. These five 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 five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the SOCIOECONOMIC ENVIRONMENTAL STUDIES series. This series describes research on the socioeconomic impact of environmental problems. This covers recycling and other recovery operations with emphasis on monetary incentives. The non-scientific realms of legal systems, cultural values, and business systems are also involved. Because of their interdisciplinary scope, system evaluations and environmental management reports are included in this series.

This report has been reviewed by the Office of Research and Development. Approval does not signify that the contents necessarily reflect the views and policies of the Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

Document is available to the public through the National Technical Information Service, Springfield, Virginia 22151.

## EPA REVIEW NOTICE

This report has been reviewed by the Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

# ABSTRACT

This report develops and demonstrates a quantitative method for the preliminary design of effluent standard surveillance systems. The principal output of the report is a procedure to be used in the state or EPA water quality programs to determine the frequency of effluent compliance monitoring visits. The procedure allocates compliance monitoring budgetary resources so as to minimize environmental damage. It utilizes a statistical model of the effluents that is obtained from self-monitoring and compliance monitoring data. The procedure is demonstrated on an example river basin using data supplied by the State of Michigan.



This report is submitted in fulfillment of Contract Number 68-01-2232, by Systems Control, Inc., under the sponsorship of the Office of Research and Development, Environmental Protection Agency. Work was completed as of January 1975.

CONTENTS

CONTENTS (Cont'd)

FIGURES

FIGURES (Cont'd)

FIGURES (Cont'd)

TABLES

TABLES (Cont'd)

TABLES (Cont'd)

## ACKNOWLEDGEMENTS

The authors wish to acknowledge the contributions of a number of individuals that aided in the development of this report. In particular, recognition is given to the continuing support and guidance provided by the Project Officer, Mr. Donald H. Lewis of the Washington Environmental Research Center, Environmental Protection Agency. From the professional staff of Systems Control, Inc., Drs. Robert Larson, Robert Schainker, Harish Rao, and John Finnemore contributed many ideas to the original formulation of the methodology. Dr. Schainker also reviewed the draft manuscript and provided many suggestions helpful in finalizing this document. Special recognition also goes to Mr. Fred Morley, Mr. Fred Cowles, Mr. Richard Christianson and their associates at the Michigan Department of Natural Resources, for their time and support in providing the data base used in the demonstration example.

# SECTION I

## CONCLUSIONS

A procedure has been developed which determines how often to simple effluent sources in a compliance monitoring program. The sampling frequencies depend on the probability each source will be in violation of its standards, as well as the environmental damage each source is expected to cause to the receiving waters.

The potential utility of the procedure was demonstrated using data from 30 industries and municipal treatment plants. The sources chosen by the procedure for monitoring with highest priority were shown to be those sources most likely to violate a standard and cause environmental damage.

The information produced by the priority setting procedure is applicable to many types of water quality studies. The statistical descriptions of the effluents can be used as inputs to water quality models. The environmental damage expected from a source and the probability that a source will be in violation of a standard can be useful in the jetting of effluent standards in "water quality limited" reaches of a river basin. The examination of these quantities quickly tells the user which sources are expected to have a major effect on water quality. The sensitivity of these quantities to changes in the standards or loadings can also be quickly determined.

# SECTION II
## RECOMMENDATIONS

The priority setting procedure developed in this report should be implemented as a user-oriented computer program. Such a program would be of great benefit to the monitoring agencies in the setting of sampling frequencies. A handbook should also be developed to describe the procedure to non-statistically trained personnel.

Notwithstanding the above recommendation, there are certain studies that can serve to increase the procedure's usefulness:

1) Geographical Considerations. In a river basin, there will exist groups of effluent sources located in close proximity to each other. When monitoring one source of the group, it may be beneficial to monitor another, since the cost of monitoring the sources concurrently will be less than the cost of monitoring them at different times. It is suggested that the priority procedure be augmented to account for such geographical considerations.

2) Scheduling of Monitoring Visits. Given the sampling frequencies, the compliance monitor must schedule his inspection crews over the monitoring period. This can be a difficult and time consuming job, especially in large regions. It is suggested that a computer program be developed that schedules the monitoring visits taking into account manpower, equipment and geographical factors.

3) Statistical Analysis. The procedure developed in this report allows the user to choose between two statistical distributions to describe each constituent of each source. The user also can specify whether the constituents of a source are statistically correlated or uncorrelated. The sampling priorities established

by the procedure are sensitive to both of these choices. There has been, however, little study on an industry by industry basis as to either the distribution of or correlation between constituents. It is suggested that there be further study into these statistical considerations.

4) Allocation Criteria. In this report, two allocation criteria are specified: (A) "Cost" of undetected violations and (B) number of undetected violators. Additional useful criteria can be specified within the framework of the present procedure. As examples, consider the following two criteria: (A') violation "cost" of undetected violators and (B') degree of violation due to undetected violators. (A) and (A') both attach a cost, as measured by a damage function, to the effect due to an effluent source's load on pollutant concentration in a stream. Criteria (A) attaches a "cost" to a pollutant even if it is not violating an effluent standard while (A') only considers "violation cost" (i.e., it is assumed that no damage is done to the environment if the standard is. not violated). The rational for using (A') over (A) is that the monitor may only be interested in damage due to standard violations and not in damage per se.

(B') is a measure of the degree of violation expected from the sources in the monitoring region. Thus. under (B'), those sources who have highest probability of being violators and which are expected to have loads most over their standard will be sampled with highest priority while under (B) only the former condition is considered.

Since these criteria may be useful to the monitoring agency, it is suggested that the priority procedure should be extended to include these additional allocation criteria.

## SECTION III
## INTRODUCTION

The Federal Water Pollution Control Act Amendments of 1972 (PL 92-500) requires the establishment of effluent limitations for all point sources by July 1, 1977. The effluent limitations* are stated as conditions on discharge permits issued to all point sources under the National Pollution Discharge Elimination System (NPDES). The Environmental Protection Agency or the state is required to establish monitoring programs to ensure that the effluent sources are in compliance with the standards. There are three ways the monitoring agency obtains information concerning the compliance of the dischargers:

(1) Self-Monitoring. The source is required to monitor its effluent levels and periodically transmit these records to the monitoring agency.

(2) Compliance Monitoring. The monitoring agency visits the source to ensure that the self-monitoring is being properly executed and reported.

(3) Ambient Monitoring. The water quality monitoring of the receiving waters.

---

*A distinction is often made between effluent limitation and effluent standard. In this report these words shall be used interchangeably to denote a restriction established by the appropriate regulatory authority on the quantities and/or concentrations of chemical or biological constituents of point source wastewater.

The self-monitoring reports are the principal source of compliance information used by monitoring agencies since the agency expense to acquire these data is minimal. Some check is, however, needed on the reliability of the self-monitoring data. The compliance monitoring program is set up to provide that check. The compliance program also has other purposes associated with the permit program, such as verifying that the plant processes described in the permit are correct, evaluating new waste removal equipment, reviewing progress toward scheduled pollution control activities, and monitoring to aid in preparing enforcement actions. The ambient monitoring is primarily used to determine water quality, discern trends in water quality, and evaluate the overall effectiveness of pollution control in a region. Under certain conditions, however, ambient monitoring may flag effluent irregularities unmeasured by other means. Through knowledge of the effluent sources that could contribute to the decline in ambient quality, action can be initiated against possible violators.

This report is concerned with that part of the compliance monitoring program that determines whether the sources are in compliance with the effluent standards. Since the monitoring agency has limited resources available for compliance monitoring, it is important that these resources be used in an efficient manner. In this report, a procedure is developed which determines how often to monitor each source in a region to obtain maximum benefit from the compliance monitoring program. The procedure utilizes information from self-monitoring, ambient monitoring, and past compliance monitoring reports.

There are two types of effluent standards that have been established under NPDES: (i) a monthly average and (ii) a daily maximum. A source is in violation if either the value of a daily composite measurement exceeds the maximum standard <u>or</u> the average of the daily composites, over the month, exceeds the average standard. In order to determine whether an effluent source is in violation of the average standard, it is necessary to make measurements over a large percentage of the month; while to determine if the maximum standard is violated, it is only necessary to determine if the standard was exceeded over a single day. Since compliance monitoring is costly to the monitoring agency and since most regions will contain many effluent sources, it is not expected, in general, that compliance monitoring resources will be available to determine whether the average standard is violated. Therefore, in this report compliance monitoring is limited to determining whether the maximum standard is violated.

The remainder of this report is organized as follows: Section IV contains a summary of the priority setting procedure developed in this report. Its purpose is to introduce the procedure to potential users. Section V develops a statistical characterization of effluent source constituents and discusses how to obtain the statistical description of the effluents that is required to initialize the priority procedure. A method is also presented which specifies how the effluent statistics can be updated as additional data become available. Section VI formulates the criterion to be optimized in the priority setting procedure, denoted the "cost" of undetected violations. Also presented in this section is a discussion of the relationship between ambient quality and effluent load. Section VII restates the priority setting problem in terms of an optimization problem and describes the method used to solve it. Section VIII gives an overall description of how all the components needed to obtain the monitoring frequencies interact and presents a simplified example showing the procedure's operation. Section IX demonstrates the priority procedure on a detailed example utilizing data supplied by the State of Michigan.

# SECTION IV

## SUMMARY

The Purpose of this summary is to introduce the procedure developed in this report to potential users, that is, the compliance monitoring staff of the state or EPA effluent monitoring programs. This summary also describes the basic considerations used in the development of this procedure.

The procedure developed in this report sets priorities as to which sources should be monitored and with what frequency. The procedure determines the sampling frequencies so that those sources that have a high probability of violating their standards and that can be expected to cause large environmental damage will be sampled with high priority. The objective in allocating the monitoring resources then is to minimize the "cost" of undetected violations, or equivalently, the expected environmental damage that would result from undetected violations. The "cost" of undetected violations for an effluent source depends on

(1) The expected frequency of a standard violation
(2) The expected magnitude of the violation
(3) The toxicity of the pollutants
(4) The assimilative capacity of the receiving waters at the discharge points.

These quantities are determined from past compliance and self-monitoring reports, effluent standards, and knowledge of the receiving water characteristics and the nature of the pollutants.

The user, at his option, can specify another allocation criterion, namely, the number of undetected violators. This criterion depends on the expected frequency of a standard violation.

Both, the "cost" of undetected violations and number of undetected violators assume that if the monitoring agency catches a violator once in the monitoring period, this component of compliance monitoring has done its job. At this point, it is up to the user to specify any follow-up actions (e.g., the monitor could elect to stay at the violator's site for a longer period or specify a given number of further visits during the monitoring period).

## RESOURCE ALLOCATION PROGRAM

The basic flow of the procedure, denoted the Resource Allocation Program, is given in Figure 4.1. The various components are briefly described below

(1) Initialize Statistical Description

Combine the self-monitoring and compliance monitoring data to obtain an initial statistical description for each pollutant of each source.

(2) Calculate Expected Damage and Probability of Violation

Use the statistical description of the effluent loads, the effluent standards, and the stream parameters to obtain, for each source, its expected environmental damage and its probability of violation of the standards.

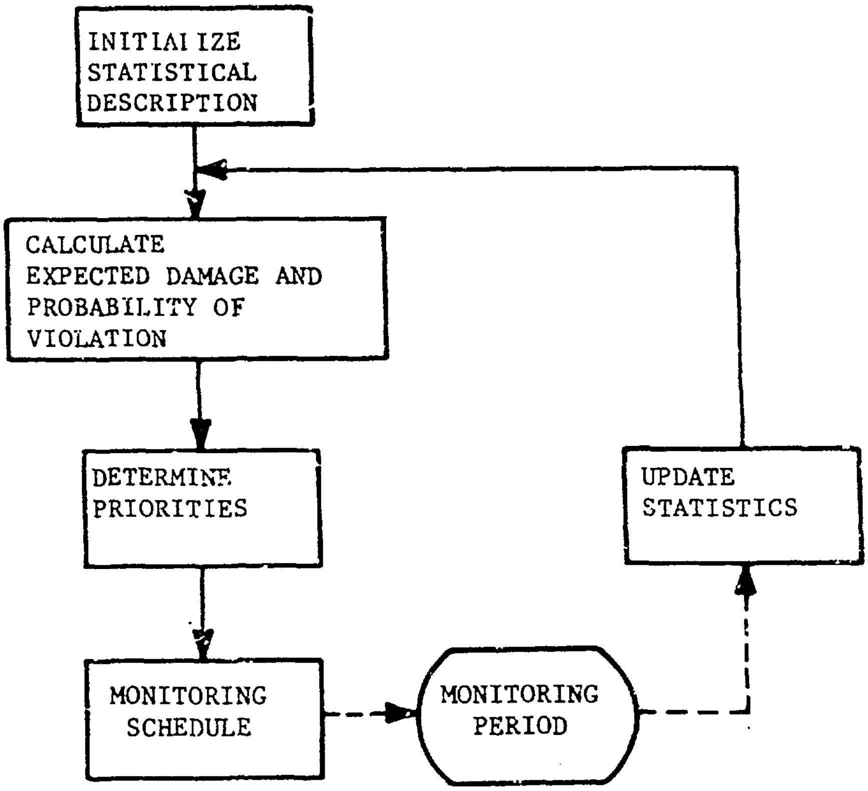


Figure 4.1 Resource Allocation Program.

(3) Determine Priorities

Allocate the monitoring resources to minimize the "cost" of undetected violations.

(4) Monitoring Schedule

Take the sampling frequencies obtained in the previous component and determine which day of the monitoring period to sample which sources.

(5) Monitoring Period

This component represents the actual time spent monitoring the sources.

(6) Update Statistics

Combine new self-monitoring and compliance data with the initial statistics to obtain an updated statistical description of the effluents for use in the next monitoring period allocation.

This procedure has been implemented as a computer program to minimize the need for data handling and hand calculations. In the remainder of this section, several of the components of the Resource Allocation Program are described in more detail.*

## Initialize Statistical Description

The daily composite value of each constituent of each source for which there is a standard is modeled by a probability density function or frequency distribution. The area under the density function between any two values of effluent specifies the fraction of the time the output of the source is between those two values. The area under the density function from zero to infinity is, clearly, always one. By allowing two types of density functions, normal and lognormal, a wide range of effluent loadings can be modeled with sufficient accuracy for determining sampling priorities. The normal

*For a description of the theoretical foundations of this procedure, refer to Sections V through VIII.

density function is the standard "bell-shaped" frequency distribution. An effluent load is distributed with a lognormal distribution if the logs of the effluent values have a normal density function. Examples of a normal and lognormal density functions are shown in Figures 4.2a and 4.2b. Both the normal and lognormal density functions are parameterized by two parameters, a mean and a standard deviation. (For the lognormal case the mean and standard deviation are of the logs of the effluent values.) These parameters are obtained for each constituent of each source from the self-monitoring and compliance monitoring data. The parameters are then fed into the next stage of the Resource Allocation Program.

Calculate Expected Damage and Probability of Violation

The monitoring frequencies depend on the environmental damage each source is expected to cause and the probability that each source is in violation of its standards. The environmental damage is related to the concentration of the water quality indicators in the receiving waters corresponding to the constitutents of the effluent. A value from 0 to 10 is given to each value of concentration depending on the degree of damage to the environment; this relationship is subjective and can be changed to meet the requirements of the user. The expected damage due to the constituents is then found by calculating the concentration of the pollutants in the receiving waters due to the source load, and then determining the environmental damage. The probability of violation of the daily standard for each constituent is simply the area under the constituent's density function to the right of the effluent standard. The environmental damage due to all the constituents from a source is the maximum of the damages due to each of the constituents, since water quality is typically limited by the pollutant causing the most damage. The probability of any of the constituents in the effluent exceeding its standard is a simple function of the probabilities that each individual constituent exceeds its standard. The expected damage and probability of violation of each source is fed into the next stage of the Resource Allocation Program.

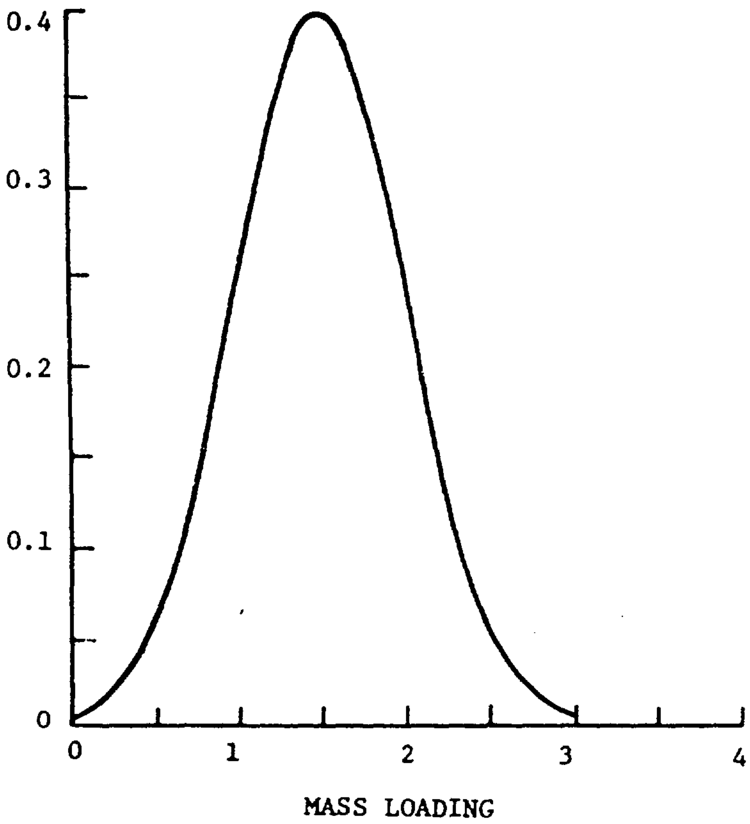


Figure 4.2a Example of normal density function.

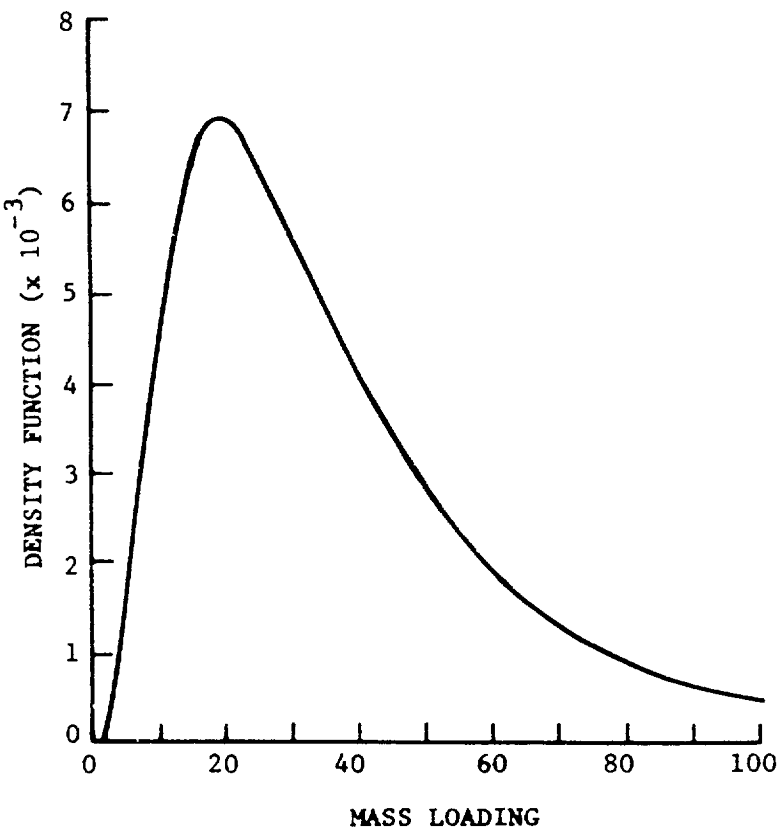


Figure 4.2b Example of lognormal density function.

## Determine Priorities

The criterion for the priority setting procedure, the "cost" of undetected violations, is defined as the expected damage that would occur due to undetected violations. This function depends on the expected damage and probability of violation of each source. As a source is sampled more times, the "cost" of undetected violations for that source decreases since the probability decreases that the source will not be found in violation on any one of the visits. The priority procedure then allocates the monitoring resources to visit those sources where the marginal return (i.e. the decrease in "cost" per dollar spent) is greatest. Therefore, given a monitoring budget or a maximum allowed "cost" of undetected violations, the priority procedure specifies the frequency with which each source should be sampled in the monitoring period. It should be noted that the criterion can be easily altered to represent only the number of undetected violators. This is done by setting all the expected damages to one. In this case, the monitoring resources will be allocated to those sources whose decrease in the probability of not detecting a violation, per unit dollar, is greatest.

Examples of the output of this stage of the Resource Allocation Program, for a hypothetical example, are given in Figures 4.3, 4.4, and 4.5. Figure 4.3 gives the initial allocation of resources along with the resources used and the "cost" of undetected violations after allocating the samples. The initial allocation is based on subjective factors such as a desire to monitor sources of a certain size at least once, or a desire to monitor certain sources in a region where water quality is known to be bad. Figure 4.4 shows the marginal return and the decrease in "cost" of undetected violations as the resources are increased. The list is ordered by the marginal return, or equivalently by the priority of monitoring the sources. The first source on the list should be monitored with highest priority, the second source with next highest priority, etc. Thus, given a limit on total resources or a maximum allowed cost of undetected violations, Figure 4.4 contains all the information needed to obtain the priorities. Figures 4.5a and 4.5b show the "Final Allocation" table for this example. In Figure 4.5a a budget limit is given, while in Figure 4.5b a maximum "cost" of undetected violation is specified.

INITIAL ALLOCATION.

| SOURCE | TIMES SAMPLED | RESOURCES USED |
|---|---|---|
| 1--JONES MANUFACTUR | 1 | 535.50 |
| 2--SAFE CHEMICAL CO | 1 | 548.00 |
| 3--SEWAGE TREATMENT | 1 | 560.00 |
| 4--NUMBERNUN CO. | 1 | 555.00 |

TOTAL RESOURCES USED 2198.50
COST OF UNDETECTED VIOLATIONS 4.27171

Figure 4.3 Initial Allocation Table

## PRIORITY LIST OF SAMPLES



| PRIORITY | SOURCE SAMPLED | MARGINAL RETURN X100 | COST OF UNDETECTED VIOLATIONS | RESOURCES REQUIRED |
|---|---|---|---|---|
| 1 | 1--JONES MANUFACTUR | .10774492 | 5.07571 | 535.50 |
| 2 | 3--SEWAGE TREATMENT | .09326524 | 4.55342 | 1095.50 |
| 3 | 3--SEWAGE TREATMENT | .07989130 | 4.10603 | 1655.50 |
| 4 | 1--JONES MANUFACTUR | .06899248 | 3.73658 | 2191.00 |
| 5 | 3--SEWAGE TREATMENT | .06843515 | 3.35334 | 2751.00 |
| 6 | 3--SEWAGE TREATMENT | .05862177 | 3.02506 | 3311.00 |
| 7 | 3--SEWAGE TREATMENT | .05021559 | 2.74385 | 3871.00 |
| 8 | 4--NUMBERWUN CO. | .04526206 | 2.49264 | 4426.00 |
| 9 | 1--JONES MANUFACTUR | .04417806 | 2.25607 | 4981.50 |
| 10 | 3--SEWAGE TREATMENT | .04301486 | 2.01519 | 5521.50 |
| 11 | 3--SEWAGE TREATMENT | .03684665 | 1.80885 | 6081.50 |
| 12 | 3--SEWAGE TREATMENT | .03156296 | 1.63209 | 6641.50 |
| 13 | 1--JONES MANUFACTUR | .02828661 | 1.48061 | 7177.00 |
| 14 | 3--SEWAGE TREATMENT | .02703693 | 1.32920 | 7737.00 |
| 15 | 3--SEWAGE TREATMENT | .02315992 | 1.19951 | 8297.00 |
| 16 | 1--JONES MANUFACTUR | .01811409 | 1.10251 | 8832.50 |
| 17 | 1--JONES MANUFACTUR | .01159902 | 1.04039 | 9368.00 |
| 18 | 1--JONES MANUFACTUR | .00742722 | 1.00062 | 9903.50 |
| 19 | 4--NUMBERWUN CO. | .00590254 | .96786 | 10456.50 |
| 20 | 2--SAFE CHEMICAL CO | .00556719 | .93735 | 11006.50 |
| 21 | 1--JONES MANUFACTUR | .00475586 | .91168 | 11542.00 |
| 22 | 2--SAFE CHEMICAL CO | .00412025 | .88931 | 12090.00 |
| 23 | 2--SAFE CHEMICAL CO | .00304938 | .87260 | 12638.00 |
| 24 | 1--JONES MANUFACTUR | .00304534 | .85629 | 13173.50 |
| 25 | 2--SAFE CHEMICAL CO | .00225683 | .84392 | 13721.50 |
| 26 | 1--JONES MANUFACTUR | .00195003 | .83348 | 14257.00 |
| 27 | 2--SAFE CHEMICAL CO | .00167027 | .82432 | 14805.00 |
| 28 | 2--SAFE CHEMICAL CO | .00123616 | .81755 | 15353.00 |
| 29 | 2--SAFE CHEMICAL CO | .00091488 | .81254 | 15901.00 |
| 30 | 4--NUMBERWUN CO. | .00076974 | .80826 | 16456.00 |
| 31 | 2--SAFE CHEMICAL CO | .00067710 | .80455 | 17004.00 |
| 32 | 2--SAFE CHEMICAL CO | .00050112 | .80181 | 17552.00 |
| 33 | 2--SAFE CHEMICAL CO | .00037087 | .79976 | 18100.00 |
| 34 | 4--NUMBERWUN CO. | .00010038 | .79922 | 18655.00 |
| 35 | 4--NUMBERWUN CO. | .00001309 | .79915 | 19210.00 |
| 36 | 4--NUMBERWUN CO. | .00000171 | .79914 | 19765.00 |
| 37 | 4--NUMBERWUN CO. | .00000022 | .79914 | 20320.00 |
| 38 | 4--NUMBERWUN CO. | .00000003 | .79914 | 20875.00 |
| 39 | 4--NUMBERWUN CO. | .00000000 | .79914 | 21430.00 |
| 40 | 4--NUMBERWUN CO. | .00000000 | .79914 | 21985.00 |

Figure 4.4 Priority List of Samples

FINAL ALLOCATION

BUDGET 10000.00

| SOURCE | MIN NO. OF SAMPLES REQUIRED | MAX NO. OF SAMPLES ALLOWED | TIMES SAMPLED | RESOURCES USED | COST OF UNDETECTED VIOLATIONS |
|---|---|---|---|---|---|
| 1--JONES MANUFACTUR | 1 | 10 | 6 | 3213.00 | .11058 |
| 2--SAFE CHEMICAL CO | 1 | 10 | 1 | 548.00 | .08667 |
| 3--SEWAGE TREATMENT | 1 | 10 | 10 | 5600.00 | .77476 |
| 4--NUMBERNUN CO. | 1 | 10 | 1 | 555.00 | .03767 |

TOTAL RESOURCES USED 9916.00
FINAL COST OF UNDETECTED VIOLATIONS 1.00988

Figure 4.5a Final Allocation Given Monetary Budget

FINAL ALLOCATION

MAXIMUM ALLOWED COST OF UNDETECTED VIOLATIONS 1.00000

| SOURCE | MIN NO. OF SAMPLES REQUIRED | MAX NO. OF SAMPLES ALLOWED | TIMES SAMPLED | RESOURCES USED | COST OF UNDETECTED VIOLATIONS |
|---|---|---|---|---|---|
| 1--JONES MANUFACTUR | 1 | 10 | 7 | 3748.50 | .07081 |
| 2--SAFE CHEMICAL CO | 1 | 10 | 1 | 548.00 | .08687 |
| 3--SEWAGE TREATMENT | 1 | 10 | 10 | 5600.00 | .77476 |
| 4--NUMBERWUN CO. | 1 | 10 | 1 | 555.00 | .03767 |

TOTAL RESOURCES USED 10451.50
FINAL COST OF UNDETECTED VIOLATIONS .97011

Figure 4.5b Final Allocation Given Maximum Allowed "Cost" of Undetected Violations

Update Statistics

After monitoring the sources over the monitoring period, new compliance monitoring and self-monitoring data become available. These data are then used in determining the priorities for the next monitoring period. The statistical descriptions (i.e. mean and standard deviation) of the effluent constituents can be updated to include this new information. Upon updating the statistics, the compliance monitor is ready to repeat the priority setting procedure so as to obtain the sampling frequencies for the next monitoring period.

Detailed examples are presented in Sections VIII.2 and IX illustrating the use of the Resource Allocation Program.

# SECTION V

# STATISTICAL CHARACTERISTICS OF EFFLUENT STREAMS

The priority setting procedure for compliance monitoring requires that the daily composite effluent loads, due to their inherent variability, be modeled statistically. Among the questions that must be addressed in developing a statistical model are:

- What probability distributions adequately model the effluent data?
- What is the statistical correlation between the various constituents of the effluent from a source?
- What is the time-varying nature of the statistics?

Section V.1 shows, for several example sets of data, that the normal and lognormal distributions adequately model the statistics of the daily composite effluent loadings. In order to decide whether to model a particular constituent by a normal or lognormal distribution, it is necessary to process a large amount of daily data. It is not expected that the individual monitoring agency will have the resources to analyze the daily data of each source in its jurisdiction. It is only postulated that the monitoring agency will have a monthly mean and maximum for each constituent of each source in its jurisdiction. It is only postulated that the monitoring which distribution can be associated with a given industrial process. Since this information is unavailable at the publication of this report, several guidelines are specified on how to choose between the normal and lognormal cases.

The normal and lognormal distributions are parameterized by a mean and a standard deviation. (For the lognormal distribution, the mean and standard deviation are of the logs of the data.) Since it is only assumed that the monthly mean and maximum, and not the sample standard deviation, are available to the monitor, the standard deviation of the normal process has to be estimated using nonstandard estimation procedures. The situation is more complicated for the lognormal case, since neither the sample mean of the logs of the data nor the sample standard deviation of the logs of the data are available. Appendix A develops approximate maximum likelihood estimates of the mean and standard deviation from the sample mean and maximum of the data for both the normal and lognormal cases. These estimates are tested on real data in Section V.1 to show that they, coupled with the associated distributions, adequately describe the statistical variations. The case is slightly more complicated for pH. The data for pH available to the monitor will include a maximum and a minimum monthly value and possibly a mean monthly value. If a mean value is given, the pH can be modeled by a mean and two standard deviations - one based on the mean and the maximum, the other based on the mean and the minimum. The estimates of the standard deviations are based on the procedures just discussed. The resulting density function has a shape shown in Figure 5.1. If a mean is not given, the mean and a single standard deviation can be estimated from the minimum and the maximum. This estimation procedure is also given in Appendix A.

There has been little study into the statistical correlation of the constituents of an effluent. As with the problem of determining the appropriate distributions, it is not expected that the monitoring agency would be able to determine the correlation of the constituents of the sources in its jurisdiction. It is therefore necessary that the correlation coefficients be obtained from industry-wide studies. Since these are unavailable at the present time, it is assumed, unless other knowledge is available, that the constituents from a source are uncorrelated. The priority setting procedure also allows for the case where the constituents are completely correlated. In Appendix B, a correlation study for a single municipal treatment plant is carried out. It is clear that no general conclusions can be reached from the analysis of one water treatment plant. The analysis has shown the variability in the correlation parameters from month to month and the problems inherent in choosing

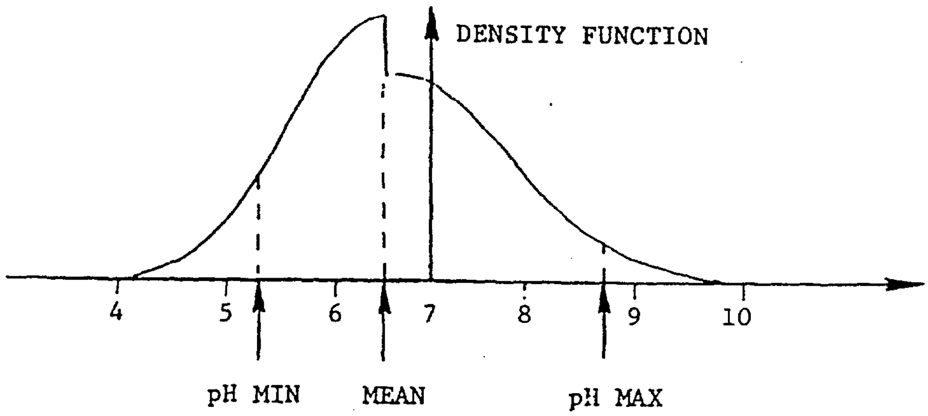


Figure 5.1 Example of probability density function of pH.

between the hypotheses of uncorrelated constituents and correlated constituents.

The time-varying nature of the statistics comes from two sources: (1) periodic variations due to weekly, monthly, or seasonal variations and (2) trends due to changes in the plant processes. The weekly and monthly variations are averaged out in the input data (i.e., monthly mean and maximum). These variations if known, should be taken into account when determining when, in a monitoring period, to monitor a particular source. The seasonal variations and trends are taken into account in the statistical characterization by discounting past information and updating the statistics as new data become available.

The specific procedures used in the Resource Allocation Program to obtain the initial statistical description of the effluent sources and to update the statistics as new information becomes available are discussed in Sections V.2 and V.3 respectively.

## V.1 CHOICE OF DISTRIBUTION

### Testing for Distribution Acceptability

This subsection addresses the problem of what probability distribution or distributions are appropriate to describe the inherent variability of effluent constituents. Based upon previous studies [1], [2] as well as operational considerations (i.e., implementation feasibility), the normal and lognormal distributions have teen chosen as candidates. A statistical testing procedure [3], namely the Kolmogorov-Smirnov (K-S) test, is used to test whether it is "acceptable" to consider the effluent data as being described by a certain probability distribution.

The statistical test whether to accept the "null hypothesis" ($H_o$), that the distribution is normal (or lognormal), is subject to a given probability of error of rejecting $H_o$ when it is true. This probability of error, denoted $\alpha$, is called the "level of significance" of the test.

If $H_o$ is accepted when $\alpha$, the allowed probability of incorrectly rejecting $H_o$, is large, then the probability that $H_o$ is true is high. The Kolmogorov-Smirnov test compares the deviations of the empirical probability distribution from the assumed distribution. The smaller the largest observed deviation, the higher is the "significance" of the null hypothesis, i.e., that the observed variables come from the assumed distribution.

The Kolmgorov-Smirnov test will now be applied to Palo Alto Municipal Waste Treatment Plant data* to determine whether the normality or log-normality assumption can be accepted, and at what level of significance. The test is done on BOD, suspended solids, and coliform data.

The BOD data for July 1973 are considered first. A plot of the observed cumulative distribution and the normal distribution with the sample mean and sample variance appears in Figure 5.2. The distributions are plotted versus

$$\xi = \frac{x - \mu}{\sigma}$$

the deviation of the loading, x, from the mean, $\mu$, normalized by the standard deviation, $\sigma$. In this case the sample mean and standard deviation are used. The solid line is the standard normal distribution, with zero mean and unit standard deviation. The points denoted by "▲" in the figure are the normalized deviations of the measurements from their mean and this is used to test the normality assumption. The lognormality assumption is tested by plotting the normalized deviations of the logs of the measurements from the mean of the logs (denoted "●" on Figure 5.2).

The K-S test determines whether the maximum deviation between the sample distribution and the assumed distribution exceeds the critical value for a given level of significance. The critical values, i.e., the maximum

---

* Data obtained from Palo Alto Wastewater Treatment Plant Automation Project [4].

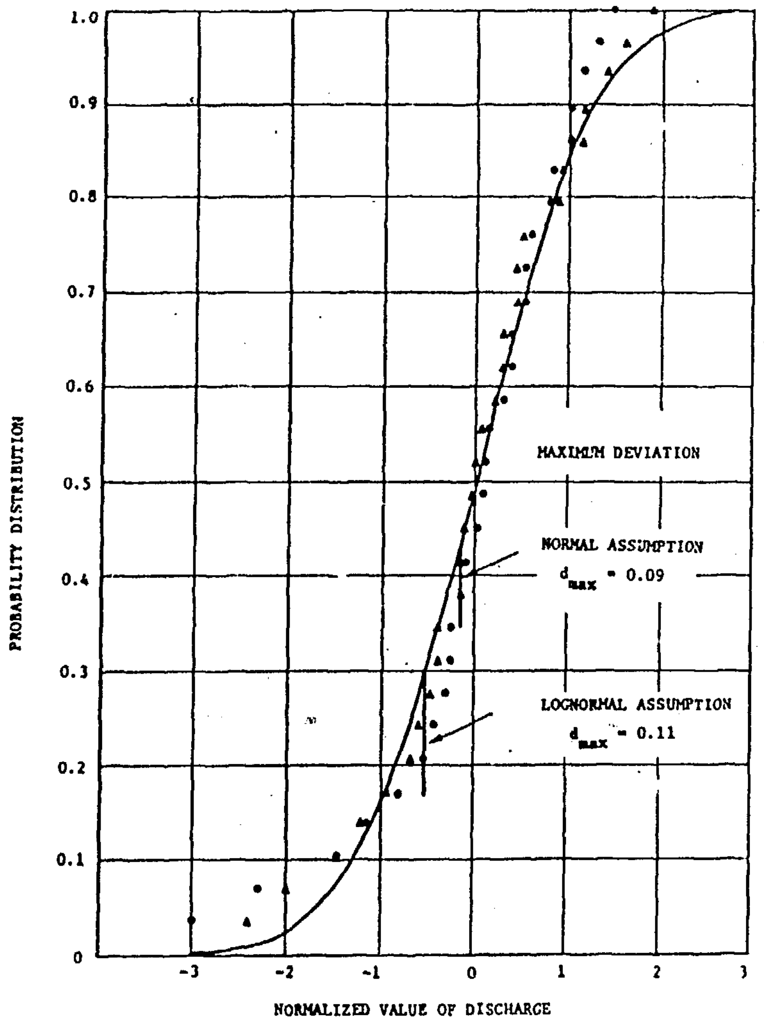


Figure 5.2 Kolmogorov-Smirnov test for BOD data.

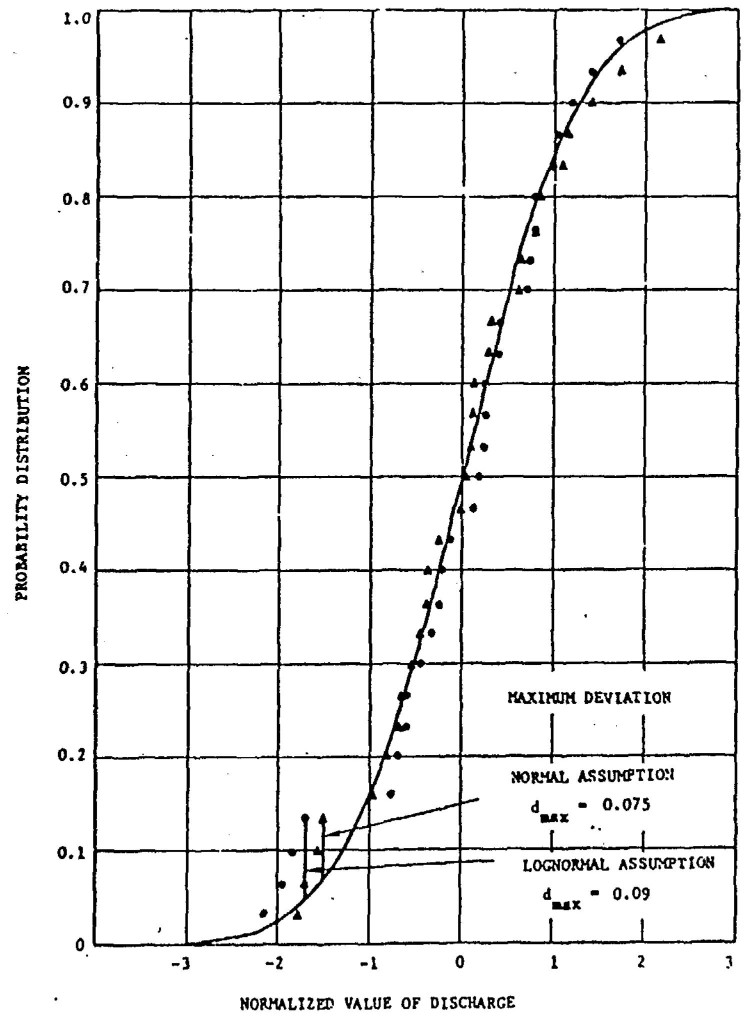


Figure 5.3 Kolmogorov-Smirnov test for suspended solids (dry month).

allowable deviation for a given level of significance for the Kolmogorov-Smirnov test, are shown in Table 5.1. The number of data points used for the plot of Figure 5.2 was 29 and, as can be seen, the maximum deviations are about 0.1 for both normal and lognormal assumptions. This shows that either of the hypotheses is acceptable at a level of significance over 20%, which is quite high.* In the statistical literature it has become customary to use 5% level of significance; thus in the present case the results are more significant than the customary required level for acceptance of $H_o$.

Table 5.1 CRITICAL VALUES, $d_\alpha(N)$, OF THE MAXIMUM ABSOLUTE DIFFERENCE BETWEEN SAMPLE AND POPULATION CUMULATIVE DISTRIBUTIONS [3].

| Sample size (N) | Level of significance ($\alpha$) | | | | |
|---|---|---|---|---|---|
| | 0.20 | 0.15 | 0.10 | 0.05 | 0.01 |
| 5 | 0.446 | 0.474 | 0.510 | 0.565 | 0.669 |
| 10 | 0.322 | 0.342 | 0.368 | 0.410 | 0.400 |
| 15 | 0.266 | 0.283 | 0.304 | 0.338 | 0.404 |
| 20 | 0.231 | 0.246 | 0.264 | 0.294 | 0.356 |
| 25 | 0.21 | 0.22 | 0.24 | 0.27 | 0.32 |
| 30 | 0.19 | 0.20 | 0.22 | 0.24 | 0.29 |
| 35 | 0.18 | 0.19 | 0.21 | 0.23 | 0.27 |
| over 35 | $\frac{1.07}{\sqrt{N}}$ | $\frac{1.14}{\sqrt{N}}$ | $\frac{1.22}{\sqrt{N}}$ | $\frac{1.36}{\sqrt{N}}$ | $\frac{1.63}{\sqrt{N}}$ |

*Since the empirical distribution is compared here to an assumed distribution with estimated rather than true parameters, the actual level of significance is somewhat lower (see Kendall and Stuart [3]).

In the case of the suspended solids data from a dry month presented in Figure 5.3, the largest deviation for both normal and lognormal assumptions are below 0.1. Therefore one can accept either of these assumptions at 20% level of confidence. For a wet month, the suspended solids data, as shown in Figure 5.4, exhibit a large deviation under the normal assumption, but this hypothesis is still acceptable at 15% level of significance; the lognormal assumption is accepted at a level of significance larger than 20%.

A set of 28 coliform measurements (Jan. 1974) are plotted in Figure 5.5 to test their distribution via the Kolmogorov-Smirnov method. Using Table 5.1 it can be seen that the normal assumption is rejected even at a low level of significance of 1%, while the lognormal assumption is accepted at a 15% level.

The conclusion is that, except for coliforms, the normal and lognormal hypotheses are both acceptable. For the coliform data the normal assumption is not adequate because of its rather large range of variability and the skewness of the frequency histogram.

## Fitting of Distributions to Real Data

This subsection compares how well the following statistical assumptions fit the data.

(1) Normal distribution - mean equals sample mean and standard deviation equals sample standard deviation.

(2) Normal distribution - mean equals sample mean and standard deviation estimated from mean and maximum value (obtained with the procedure of Appendix A).

(3) Lognormal distribution - mean of logs equals sample mean of logs, standard deviation of logs equals sample standard deviation of logs.

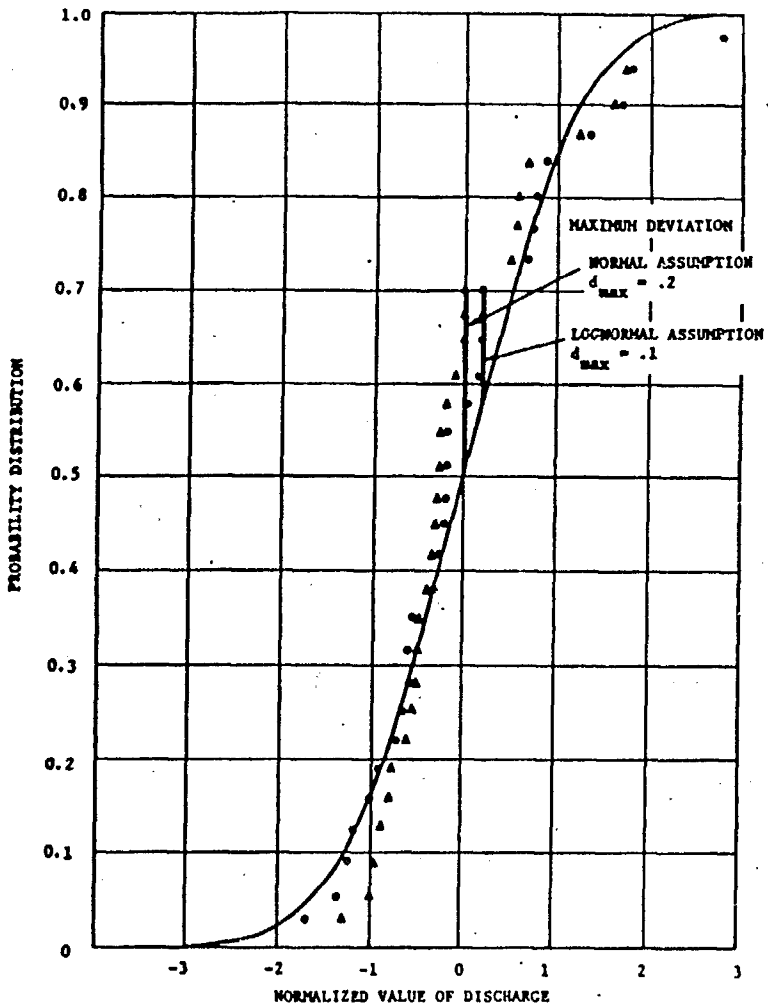


Figure 5.4 Kolmogorov-Smirnov test for suspended solids (wet month).

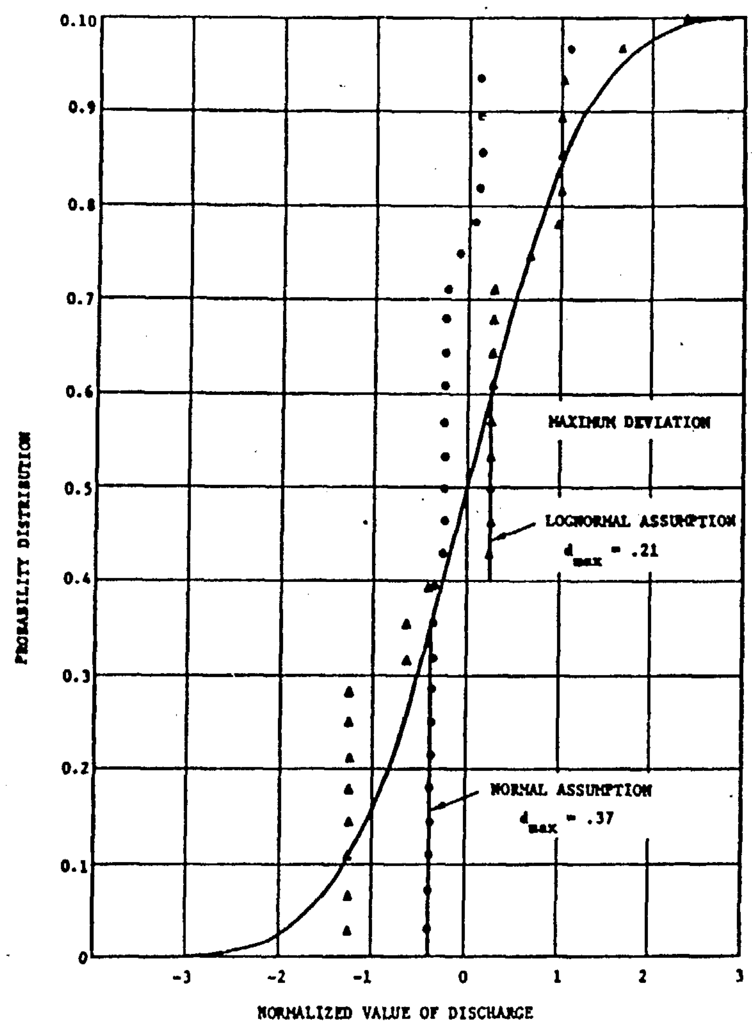


Figure 5.5 Kolmogorov-Smirnov test for coliform data.

(4) Lognormal distribution - mean of logs equals sample mean of logs, standard deviation of logs estimated from mean and maximum value of logs (obtained with the procedure of Appendix A).

To determine which distribution fits the data best, the data are plotted on normal probability paper. The normal distribution then appears as a straight line. The lognormal assumption is also a straight line if the distribution of the logs of the data is plotted. This technique was used, as opposed to the more sophisticated tests such as the K-S test, for the following reasons:

- It gives a simple visual test of the various assumptions.
- It can be easily used to determine if the data agree with the assumed distribution for large values of the constituent (i.e., at values where a violation or damage will occur).

This procedure is demonstrated on daily data of both effluent concentration and effluent loadings over either a six-month or twelve-month period for the non-fertilizer phosphorus chemicals industry and the inorganic chemicals, alkali, and chlorine industries [5], [6]. Table 5.2 describes the various cases and includes the sample mean, sample standard deviation, maximum and estimated standard deviation (from mean and maximum). (For the lognormal cases, the statistics are of the logs of the data.) The figures are distributions plotted on normal probability paper so that a normal process will lie close to a straight line. The data are normalized so that the sample mean equals zero and the sample standard deviation equals one. For each case, the following two normal distributions are compared: the means are equal to the sample mean for both cases; the standard deviation for one case is equal to the sample standard, deviation, and for the other case, is equal to the estimated standard deviation from the sample mean and maximum. Figures 5.6a through 5.12a

Table 5.2 DESCRIPTION OF EXAMPLE CASES.

| Figure | Plant # | Constituent | Time period (months) | Normal | Lognormal | Mean | Stan. dev. | Max | Est. stan. dev. |
|---|---|---|---|---|---|---|---|---|---|
| 5.6b | 159 | $PO_4$(ppm) | 12 | X | | 24.0 | 3.40 | 34.5 | 3.31 |
| 5.6c | | | | | X | 1.38 | .063 | 1.54 | 0.56 |
| 5.7b | 159 | $PO_4$ (kg/day) | 12 | X | | 235 | 54.6 | 370 | 48.0 |
| 5.7c | | | | | X | 2.36 | .110 | 2.57 | .077 |
| 5.8b | 030 | Cl-ion (kg/day) | 6 | X | | 3060 | 1220 | 6450 | 1290 |
| 5.8c | | | | | X | 3.44 | 2.21 | 3.81 | 2.14 |
| 5.9b | 144 | Hg ($10^{-3}$ ppm) | 6 | X | | 3.78 | 2.05 | 10.5 | 2.55 |
| 5.9c | | | | | X | .511 | .249 | 1.02 | .194 |
| 5.10b | 144 | Hg ($10^{-3}$ kg/day) | 6 | X | | 17.6 | 10.5 | 47.0 | 11.1 |
| 5.10c | | | | | X | 1.16 | .306 | .167 | .193 |
| 5.11b | 144 | Cl-ion ($10^2$ ppm) | 6 | X | | 6.50 | 6.80 | 29.5 | 8.7 |
| 5.11c | | | | | X | .544 | .529 | 1.47 | .352 |
| 5.12b | 144 | Cl-ion ($10^3$kg/day) | 6 | X | | 3.26 | 3.47 | 16.5 | 5.00 |
| 5.12c | | | | | X | .267 | .494 | 1.22 | .360 |

show the corresponding histograms. From these histograms, we see that the density functions in Figures 5.6a, 5.7a, and 5.8a are of the normal shape, the density function is Figures 5.9a and 5.10a are somewhat of a normal shape, and the density functions in Figures 5.11a and 5.12a are far from normal. Examining the figures, the following conclusions can be drawn:

- The data in Figures 5.6, 5.7, and 5.8 fall closer to the normal (as opposed to lognormal) distribution. Good fit to the data is obtained, under the normal assumption, by using either estimate of the standard deviation.

- The data in Figures 5.9 and 5.10 are fit equally well by the normal or lognormal distributions. The fit to the data using the estimated standard deviation (from mean and maximum) is better for large values of the constituent than the fit obtained using the sample standard deviation.

- The data in Figures 5.11 and 5.12 fit the lognormal distribution better.

From these few examples, it is not possible to make any general statements assigning either normal or lognormal distributions to an industry or a constituent. However, the following tentative conclusions can be made:

- The normal and lognormal distributions with the standard deviation estimated from the mean and maximum give a good fit to the data for many cases.

- A better fit for large values of constituent is obtained when the standard deviation is estimated from the mean and maximum as opposed to using the sample standard deviation.

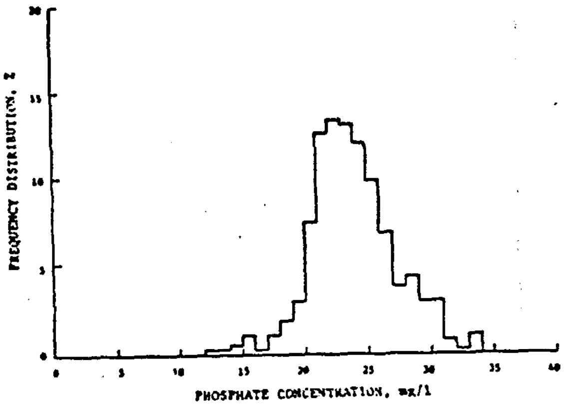


Figure 5.6a Histogram of phosphate concentration data at plant 144.

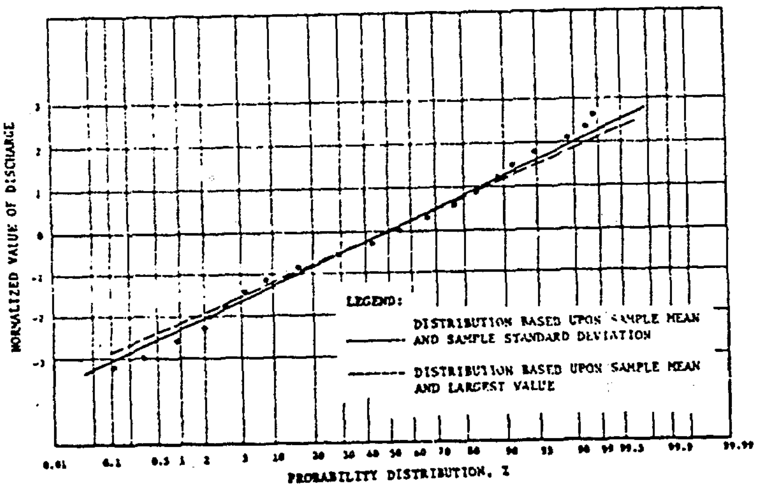


Figure 5.6b Normal.

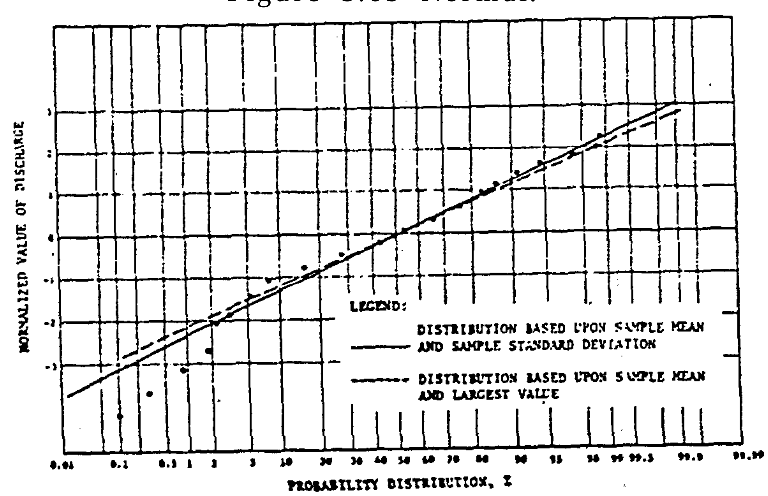


Figure 5.6c Lognormal.

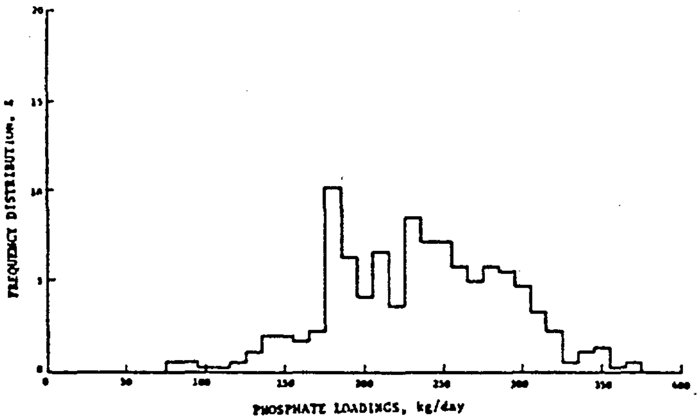


Figure 5.7a Frequency distribution of effluent phosphate daily discharge at plant 159.

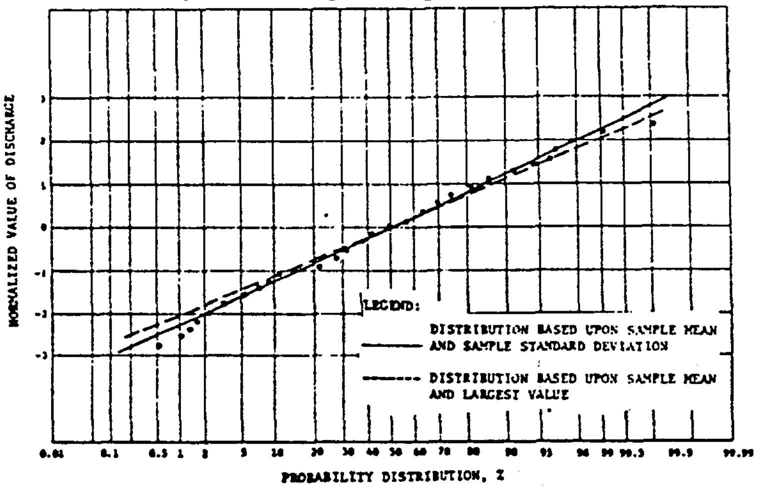


Figure 5.7b Normal.

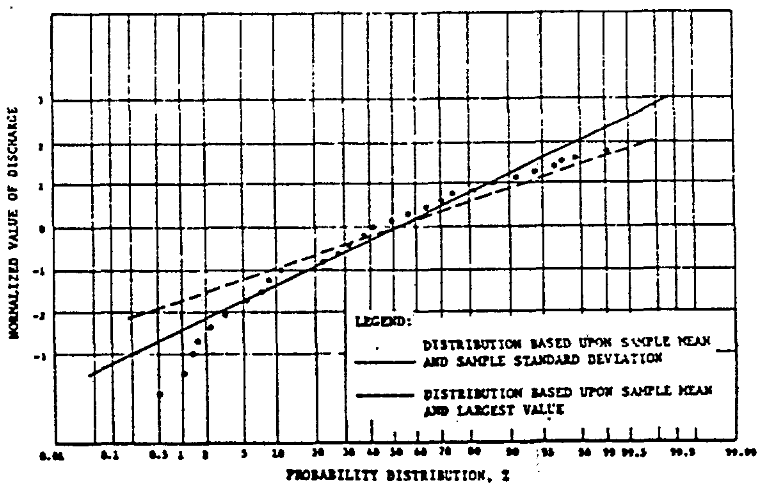


Figure 5.7c Lognormal.

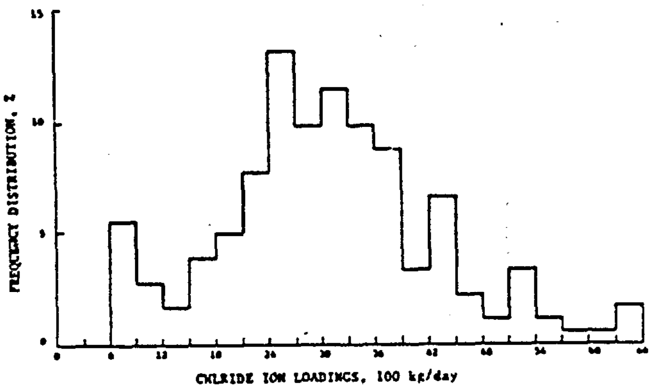


Figure 5.8a Frequency distribution of effluent chloride ion discharge at plant 030.

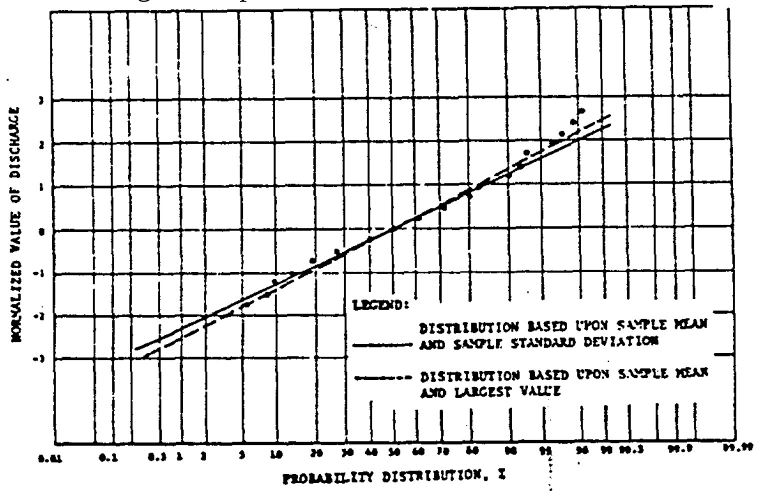


Figure 5.8b Normal.

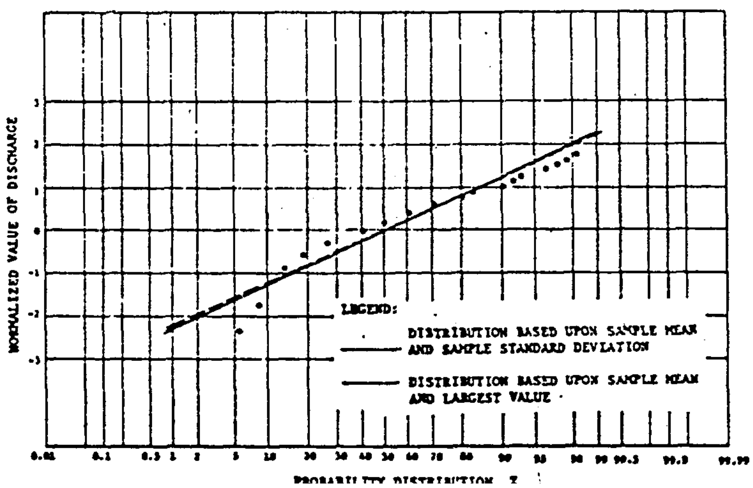


Figure 5.8c Lognormal.