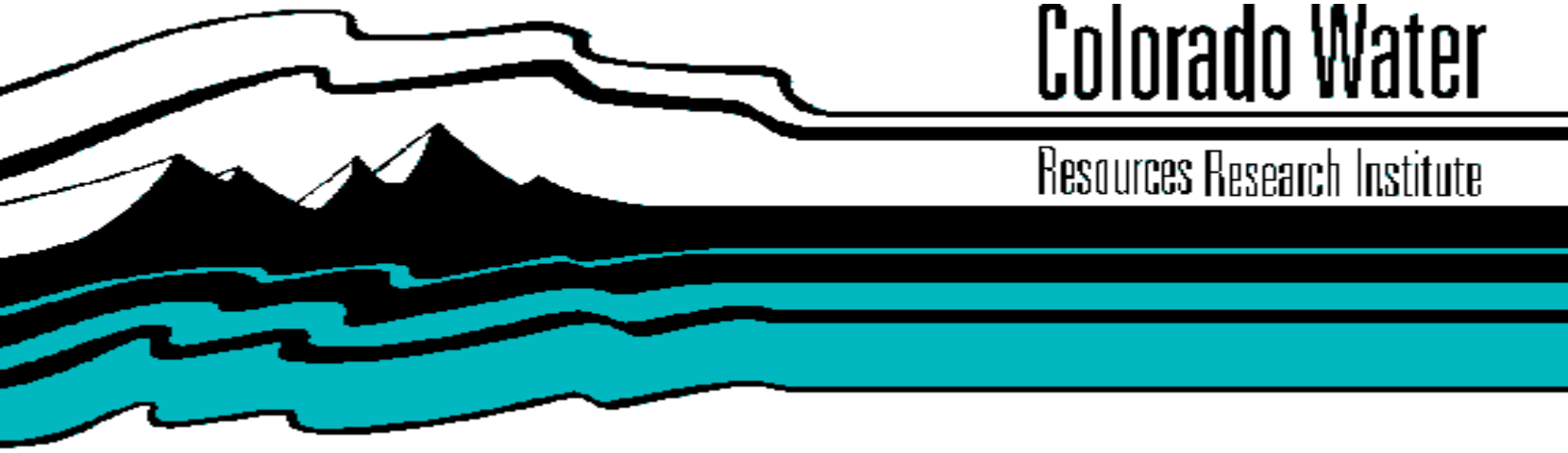
DETECTION OF WATER QUALITY CHANGES THROUGH OPTIMAL TESTS AND RELIABILITY OF TESTS

by

Roy W. Koch, Thomas G. Sanders, and Hubert J. Morel-Seytoux



Completion Report No. 104

DETECTION OF WATER QUALITY CHANGES THROUGH OPTIMAL TESTS AND RELIABILITY OF TESTS

Completion Report

OWRT Project No. B-186-COLO

by

Roy W. Koch
Thomas G. Sanders
Hubert J. Morel-Seytoux

Department of Civil Engineering

Colorado State University

submitted to

Office of Water Research and Technology

U.S. Department of the Interior
Washington, D.C. 20240

September, 1980

The work upon which this report is based was supported (in part) by funds provided by the United States Department of the Interior, Office of Water Research and Technology, as authorized by the Water Resources Act of 1978, and pursuant to Grant Agreement No.(s) 14-34-0001-8069.

COLORADO WATER RESOURCES RESEARCH INSTITUTE
Colorado State University
Fort Collins, Colorado

Norman A. Evans, Director

ABSTRACT

The detection of change in a hydrologic variable, particularly in water quality, is a current problem. A method of formulating this problem in a mathematical programming context is presented. The method is based on using weighted linear combinations of water quality variables from different locations with the weighting factors being adjusted so that the time required to detect the change is minimized. The basis of the technique, then, is a trade-off of time by adding information from other locations. The results of example applications show that significant savings in time can be achieved by using this method.

Since the detection method is based on sample statistics developed from historic data, uncertainty exists as to how accurately the optimal values of the time required to detect the change and the weighting factors reflect the true, but unknown, values. A method of evaluating the reliability of these estimates is presented using an analytical solution of the optimization problem. Through this approach, expressions are obtained relating explicitly the optimization variables to the random variables of the problem giving a clear picture of the interrelationships. Simulation is then used to evaluate the behavior of the optimization variables through these explicit relations.

TABLE OF CONTENTS

TABLE OF CONTENTS
(continued)

LIST OF FIGURES

LIST OF FIGURES
(continued)

LIST OF TABLES

LIST OF TABLES
(continued)

INTRODUCTION

The time variation of hydrologic variables such as water discharge or water quality requires that they be treated as, at the very least, random variables and possibly as stochastic processes when describing possible future occurrences. As a result of this characteristic, the detection of a change in one of these variables requires consideration of this random nature and the laws of probability must be applied to the problem. It is the purpose of this study to develop a test for detection of changes in a hydrologic variable, particularly applied to water quality. Further, since our knowledge of the future is incomplete, such a test can only be optimal for the sample upon which it is based. These uncertainties will be investigated in terms of how they affect the reliability of the test.

Detection of Changes

In general, the ability to detect a change in a hydrologic variable reduces to a statistical problem depending in major part on the variability of the variable of interest. In a statistical sense, the number of observations dictates the level of change that can be detected and, conversely, the degree to which we would like to be able to detect a change dictates the number of observations required. In a practical sense, when related to the hydrologic problem, the number of observation is translated into time, i.e., days, months or years and often becomes the major factor constraining the problem.

Planning and management decisions required of an agency or individual must often be made based on short data records as a result of economic and political pressures. Thus, it is necessary to find methods of evaluating hydrologic change within the shortest time horizon to

respond to these pressures and provide reliable results. This problem can be illustrated in many cases, particularly when water quality is concerned. Significant land use changes such as surface mining may have an effect on the water quality for certain beneficial uses downstream. It is necessary that any changes which occur be quantified within a short time period in order to adjust any regulations on future operations. Similarly, when treatment of municipal pollutants is considered, the degree to which our past efforts have changed the water quality should be evaluated before proceeding to newer or higher levels of treatment facilities. Any practical methodology for evaluating hydrologic changes could then be applied to many timely planning and management problems.

Reliability of the Detection Method

Since no definitive statements can be made as to the future response of hydrologic systems due to the inability to characterize the inputs deterministically and the wide spatial variability of the processes, they must be viewed at least in part as random variables. This introduces a level of uncertainty in the assessment of the future values of the variable. Any statistical test will be based on certain characteristics of the past, recorded history of the variable. Thus, any test will be only as good as the degree to which the sample reflects the actual level and variation of the variable. In addition, the applicability of the test will depend on how closely the variable agrees with any assumptions made in deriving the test such as underlying distributions and/or time dependent structure. It is then necessary to evaluate the test in light of these known, but necessary, shortcomings for an idea of its reliability.

DETECTION OF CHANGE IN A WATER QUALITY VARIABLE

The methodology to be applied in detecting a change in a hydrologic variable is one which is adapted from Morel-Seytoux and Saheli (1973). This approach is a generalization of the more traditional statistical approaches to the problem but, in theory, works in much the same way. So, prior to developing this method, a brief review of some of the simpler methods is presented to introduce the concepts behind the basic problem of detecting a change in a random variable.

Before presenting the underlying theory, however, it is necessary to define the detection problem more precisely. There are any number of characteristics of a hydrologic variable which could change due to environmental changes. However, in this particular case, attention will be focused on the most basic statistic, the mean value of the variable. Also, in general, the annual values will be used.

Traditional Methods for Detection of Change

The problem of detecting a change in the mean of a random variable is a hypothesis testing problem, the test being whether or not a sample mean belongs to some particular parent population. Due to its ease of use and the abundance of theory associated with it, the normal distribution is applied to these problems whenever possible. Often, for hydrologic applications, either the original variable or some relatively simple transformation of it will be approximately normally distributed and thus this body of theory can be employed.

For a single random variable, X, normally distributed with mean, μ_x and standard deviation σ_X, a test variable for a one-sided test of size α (or level of significance) for the sample mean, $\bar{X}$, is

$$Z = \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \tag{1}$$

where n is the number of observations.

The test proceeds as follows:

1) observe a sample of length n and compute $\bar{X}$,

2) compute Z from Equation (1),

3) compare Z with $z_{1-\alpha}$, the critical value of the test variable,

4) if $Z \leq z_{1-\alpha}$ then accept the hypothesis that the sample mean is from the same population with mean, μ_X.

Now, if the value of the test variable Z, is equal to the critical value and a selected change in the mean, $k\mu_X$ is chosen so that the sample mean, $\bar{X}$ is equal to the population mean plus the specified change, $\mu_X + k\mu_X$, then there are two unknowns in the problem, the sample size n and the fractional change, k. Making these substitutions into (1) and solving for n gives:

$$n = \left(\frac{z_{1-\alpha}\ \sigma_X}{k\mu_X}\right)^2 \tag{2}$$

Noting that the coefficient of variation, C_v is the ratio of the standard deviation to the mean, Equation (2) can be written as:

$$n = \left(\frac{z_{1-\alpha}}{k}\right)^2\ C_v^2 \tag{3}$$

Thus, the number of observations (years in the context of this study) required to detect a change in the mean is inversely proportional to the square of the fraction of change to be detected and directly proportional to the square of both the critical value of the test variable,

$z_{1-\alpha}$, an indication of the accuracy of the test, and the coefficient of variation, a measure of the degree of variation of the variable relative to the mean value. Thus, either a small change ($k << 1$) or a large variation in the variable (C_v large) will lead to a large sample size required to detect the change. Certain changes are necessary when the population parameters are not known. However, this is the essence of the problem. If the coefficient of variation could be decreased, then the test would perform better, that is fewer observations would be required to detect a change of a given level.

A sophistication of the univariate case is to use two variables which are statistically correlated. If this is the case, given some information about a variable, X, then we have some information about the other variable, Y, as a result of this correlation. If a linear relation can be assumed between X and Y and both are normally distributed then regression theory based on the bivariate normal distribution can be applied to the problem.

A relation similar to (2) can be developed as:

$$n = \left(\frac{z_{1-\alpha} \, \sigma_{Y/X}}{k\mu_Y} \right)^2 \tag{4}$$

In this case, however, the conditional variance, $\sigma^2_{Y/X}$ is used. This is given by the expression

$$\sigma^2_{Y/X} = (1-\rho^2_{XY}) \, \sigma^2_Y \tag{5}$$

where ρ_{XY} is the correlation coefficient between X and Y.

Since ρ_{XY} varies between -1 and 1, the conditional variance is always smaller than the marginal variance. Thus, it would be expected that the number of observations (or years of record) computed from Equation (4) would be less than that from Equation (2).

Although it has not been considered to this point and is usually ignored in simple applications, there is one additional aspect of the hypothesis testing problem that should be mentioned: the power of the test. The size of the test, α, indicates the probability of rejecting the hypothesis that the mean from the new sample is the same as the population mean when they are equal, called the Type I error. There is, however, another error which must be considered, the probability of accepting the mean of the new sample as being equal to the population mean when it is not: the type II error, usually signified by β. The complement of this probability, $1-\beta$, is the power of the test (see Figure 1). So, as the size is decreased the power also decreases. The size and power should both be specified before performing the test, however it is necessary to know the alternate value of the mean in order to set the power.

The major aspects of the traditional hypothesis testing problem can be summarized as follows:

1) For a given level of change in the mean, there is a certain sample size required to detect this change for a population with a given coefficient of variation.

2) As the magnitude of the change decreases, a larger sample size is required to detect it for the same coefficient of variation.

3) For a given level of change, if the coefficient of variation is smaller, the number of observations required to detect the change is smaller.

These principles have been employed in the formulation of the test to be described below.

Design of a Test for Detection of Change

Using the concepts of traditional hypothesis testing as a starting point, Morel-Seytoux and Saheli (1973) have developed a test for the

α Size of the Test (Level of significance)

$1-\beta$ Power of the Test, P [Accept H_1 | H_1 True]

x_c Critical Value of the Variable x

μ_0 Mean of the Null Hypothesis, H_0

μ_1 Mean of the Alternate Hypothesis, H_1

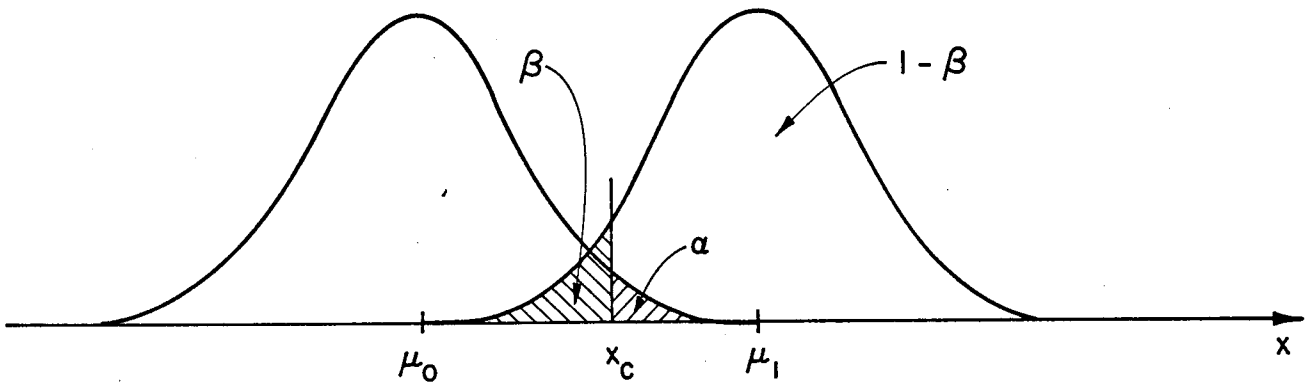


Figure 1. Graphical Representation of the Statistical Theory.

detection of a change in annual snowmelt runoff due to weather modification. This basic approach has been adopted in a slightly modified manner in this study for the problem of the detection of changes in water quality variables.

The test considered in the remainder of this discussion is based on the target-control concept, that is, the regression type approach discussed above. In addition, a regional approach is taken where variables are formed as weighted linear combinations of the annual values at various locations both for a target and a control area. Using this approach, an attempt is made to take advantage of the information available from a number of locations. Further, although a large number of locations (or variables) appear in the original data, they are reduced to only two, the weighted linear combinations in both the target and control areas, which allows the use of the well developed body of theory regarding the bivariate normal distribution. Thus, a multivariate problem becomes a bivariate problem. It is then reasonable to assume that a judicious selection of the weighting factors can be made which would result in a decrease in the coefficient of variation in this linear combination and provide a better test, that is one which would be able to detect a given change in a shorter period of time. Selection of the weighting coefficients then becomes the problem.

Based on the assumptions that the sample is time independent and normally distributed, the following equation can be derived (see Appendix A) for the number of observations (years in this case) necessary to detect a change of 100 k percent in the mean with size α and power $1-\beta$:

$$N = (z_{1-\alpha} + z_{1-\beta})^2 \left(\frac{\sigma_{Y/X}}{k\mu_Y}\right)^2 \qquad (6)$$

where N is the number of years required to detect the change in the mean

$z_{1-\beta}$ is the standard normal deviate defined by $P[Z \leq z_{1-\beta}] = 1-\beta$

$z_{1-\alpha}$ is the standard normal deviate defined by $P[Z \leq z_{1-\alpha}] = 1-\alpha$

$\sigma_{Y/X}$ is the conditional variance of the variable Y given X

k is the fractional change in the mean

μ_Y is the mean of the variable Y.

It is worth noting that this expression, unlike the simple cases presented earlier, includes the power of the test.

Now, rather than single variables X and Y, weighted linear combinations of variables will be used in the expression as

$$Y^* = \sum_{i=1}^{NT} w_i Y_i \qquad (7)$$

$$X^* = \sum_{i=NT+1}^{NT+NC} w_i X_i \qquad (8)$$

where w_i is the weighting factor

X_i is the value of the variable (annual) in the control area

Y_i is the value of the variable in the target area

NT is the number of stations in the target area, and

NC is the number of stations in the control area.

Recalling the expression for the conditional variance given in Equation (5), the expression in Equation (6) can be written for the linear combinations as:

$$N = (z_{1-\alpha}+z_{1-\beta})^2 \ (1-\rho^2_{X^*Y^*}) \left(\frac{\sigma_Y^*}{k\mu_Y^*}\right)^2 \tag{9}$$

Finally, using the expressions for $\rho_{X^*Y^*}$, σ_Y^* and μ_Y^* in expanded form (see Appendix A) results in the equation

$$N = \left(\frac{z_{1-\alpha} \quad z_{1-\beta}}{k \sum_{i=1}^{NT} w_i\mu_{Y_i}}\right)^2 \left[\sum_{i=1}^{NT}\sum_{j=1}^{NT} w_i w_j \ \mathrm{Cov}(Y_iY_j)\right. \tag{10}$$

$$\left. - \left(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \ \mathrm{Cov}(Y_iX_k)\right)^2 \Big/ \sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NC} w_k w_\ell \ \mathrm{Cov}(X_kX_\ell)\right]$$

where $\mathrm{Cov}(Y_iY_j)$ is the covariance between Y_i and Y_j.

With this expression in terms of the weighting factor and parameters of the variables (means and covariance), a method of selecting the weighting factors can be devised. Morel-Seytoux and Saheli (1973) minimized this expression which effectively minimizes the number of observations (time in this case) necessary to detect the selected percent change, k, with respect to the weighting factors, w_i, i = 1, 2, ... NT + NC.

Rather than proceed with an unconstrained problem, however, Morel-Seytoux and Saheli (1973) introduced the following constraints:

$$\sum_{i=1}^{NT} w_i\mu_{Y_i} = \sum_{i=1}^{NT} \mu_{Y_i} = \mu_Y^* \tag{11}$$

$$\sum_{i=NT+1}^{NT+NC} w_i\mu_{X_i} = \sum_{i=NT+1}^{NC} \mu_{X_i} = \mu_X^* \tag{12}$$

These constraints force the expected value of the linear combination to be equal to the expected value of the simple sum of the mean values at the various locations which specified at least one point on the distribution of each linear combination. One further complication was introduced as well; the capability to select a number of stations $n < NT$ in the target area and $m < NC$ in the control area which would have nonzero weighting factors. This feature could be represented by introducing an additional variable, δ, into the constraints giving:

$$\sum_{i=1}^{NT} w_i \delta_i \mu_{Y_i} = \sum_{i=1}^{NT} \delta_i \mu_{Y_i} \tag{13}$$

$$\sum_{i=NT+1}^{NT+NC} w_i \delta_i \mu_{X_i} = \sum_{i=NT+1}^{NC} \delta_i \mu_{X_i} \tag{14}$$

where δ_i is a variable with the property

$$\delta_i = \{^{1 \text{ if } w_i \neq 0}_{0 \text{ if } w_i = 0}\} \quad i = 1, 2, \ldots NT + NC$$

These constraints now imply that only those locations with nonzero weighting factors should be included in the constraint to maintain their physical significance.

Thus, a method of determining the weighting factors is obtained in the form of a mathematical programming problem. This problem has a nonlinear objective function and two equality constraints which are unlinked. Noting the form of Equation (11), it can be seen that the term $\sum_{i=1}^{NT} w_i \mu_{Y_i}$ in Equation (10) has a constant value. Thus, this problem is reduced to minimizing the square of a variance. Since a variance is always positive, the objective function, Equation (10) can be seen to be positive definite and, therefore, a minimum is a global

minimum. A final note on the problem formulation; the percent change in the mean, k, has been assumed to be the same at all locations, that is if k is 0.10, a 10 percent change is assumed at each station.

Mathematical Programming Aspects

The problem formulation as presented above has several features which prevent the use of the more "standard" mathematical programming approaches including:

1) The highly nonlinear nature of the objective function, and

2) The provision for modification of the constraint equations as the optimization proceeds.

For large scale problems, the direct methods of optimization where an initial point is specified and solution proceeds to the minimum (or maximum) by successive stepwise improvement (Beightler, Phillips and Wilde, 1979) have proven to be the most efficient. Thus, a method for solution of this problem has been developed using as its basis these direct procedures with modifications where necessary to account for the specific nature of the problem.

The algorithm for solution of this problem is based on the iterative use of a Quadratic Programming (QP) code, Jønch-Clausen and Morel-Seytoux (1978). The nonlinear objective functon is expanded about a feasible point in a second order Taylor Series approximation. Using this approximation, the problem is now a QP problem, i.e., a quadratic objective function with linear constraints (Morel-Seytoux, 1976) and the optimization proceeds as usual with this sort of problem. However, at each new point in the improvement of the objective function, the approximation is checked to see if it is still sufficiently close to the

actual function. If not, it is reapproximated at this new point before continuing with the optimization. Details of the quadratic expansion of the objective function are presented in Appendix B.

The aspect of the changing constraints during the optimization process due to the physical considerations of the problem required further modifications in the standard QP approach. This feature influences several phases of the optimization algorithm. First, since the number of nonzero variables (weighting coefficients in this case) is selected a priori, when this number is reached in the program, no move can be made to decrease the objective function which increases a zero valued variable but does not drive a nonzero valued variable to zero. Thus, movements are considerably more restricted in this case. In addition, if a move requires a variable to be changed to a zero value, the constant factor in the constraint equations change. Proceeding with this move then requires the constraint equations to be reformulated temporarily to allow for determining the values of the other variable under the new constraint set. If these new values produce a decrease in the objective function, the move is accepted and all changes are made permanent. If not, the variable with the next greatest affect in decreasing the objective function is selected and the process is repeated. The fact that the two constraint equations (11) and (12) do not have any variables in common makes this a relatively easy process. A detailed description of the optimization program, IITQP, has been prepared in the form of a Users Manual (Koch and Morel-Seytoux, 1980) including description of the input data requirements, examples of output and a listing of the program.

The mathematical programming problem to determine the weighting coefficients has thus been formulated and is relatively specific to the problem at hand.

Example Applications

It has been previously demonstrated (Morel-Seytoux and Saheli, 1973) that the technique presented using weighted linear combinations in a target-control test allows for detection of change in a much shorter time horizon than would a traditional statistical test. However, to assess the utility of this approach versus traditional methods in the analysis of changes in water quality, an example application is presented. The selection of an area for this application was based on several criteria. First, an adequate data base was necessary. This requires a fairly long period of annual stream flow and quality data at a number of stations within reasonable proximity to each other. This was the primary requirement. Second, an area which would likely be subject to future development pressures that could result in changes in water quality, such as a major change in land use, was sought. Finally although not a strong criteria, an area with some proximity and therefore interest to the state of Colorado was given consideration.

After evaluation of several candidate study areas, the upper Colorado River Basin of Colorado, Utah and Wyoming was chosen. This area meets all of the requirements, particularly in terms of data availability. Further, large deposits of coal and oil shale are found in some tributary basins leading to the possibility of large scale mining activities in these areas in the near future. In addition, as the headwaters of the Colorado River, this area provides much of the water available to the arid southwestern United States.

Description of the Study Area. The upper Colorado River basin encompasses most of western Colorado, southwestern Wyoming and eastern Utah. Major rivers in this area include the San Juan, Dolores, Gunnison, Colorado, White, Yampa, Green and Duchesne. These streams all have their headwaters in high mountainous areas and the flow regimes are dominated by snowmelt runoff in the spring and early summer months. The lower elevation areas are quite arid and tend to produce little runoff to the streams except in cases of intense thundershower activity which are generally very localized.

Selection of Stations. Six stations were selected for the example application based primarily on the availability of data. These stations along with their drainage areas, periods of record and mean annual discharge are presented in Table 1. Each of these stations has a daily record of stream discharge and conductivity (EC) collected by the U.S. Geological Survey (USGS) for at least the period 1964 through 1979. Thus, 16 years of annual data on these two variables are available at all six of these stations. Finding a longer concurrent data base of both flow and quality for this many stations in this area was not possible.

Due to their proximity to the energy resource areas, the Duchesne, Green and White River stations were chosen as the target area stations in this example. The remaining three stations, the Colorado, Gunnison and Dolores Rivers, are the control area stations.

Station Characteristics. The statistical characteristics of the relatively short samples of annual flow and quality are of interest for two reasons. First, they are part of the input required for the optimization code which determines the appropriate weighting factors in the linear combinations. Second, there were certain assumptions made in

Table 1

USGS Gaging Stations Selected for Use in the Example Application

Station Name	USGS Station Number	Period of Record		Drainage Area (sq.mi.)	Mean Annual Discharge (cfs)
		Chemical Analyses	Streamflow		
Gunnison River near Grand Junction, CO	09152500	10/31-9/79, 4/49-9/79	Concurrent	7928	2590
Dolores River near Cisco, UT	09180000	3/51-9/59, 10/64-9/79	Concurrent	4580	688
Green River near Jensen, UT	09261000	6/47-9/52, 4/62-9/79	Concurrent	25400	4940
Duchensne River near Randlett, UT	09302000	12/50-9/51, 11/56-9/79	Concurrent	3920	770
White River near Watson, UT	09306500	12/50-9/79	Concurrent	4020	655
Colorado River near Cameo, CO	09095500	10/33-9/79	Concurrent	8050	3560

deriving the test including independence of observations, constant variance and an underlying normal distribution. Certain statistics of the sample data set can be used to evaluate how well the data meet these assumptions.

Due to the small sample size, constant variance is assumed for all stations since not enough data are available to test this assumption. The assumption of normality was evaluated by two simple procedures: testing of the coefficient of skewness and plotting the data on normal and lognormal probability paper. For normally distributed data the coefficient of skewness should not be statistically different from zero while for lognormally distributed data the skewness coefficient of the logs should not be statistically different from zero. Further, depending on whether the data are normally or lognormally distributed, they should plot as a straight line on normal or lognormal probability paper respectively. This is, effectively, another less rigorous test of the skewness. Independence of the series is tested through the first autocorrelation coefficient. For independent data, this value should not be statistically different from zero.

In Table 2 relevant statistics, mean ($\bar{Q}$, $\overline{EC}$), standard deviation (S_Q, S_{EC}), coefficient of variation (C_v), skewness coefficient (g_Q, g_{EC}) and lag one serial correlation coefficient (r_1), for the untransformed data are presented while similar statistics for the log transformed data are presented in Table 3. As previously stated, the coefficient of skewness for normally distributed data should not be statistically different from zero. A test presented by Salas et al. (1980), for skewness of small samples taken from normal distributions show that a value in the range of 1.3 to 1.4 would not be unexpected for a sample size of 16 observations. All of the annual data for both streamflow and

Table 2

Basic Statistics of Streamflow and Conductivity Data[1/]
for the Stations Used in the Example Application

Station	Streamflow (cfs)					Conductivity (μmhos/cm)				
	$\bar{Q}$	S_Q	C_v	g_Q	r_1	$\overline{EC}$	S_{EC}	C_v	g_{EC}	r_1
Gunnison River	2140	737	0.34	0.35	0.10	1116	216	0.19	0.01	0.52
Dolores River	970	1028	1.06	2.74	-0.12	2939	925	0.31	1.17	0.21
Green River	4191	988	0.24	-1.31	0.48	650	47	0.07	-0.46	0.36
Duchesne River	541	266	0.49	0.50	0.19	1471	455	0.31	-2.05	0.11
White River	621	135	0.22	-0.69	0.03	798	91	0.11	-0.20	0.68
Colorado River	3487	793	0.23	-0.18	0.26	936	97	0.10	0.69	0.48

[1/] Based on the 16 year period 1964 to 1979.

Table 3

Basic Statistics of the Logarithmic Transformations of Streamflow and Conductivity Data[1/] for the Stations Used in the Example Application

Station	Streamflow					Conductivity				
	$\bar{Q}$	S_Q	C_v	g_Q	r_1	$\bar{EC}$	S_{EC}	C_v	g_{EC}	r_1
Gunnison River	7.61	0.37	0.05	-0.16	0.16	7.00	0.20	0.03	-0.38	0.52
Dolores River	6.54	0.81	0.12	0.43	-0.22	7.94	0.30	0.04	0.34	0.14
Green River	8.30	0.33	0.04	-2.54	0.56	6.48	0.07	0.01	-0.70	0.37
Duchesne River	6.15	0.60	0.10	-1.14	0.32	7.35	0.15	0.02	0.59	0.12
White River	6.42	0.25	0.04	-1.52	0.02	6.68	0.12	0.02	-0.46	0.69
Colorado River	8.13	0.25	0.03	-0.71	0.24	6.84	0.10	0.01	0.64	0.50

[1/] Based on the 16 year period 1964 through 1979.

conductivity fall within this range with the exception of annual streamflow in the Dolores River which is strongly skewed to the right. No significant overall gain is made by transforming the data logarithmically, thus, the untransformed data will be used in this analysis. Appendix C presents plots of the data both on probability and log probability paper for all of the stations. Visual inspection of these plots would also support the conclusion of normality.

The assumption of independence is evaluated by testing whether the lag one autocorrelation coefficient is statistically different from zero. Using a transform due to Fisher (Yevjevich, 1972 and Haan, 1977), the 95 percent two sided confidence interval for a sample size of 16 is (-0.49, 0.49). Inspecting the lag one autocorrelation coefficients presented in Table 2, most fall well within the 95 percent confidence limits. Log transformation does not affect these values appreciably in either direction. Thus, for such a small sample size, the hypothesis of independence in the data is accepted in general. This assumption will be further discussed and tested when reliability of the test is considered.

In addition to the statistical characteristics at each station, the interrelationships between stations are of interest and are necessary as input to the optimization problem. The covariance matrix for annual values of EC are given in Table 4. All of the values in this table are positive with the exception of the covariance between EC at the White River and Dolores River Stations. This value is quite small, however, and indicates very little correlation between these two stations. In general, then, there is a positive correlation between all combinations of stations as would be expected. In Table 5, the covariance matrix between EC and streamflow (Q) in the three target

Table 4

Covariance Matrix for Annual Conductivity in the Study Area

	Target Locations			Control Locations		
	Duchense River	Green River	White River	Colorado River	Gunnison River	Dolores River
Duchesne River	206776	5407	6765	8794	23545	102316
Green River	5407	2178	2332	1489	6473	25740
White River	6765	2332	8245	6073	16547	-5406
Colorado River	8794	1489	6073	9344	16689	5827
Gunnison River	23545	6473	16547	16689	46646	41908
Dolores River	102316	25740	-5406	5827	41908	856035

Table 5

Covariance Matrix for Annual Conductivity and Annual Streamflow
for the Target Stations in the Study Area

		Conductivity			Streamflow		
		Duchesne River	Green River	White River	Duchesne River	Green River	White River
Conductivity	Duchesne River	206776	5407	6765	-25908	-129064	-29861
	Green River	5907	2178	2332	-1603	-10342	-3473
	White River	6765	2332	8245	5107	-47408	-6247
Streamflow	Duchesne River	-25908	-1603	5107	75206	130959	17073
	Green River	-129064	-10342	-47408	130959	1041706	100650
	White River	-29861	-3473	-6247	17073	100650	19393

area stations is listed. Here, a negative correlation results between EC and Q in all cases except one where the computed correlation is very low while a positive correlation is exhibited for only EC or only Q. This is as expected since, as discharge is increased, concentration and therefore conductivity tends to decrease. These data will be the basis for two applications of the methodology presented.

To provide a better feeling for the interrelationships between variables, Tables 6 and 7 present the correlation matrices between the stations.

Table 6

Correlation Matrix for Annual Conductivity in the Study Area

	Target Area			Control Area		
	Duchesne River	Green River	White River	Colorado River	Gunnison River	Dolores River
Duchesne River	1.0	0.25	0.16	0.20	0.76	0.24
Green River	0.25	1.0	0.55	0.33	0.64	0.60
White River	0.16	0.55	1.0	0.69	0.84	-0.06
Colorado River	0.20	0.33	0.69	1.0	0.80	0.07
Gunnison River	0.76	0.64	0.84	0.80	1.0	0.21
Dolores River	0.24	0.60	-0.06	0.07	0.21	1.0

Table 7

Correlation Matrix for Annual Conductivity and Annual Streamflow for Stations in the Target Area

	Conductivity			Streamflow		
	Duchesne River	Green River	White River	Duchesne River	Green River	White River
Duchesne River	1.0	0.25	0.16	-0.21	-0.28	-0.47
Green River	0.25	1.0	0.55	-0.13	-0.22	-0.53
White River	0.16	0.55	1.0	0.21	-0.51	-0.49
Duchesne River	-0.21	-0.13	0.21	1.0	0.47	0.45
Green River	-0.28	-0.22	-0.51	0.47	1.0	0.71
White River	-0.47	-0.53	-0.49	0.45	0.71	1.0

Application No. 1. The first application is a direct analogy to the original use of this technique by Morel-Seytoux and Saheli (1973) where it was applied to streamflow data to detect changes resulting from weather modification. In this case, however, conductivity rather than streamflow is the variable. Using the six stations described earlier, partitioned into target and control areas as presented in Table 4, the optimization routine was applied to determine the values of the weighting factor which would detect the change of 10 percent (k=0.10) in the minimum time, with a significance level, α, of 0.05, and a power, $1-\beta$, of 0.50. With equal weights, the objective function indicated

7 years was required to detect the change. The results of this analysis are given in Table 8. With these weighting factors (constrained to be nonnegative) the minimum amount of time required to detect this 10 percent change is 1 year (rounded to the next highest integer value). Of note in these results are the fact that two of the target stations have little or no effect on the test as their respective weighting factors approach zero.

Application No. 2. Another approach to this problem of detection of change is to use alternative variables rather than different areas. That is, choosing variables that are related such as EC and Q and assuming that one variable, the target, will change while the other, the control, will not. In this case, EC was chosen the target variable and Q the control variable and only the data for the three stations that were previously defined as the target area were used. Applying the optimization routine to detect a 10 percent change for an α of 0.05 and $1-\beta$ of 0.50, the weighting factors were determined. For equal weights, the time required to detect the change was 7 years. The results are presented in Table 9. In this case, the minimum time was also 1 year. Again several of the weighting factors were very small or zero values.

Since both of these applications included stations with very low coefficients of variation, especially the Green River, a third example application is presented.

Application No. 3. The additional example application is given to illustrate how the method can markedly decrease the time required to detect a change when there is relatively high variability in the variables of interest. In this example four stations were selected; two as target stations and two as controls. These stations are all located

Table 8

Results of Example Application No. 1 EC in the Target Area vs. EC in the Control Area

Station	Weighting Factor
Target Area	
Duchesne River	0.02
Green River	3.21
White River	0.00
Control Area	
Colorado River	1.00
Gunnison River	1.00
Dolores River	1.00

Table 9

Results of Example Application No. 2 EC as the Target Variable vs. Q as the Control Variable

Station	Weighting Factor
Target Variable, EC	
Duchesne River	0.097
Green River	3.56
White River	0.58
Control Variable, Q	
Colorado River	0.00
Gunnison River	0.00
Dolores River	1.00

in the White River Basin in western Colorado. The two target stations, Yellow Creek and Piceance Creek are underlain by large deposits of oil shale and thus are likely to be subject to intense development pressures in the future. The two control stations, the north and south forks of the White River are relatively undeveloped and are likely to provide a stable control area. Table 10 presents the basic characteristics for streamflow at these four stations while the interstation relationships are given in Table 11. Only short concurrent record is available, however the thrust of this example is to demonstrate the utility of the detection method. The reliability of the results are addressed in the following section.

A preliminary inspection of Tables 10 and 11 reveals that the coefficients of variation of these variables (annual streamflow) are considerably higher than those used in either Application No. 1 or No. 2. In addition, there is fairly low correlation between stations in the target area while the control stations are highly correlated. Further, one control station, the North Fork of the White River is consistently more highly correlated with the target stations than is the other control station.

Applying the optimization routine to this example produced a more marked savings in time for detection of a 10 percent change in the mean annual flow for the same size and power as the previous two applications. When equally weighted, 22 years were required for detection of the specified change. Upon selection of the optimal weighting factors, only 4 years were required. The results of the optimization procedure are given in Table 12. From these results it is noted that the second control station, the south fork of the White River was given a zero weight and thus was not included in the detection scheme.

Table 10

Statistical Characteristics of Annual Streamflow
at the Four Stations Used in Application No. 3

Station Name	USGA No.	Mean (cfs)	Standard Deviation (cfs)	Coefficient of Variation
Yellow Creek	09306255	1.79	0.50	0.28
Piceance Creek	09306200	22.08	7.28	0.33
North Fork White River	09303000	282.0	80.59	0.29
South Fork White River	09304000	242.6	74.72	0.31

Table 11

Correlation Coefficient Matrix for Annual Streamflow at the
Four Stations Used in Application No. 3

	Yellow Creek	Piceance Creek	N. Fork White River	S. Fork White River
Yellow Creek	1.00	0.17	0.79	0.61
Piceance Creek		1.00	0.53	0.38
N. Fork White River			1.00	0.93
S. Fork White River				1.00

Table 12

Results of Example Application No. 3 Annual Discharge in the Target Area vs. Annual Discharge in the Control Area

Station	Weighting Factor
Target Stations	
Yellow Creek	8.75
Piceance Creek	0.37
Control Stations	
N. Fork White River	1.00
S. Fork White River	0.00

Evaluation of the Method

To evaluate this method against the more traditional approaches, Table 13 has been prepared. Presented here are the number of observations, N, that would be required to detect a 10 percent change in EC if only a single station in the target area were used. This is computed using Equation (3) with data taken from Table 2. From Table 13, two of the stations selected have very low coefficients of variation and therefore require few observations to detect a change. Further inspection of the results presented in Tables 7 and 8 reveal that those stations have indeed been assigned the largest weights, particularly in the second example. In the first example, little change was noted in the objective function with further changes in the variables. As a result the process was halted. Similar results are noted for the third example from Table 13. Also, in each case, the weighted linear combinations do provide a shorter time required for detection of the changes over any single station using the traditional statistical technique.

Table 13

Results of Detection of Change Problem Using a Traditional Approach[1/] on Target Area Station from Applications No. 1, No. 2 and No. 3

Station	Coefficient of Variation C_v	Number of Years Required for Detection of Change N
Duchesne River	0.31	26
Green River	0.07	2
White River	0.11	5
Yellow River	0.28	21.22
Piceance Creek	0.33	29.47

1/

$$N = \left(\frac{z_{1-\alpha} \quad z_{1-\beta}}{k}\right)^2 \quad C_v^2 \qquad (6)$$

$k = 0.01$, $z_{1-\beta} = 0.00$, $z_{1-\alpha} = 1.645$

Several of the implications of this approach can be noted. First, the relative magnitude of the weighting factors, particularly as they approach zero, indicate the value of any particular station in providing information for detecting changes. In both examples, certain stations were given no weight indicating they added nothing additional to the information contributed by the other stations. The criteria for station weighting seems to be as follows. In the target area, the stations with the lowest coefficient of variation are weighted highest unless there is a high coefficient of correlation between the stations. Then, the weight is not so high as seen in Example 2 where both the Green and White Rivers have low coefficients of variation but, with a relatively high correlation between them, only the smallest C_v is weighted

heavily. This result agrees with that of Matalas and Langbein (1963) who showed that little information is added to a single station by considering another station to which it is highly correlated. In the control area, the weighting seems to depend on both the correlation between the control and target stations and on the correlation within the control area. The station with relatively high correlations with the target area is weighted heavily, particularly if it is also highly correlated with the other control stations. This is also borne out in Example 2 and seems to follow from Matalas and Langbein (1963) as well.

In summary, using the time required to detect a change of a selected magnitude with a given size and power as the criteria, the method presented can detect the change within a shorter time period than traditional tests. It can be expected to perform much better than the traditional methods with more variable data while, if the relative variation is quite small, the results may not be significantly better. Thus the method has its greatest application when applied to variables which are highly variable. In addition, the use of highly correlated stations between the target and control areas tend to make the results better while high correlation within the target or control areas result in the exclusion of some of the stations from the analysis by assigning them zero weighting factors.

Implications of the Method

In addition to the obvious use of this technique for developing new test variables as weighted linear combinations of other variables which have the ability to exhibit changes more quickly, other uses may arise. For example, under certain financial constraints, an agency may need to cutback on monitoring activities in an area. The weighting factors

resulting from this analysis provide an objective means of assessing the relative importance of each station in the overall network at least with regard to its ability to serve a change detection function.

ASSESSMENT OF RELIABILITY

Thus far, a test for detecting a change in a hydrologic variable has been presented and applied to a sample case. It is demonstrated that this method allows for the detection of a change in a much shorter time horizon than do the more standard statistical tests. It has also been noted, however, that the theory underlying this test is developed with two restrictions: the parameters of the distribution are known and the random variables have certain characteristics including normal distributions, time independence and constant variance. However, to actually apply the test, sample estimates of population parameters are substituted in all of the equations. The reason for this will be explained later. Further there is some reason to question the validity of the assumption of time independence even in annual data. The validity of the test will then depend upon the effects of these simplifications.

Specific to the problem at hand, the interest in the affects of the approximations relates to the reliability of the test in its application to a real world situation. In solving the optimization problem, weighting factors are determined which, in fact, are only optimal for the sample data used to develop the coefficients in the objective function and constraints. Since, as sample values, these estimates are random variables the "optimal" number of years is also a random variable as are the weighting coefficients and are therefore characterized by some distribution. It is then necessary to determine at least the first few moments of these variables to assess, approximately, the reliability of the resulting estimates of N and w. In addition, since the weighting factors are no longer constants but random variables, the

distribution of the weighted linear combinations is not known. Therefore the test may not have the purported size and power originally intended. Any time dependence in the data will only serve to increase these affects.

The problem of the distribution of an optimized random variable can also be viewed as being of interest in general terms. The problem being posed is an optimization problem with random variables appearing as coefficients in the objective function and constraint equations. In many cases, e.g., Morel-Seytoux (1976), especially engineering applications, mathematical programming problems have coefficients or "constant" values which represent outcomes of various natural processes such as streamflow or precipitation. These variables are best described by distributions rather than specific numbers due to their apparent stochastic nature. This problem can be dealt with by using either the expected value of the random variable or some quantitative level which was considered "safe" for the purposes at hand. The technique, however, ignores the fact that the optimal value of the objective function is a function of these random variables and is itself a random variable. Ideally, then, this optimal value should be expressed as a function of these random coefficients from which, at least theoretically, its distribution could be derived analytically given that the distribution of the coefficients are known. It is this technique that is pursued in the following discussion to assess the reliability of the detection method, or in more general terms, the reliability of optimized random variable.

Method of Analysis

As stated above, the ideal approach is to express the objective function in terms of the random variables of the problem and from this expression derive its distribution. In certain situations at least the first step, expressing the objective function in terms of the random variables of the problem, can be accomplished. The method, however, depends on exactly how the problem is posed. In the case where the variables of the optimization problem are free to take on both positive or negative values and the constraints are equalities, the traditional Lagrange multiplier approach can be applied to develop unique expressions for the variables and objective function in terms of the random coefficients. An additional complication occurs when the optimization variables are constrained to be non-negative, a case which often occurs in optimization problems resulting from physical considerations in engineering, planning or management applications. In these situations, the possibility exists for piecewise solutions depending on the values assumed by the random variables. This is due to the form of the stationarity conditions in this case. Appendix D gives a complete mathematical description of these two cases. Thus, for the second situation, where non-negativity is imposed, all of the combinations which could satisfy the stationarity conditions must be investigated to establish the range of values of the random variable for which each combination does produce the minimum. Depending, then, on the number of optimization variables in the problem, there may be a number of combinations to be investigated. A more detailed description of this problem is presented in Appendix D.

The merit of this approach is obvious when the alternative method is evaluated. Rather than solve for explicit expressions, a complete simulation can be undertaken where those coefficients and "constants" in the problem which are by nature random, are varied over ranges of their values and, at each set of values the optimization problem is solved. This method, however, has several shortcomings. First, a great many simulations may be required to develop a reliable picture of how the objective function and optimization variables react to the variation in the random variables. As a result, the true interactions in the problem may not be discovered. It has been shown, Morel-Seytoux (1975), that the objective function may be described by a piecewise rather than a continuous relationship, the actual form of which may not be evident from the random selection of values of the random variables. In addition, this approach may be very costly, requiring the optimization problem to be solved many times to achieve a reasonably accurate indication of the interactions. If analytical expressions can be derived, they would provide explicit relations defining how the random variable affects the problem.

A more complex problem is the derivation of the distribution of the objective function and variables once the analytical expressions have been developed. If the expressions from the analytical solution are very complex, it may not be possible to derive these distributions analytically. At this point, simulation could be applied to establish the approximate distributions of the objective function and variables knowing, from the explicit relationships previously derived, the interactions of the random variables in determining their values. Therefore, at best, a completely analytical solution can be obtained and, at worst, a hybrid analytical-simulation approach can be applied.

Reliability of the Detection Problem

To demonstrate the approach suggested above, an assessment of the reliability of the detection problem is presented. However, due to the complexity (see Appendix C) of the original objective function, it was not deemed suitable for a test of this methodology. Instead, a simpler problem is posed without the conditional distribution being considered. In this case, it is assumed that there are several locations where change is likely to occur and a test is developed by using a weighted linear combination, $Y^* = \sum_{i=1}^{N} w_i Y_i$, of these locations as the test variable. The optimization problem for selection of the weighting coefficients becomes:

$$\underset{\underline{w}}{\text{Min}} \left\{ N = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{k \, \mu_Y^*} \right)^2 \quad \sigma_Y^{2*} \right\} \tag{14}$$

subject to:

$$\sum_{i=1}^{M} w_i \mu_{Y_i} = \sum_{i=1}^{M} \mu_{Y_i} = \mu_Y^*$$

where M is the number of stations and all other variables are as previously defined.

Noting that, due to the constraint equation, only the variance, σ_Y^{2*}, can vary, the problem reduces to one of minimizing the variance of a weighted linear combination of variables. This is, then, a standard QP problem with one equality constraint. If sample estimates of the parameters are substituted and the expression for the variance is expanded, the problem is stated as:

$$\underset{\underline{w}}{\text{Min}} \left\{ N = \left(\frac{z_{1-\alpha} + z_{1-\beta}}{k \quad \bar{Y}^*} \right)^2 \quad \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j \, S_{Y_i Y_j} \right\} \tag{15}$$

subject to:

$$\sum_{i=1}^{M} w_i \bar{Y}_i = \sum_{i=1}^{M} \bar{Y}_i = \bar{Y}^*$$

where $S_{Y_i Y_j}$ is the sample estimate of covariance between Y_i and Y_j

In this problem, the covariances and means, being estimates based on sample data, are random variables. Both of the cases, when $\underline{w}$ is a free variable and when $\underline{w}$ is non-negative, are considered below. To further simplify matters, but without loss of generality, a case with only two stations is considered.

Free Variable Case. When non-negativity conditions are not placed on the variables in the problem (the weighting factors), any value can be assumed. In this case, only one solution results from solving the optimization problem since only one set of stationarity conditions applies (see Appendix D). From this procedure, the following expressions are obtained

$$w_1 = \frac{\bar{Y}^* \left(\frac{S_{Y_2}^2}{\bar{Y}_2} - \frac{S_{Y_1 Y_2}}{\bar{Y}_1} \right)}{\left(S_{Y_1}^2 \frac{\bar{Y}_2}{\bar{Y}_1} - 2 S_{Y_1 Y_2} + S_{Y_2}^2 \frac{\bar{Y}_1}{\bar{Y}_2} \right)} \tag{16}$$

$$w_2 = \frac{\bar{Y}^* \left(\frac{S_{Y_1}^2}{\bar{Y}_1} - \frac{S_{Y_1 Y_2}}{\bar{Y}_2} \right)}{\left(S_{Y_1}^2 \frac{\bar{Y}_1}{\bar{Y}_2} - 2 S_{Y_1 Y_2} + S_{Y_2}^2 \frac{\bar{Y}_1}{\bar{Y}_2} \right)} \tag{17}$$

$$N^{*} = \left(\frac{z_{1-\beta}+z_{1-\alpha}}{k}\right)^2 \left\{ S_{Y_1}^2 \left(\frac{S_{Y_2}^2}{\bar{Y}_2} - \frac{S_{Y_1Y_2}}{\bar{Y}_1}\right)^2 + 2\, S_{Y_1Y_2} \left(\frac{S_{Y_2}^2}{\bar{Y}_2} - \frac{S_{Y_1Y_2}}{\bar{Y}_1}\right)\left(\frac{S_{Y_1}^2}{\bar{Y}_1} - \frac{S_{Y_1Y_2}}{\bar{Y}_2}\right) \right.$$

$$\left. + S_{Y_2}^2 \left(\frac{S_{Y_1}^2}{\bar{Y}_1} - \frac{S_{Y_1Y_2}}{\bar{Y}_2}\right)^2 \right\} \Big/ \left(S_{Y_1}^2 \frac{\bar{Y}_1}{\bar{Y}_2} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \frac{\bar{Y}_2}{\bar{Y}_1}\right)^2 \qquad (18)$$

where $S_{Y_1}^2$ is the sample variance for Y_1,

$S_{Y_2}^2$ is the sample variance for Y_2, and

$S_{Y_1Y_2}$ is the sample covariance between Y_1 and Y_2

Thus, the variables and objective function are expressed explicitly in terms of the random variables of the problem. In this case, however, all of the coefficients are random variables; a more complex situation than might be expected in a typical mathematical programming formulation of an engineering problem.

Recalling that it has been assumed that the underlying variables are normally distributed, an attempt might be made to derive the distributions of the expressions in Equations (16), (17) and (18). However, due to their extreme complexity, particularly for (18), simulation should be entertained as the means of estimating these distributions.

Non-Negative Variable Case. Imposing non-negativity conditions on the weighting factors requires that they take on a value no less than zero. This changes the results of the optimization to the following (see Appendix D):

$$w_1 = \begin{cases} 0 & , \text{ if } \dfrac{\bar{Y}_1}{\bar{Y}_2} \leq \dfrac{S_{Y_1Y_2}}{S_{Y_2}^2} \\ \dfrac{\bar{Y}^*\left(\dfrac{S_{Y_2}^2}{\bar{Y}_2} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_1}\right)}{\left(S_{Y_1}^2 \dfrac{\bar{Y}_2}{\bar{Y}_1} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \dfrac{\bar{Y}_1}{\bar{Y}_2}\right)} & , \text{ if } \dfrac{\bar{Y}_1}{\bar{Y}_2} > \dfrac{S_{Y_1Y_2}}{S_{Y_2}^2} \end{cases} \tag{19}$$

$$w_2 = \begin{cases} 0 & , \text{ if } \dfrac{\bar{Y}_1}{\bar{Y}_2} \geq \dfrac{S_{Y_1}^2}{S_{Y_1Y_2}} \\ \dfrac{\bar{Y}^*\left(\dfrac{S_{Y_1}^2}{\bar{Y}_1} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_2}\right)}{\left(S_{Y_1}^2 \dfrac{\bar{Y}_2}{\bar{Y}_1} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \dfrac{\bar{Y}_1}{\bar{Y}_2}\right)} & , \text{ if } \dfrac{\bar{Y}_1}{\bar{Y}_2} > \dfrac{S_{Y_1}^2}{S_{Y_1Y_2}} \end{cases} \tag{20}$$

$$N^* = \begin{cases} \left(\dfrac{z_{1-\beta}\, z_{1-\alpha}}{k}\right)^2 \left(\dfrac{S_{Y_2}}{\bar{Y}_2}\right)^2 & , \text{ if } w_1 = 0 \\ \left(\dfrac{z_{1-\beta}\, z_{1-\alpha}}{k}\right)^2 \left(\dfrac{S_{Y_1}}{\bar{Y}_1}\right)^2 & , \text{ if } w_2 = 0 \\ \left(\dfrac{z_{1-\beta}+z_{1-\alpha}}{k}\right)^2 \left\{ S_{Y_1}^2 \left(\dfrac{S_{Y_2}^2}{\bar{Y}_2} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_1}\right)^2 + 2S_{Y_1Y_2}\left(\dfrac{S_{Y_2}^2}{Y_2} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_1}\right)\left(\dfrac{S_{Y_1}^2}{\bar{Y}_1} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_2}\right) \right. & \\ \left. + S_{Y_2}^2 \left(\dfrac{S_{Y_1}^2}{\bar{Y}_1} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_2}\right)^2 \right\} \Big/ \left(S_{Y_1}^2 \dfrac{\bar{Y}_1}{\bar{Y}_2} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \dfrac{\bar{Y}_2}{\bar{Y}_1}\right)^2 & , \\ & \text{if } w_1, w_2 > 0 \end{cases} \tag{21}$$

In this case, the same solution results, as it must, when both weighting factors are positive. The difference exists in that neither weight is permitted to assume a negative value. The regions where the non-negativity conditions apply are defined in terms of the random variables of the problem in Equations (19) and (20).

Further Evaluation of the Method. Once the optimization problem has been solved explicitly, the optimal values of the variables are expressed in terms of the parameters of the problem which, in many cases, are best described as random variables. The explicit analytical relationships such as Equations (16) through (21) in general allow a much deeper insight into the problem and the interrelationships of the variables. For the detection of change problem being investigated in this particular case, an assessment of how the various sample statistics affect the optimal solution can be undertaken.

The explicit relationships presented in Equations (16) through (21) can be used in two ways. First, they can be used to evaluate the inter-relationships of the variables and the relative importance of any particular statistic such as the cross correlation between stations. This analysis leads to a better understanding of how the detection method actually works. A second use of the explicit relationships is to evaluate the variation of the optimal values of the problem (objective function and weights) with the random variability of the statistics. This leads to an assessment of the reliability of the detection method. This second use is discussed in detail later while the use of the explicit relationships to gain insight into the method is the subject of the remainder of this section.

Using Equations (16) and (17) for the weighting coefficients, further algebraic manipulation results in the following expressions

$$w_1 = \frac{\left(\frac{C_{v_2}}{C_{v_1}} - r_{Y_1Y_2}\right) \frac{\bar{Y}^*}{\bar{Y}_1}}{\left(\frac{C_{v_2}}{C_{v_1}} + \frac{C_{v_1}}{C_{v_2}} - 2r_{Y_1Y_2}\right)} \tag{22}$$

$$w_2 = \frac{\left(\frac{C_{v_1}}{C_{v_2}} - r_{Y_1Y_2}\right) \frac{\bar{Y}^*}{\bar{Y}_2}}{\left(\frac{C_{v_2}}{C_{v_1}} + \frac{C_{v_1}}{C_{v_2}} - 2r_{Y_1Y_2}\right)} \tag{23}$$

where C_{v_i} is the sample coefficient of variation of variable Y_i and

$r_{Y_1Y_2}$ is the sample correlation coefficient between Y_1 and Y_2

The weighting factors have thus been expressed in terms of the coefficients of variation, the correlation coefficient between stations and the mean values.

Many inferences about the behavior of the weighting factors can be drawn from these expressions. First, the magnitude of the weighting factors depends on the actual magnitude of the mean value of the variables being used in the problem through the ratios of the sum of the means, $\bar{Y}^*$, to the mean, $\bar{Y}_i$. If the mean $\bar{Y}_i$ is small relative to the sum of the means then the weighting factor, w_i, corresponding to this variable will be large and conversely if the mean is large relative to the sum, the weighting factor will be smaller.

The affect of interstation correlation is also of interest in this problem. Using Equations (22) and (23), the influence of this characteristic on the weighting factors can also be evaluated. First,

it can be seen that if the cross correlation coefficient, $r_{Y_1Y_2}$, is negative, then no negative weighting factor can result. Further, if $r_{Y_1Y_2}$ is positive, the occurrence of negative weighting factors depends on the relationship between the ratio of the coefficients of variation and the correlation coefficient. This provides another means of expressing the conditions imposed by the non-negativity conditions. Also, it is apparent that only one of the weighting factors can be negative in the free variable case since $r_{Y_1Y_2}$ is bounded by one and one coefficient of variation will be greater than the other leading to a ratio greater than one. This is consistent with the constraint used in the problem. It is also apparent, that if a negative weight occurs, it will be associated with the variable that has the largest coefficient of variation.

Special cases of cross correlation can also be evaluated. In the case when the correlation coefficient approaches zero, i.e., there is very little correlation between stations, the weighting factors depend only on the coefficients of variation and mean values. In this case, the ratio of the weights is inversely proportional to the ratio of the square of the coefficients of variation and inversely proportional to the ratio of the means as given below

$$\frac{w_1}{w_2} = \left(\frac{C_{v_2}}{C_{v_1}}\right)^2 \left(\frac{\bar{Y}_2}{\bar{Y}_1}\right) \tag{24}$$

It can also be noted that no negative weights can result from the case when the stations are uncorrelated. When the stations are very highly correlated, that is the correlation coefficient approaches 1.0, the weights are again a function of the coefficients of variation and the means as:

$$\frac{w_1}{w_2} = -\left(\frac{C_{v_2}}{C_{v_1}}\right) \quad \frac{\bar{Y}_2}{\bar{Y}_1} \tag{25}$$

In this case it is apparent that one weight is always negative in the free variable case or one weight is always zero in the non-negative case. Thus, to summarize, when there is little correlation between the stations, it is very likely that both weighting factors will be positive indicating that both are useful in assessing the detection of change. However, when the stations are highly correlated one of the weights is always negative or zero depending on whether non-negativity conditions are imposed.

Simulation Study

Due to the complexity of the explicit expressions for the objective function and weighting factors in terms of the sample estimates of the parameters, an analytical derivation of their distributions would be, at best, very difficult, and may not be possible. As a result, a simulation (data generation) study was undertaken at this point to evaluate certain characteristics of the test including:

1) the distribution of N^*,
2) the distributions of w_1 and w_2,
3) the actual size and power of the test.

The first two evaluations, the distributions of the objective function and variables, are of general interest to any stochastic optimization problem whereas the last analysis is specific to the detection of change methodology under investigation in this study.

Extrapolating from the results of the example applications and, knowing the assumptions used to derive the test, several aspects of the problem can be identified as being of interest in terms of how they

affect the properties presented above. First, the length of record used to develop the sample statistics (means and covariances) which are in turn used to determine the optimal value of the objective function and the weighting factors has an effect. It is well known that, as the sample size increases, the parameter estimates are more reliable, that is they have a smaller variation, and hence the results of the optimization problem would also have a smaller variation and be more reliable. It might also be postulated that the actual size and power of the test would be affected by the sample size of the original data. Finally, it is of interest to investigate whether imposing non-negativity conditions on the variables affects the results in any way.

Simulation Approach. For analyzing the affects of sample size and the non-negativity conditions on the variables and characteristics of the test required data generation from a multivariate normal (MVN) distribution.

The simulation procedure had the following steps:

I Generation of data sets, computation of statistics and optimal variables.

1) Generate a set of data of length N for each site by a MVN (in this case only a bivariate normal was required) with a selected mean vector, $\underline{\mu}$ and covariance matrix, Σ.

2) From this data set, compute sample estimates of the parameters, $(\hat{\underline{\mu}}, \hat{\Sigma})$.

3) Using these statistics, compute the optimal value of the objective function, N^*, and weighting factors, w_1 and w_2.

4) Compute the critical value of the weighted linear combination as $\bar{Y}^*_{cr} = z_{1-\alpha} S_{\bar{Y}^*} + \bar{Y}^*$

In the case of an application of the detection of change methodology, no further steps would be required. To test the reliability of the method, however, additional steps are necessary.

II. Evaluation of the size of the test.

1) Since the test is based on detection of change in mean values within N^* years and the weights are now determined, generate L1 means from the MVN distribution. The distribution of these means is MVN $(\underline{\mu}, \frac{1}{N^*} \Sigma)$

2) Using the weights and critical value determined from the original set of data compute $\bar{Y}^*$ from the generated data and perform the test as:

a) If $\bar{Y}^* \leq \bar{Y}_{cr}$, accept the hypothesis that the means are the same

b) If $\bar{Y}^* > \bar{Y}_{cr}$, reject this hypothesis

3) Count the number of rejections and divide them by the number of tests, L1, to estimate the size, $\hat{\alpha}$.

III Evaluation of the power of the test.

1) Generate L2 means from a MVN $(\underline{\mu} + k\underline{\mu}, \frac{1}{N^*} \Sigma)$, that is from a distribution with the mean 100k percent larger but with the same covariance matrix.

2) Using the same weights, optimal number of years and critical value perform the same test as in the evaluation of the size.

3) Count the number of rejections and estimate the power, $\hat{1-\beta}$, by dividing this number by the number of tests, L2.

This procedure then gives an estimate of all of the characteristics based on one set of generated data of length N. To determine the behavior of these estimates, that is N^*, w_1, w_2, $\hat{\alpha}$ and $\hat{1-\beta}$, many sets must be generated and evaluated in the same fashion.

The behavior of these estimates can then be evaluated in a relative frequency context. This is accomplished by computing the mean, variance and skewness of each of the estimates as well as plotting histograms based on the many sets of data generated. The affect of record length

is evaluated by performing simulation runs for various sample sizes, N. In this analysis, samples of length 10, 25 and 50 observations were generated. The effect of imposing non-negativity constraints is evaluated by direct comparison of simulations where each approach was used to determine the weighting coefficients. For each case, a change of 10 percent was considered with a size, α, of 5 percent and a power, $1-\beta$, of 50 percent.

Since data generation can be very costly, it is necessary to estimate the number of samples required for the analysis and this number should be as small as possible without affecting the accuracy of the results. It is then desirable to estimate the number of data necessary to determine whether the given characteristic falls within some specified limit with some fairly large probability. In this case, a sample length of 500 is necessary to say that the size of the test is 0.05 ± 0.02 while a sample size of 100 is required to determine where the power is 0.50 ± 0.10. The details of these computations are presented in Appendix E. For the simulation study, then, for each set of generated data of length N, 500 means were generated to test the size while 100 were generated to test the power. In all, 500 sets were generated to assess the average characteristics of N^*, w_1, w_2, $\hat{\alpha}$ and $\hat{1-\beta}$.

To summarize the data generation study, Table 14 has been prepared. All of the simulation runs are presented along with a description of the purpose of each one.

Selection of Stations. To assess the behavior of the objective function, variables and characteristics of the test, two stations were selected from those used in the example application of the complete detection problem presented earlier. The two stations selected were the

Table 14

Summary of Data Generation Runs for Reliability Study

Run No.	Sample Size (years)	Comments
1	25	Free Variable Case
2	10	Non-negativity Case
3	25	Non-negativity Case
4	50	Non-negativity Case

White River and the Duchesne River. Annual EC at these two stations has a relatively low sample cross correlation coefficient (r = 0.16) and the coefficients of variation are quite different; 0.11 for the White River and 0.31 for the Duchesne River. The sample statistics from 16 years of data at these two locations served as the basis for the simulation, that is, they were assumed to be the population values and used as the parameters in the distributions used for data generation.

Simulation Results. Four simulation runs were made as detailed in Table 14. From these runs, a great deal of insight is gained regarding the behavior of the objective function and variables of the problem as well as the characteristics of the statistical test.

The effects of the original sample size on the value of the objective function and variables of the problem are summarized in Table 15. As the number of observations used to compute the sample statistics (means and covariances), and thus the weighting factors increased, little change was noted in the overall mean value of the objective function or any other variables. This is shown graphically in Figures 2, 3 and 4. There was, however, a marked affect on the variation of the optimal value of the objective function and weighting factors as shown in Figures 5, 6 and 7. The standard deviation of all three variables decreases markedly as the sample size used to compute the statistics increases. This points to considerably more reliable estimates of the variables with a larger sample size as would be expected. The decrease of standard deviation of N^* with sample size is approximately proportional to $1/\sqrt{N}$. The standard deviation of the weighting factors, w_1 and w_2 also decrease with increasing sample size but the decrease is much slower after a period of from 25 to 30 years.

Table 15

Comparison of Objective Function and Optimization Variables Based on Sample Size

Run No.	Sample Size (years)	Variable	Mean	Standard Deviation	Skewness
2	10	N^*	3.396	1.582	0.774
		w_1	2.717	0.274	-1.061
		w_2	0.070	0.124	2.585
3	25	N^*	3.478	0.932	0.561
		w_1	2.696	0.173	-0.439
		w_2	0.075	0.085	1.035
4	50	N^*	3.335	0.667	0.384
		w_1	2.680	0.144	0.008
		w_2	0.087	0.073	0.456

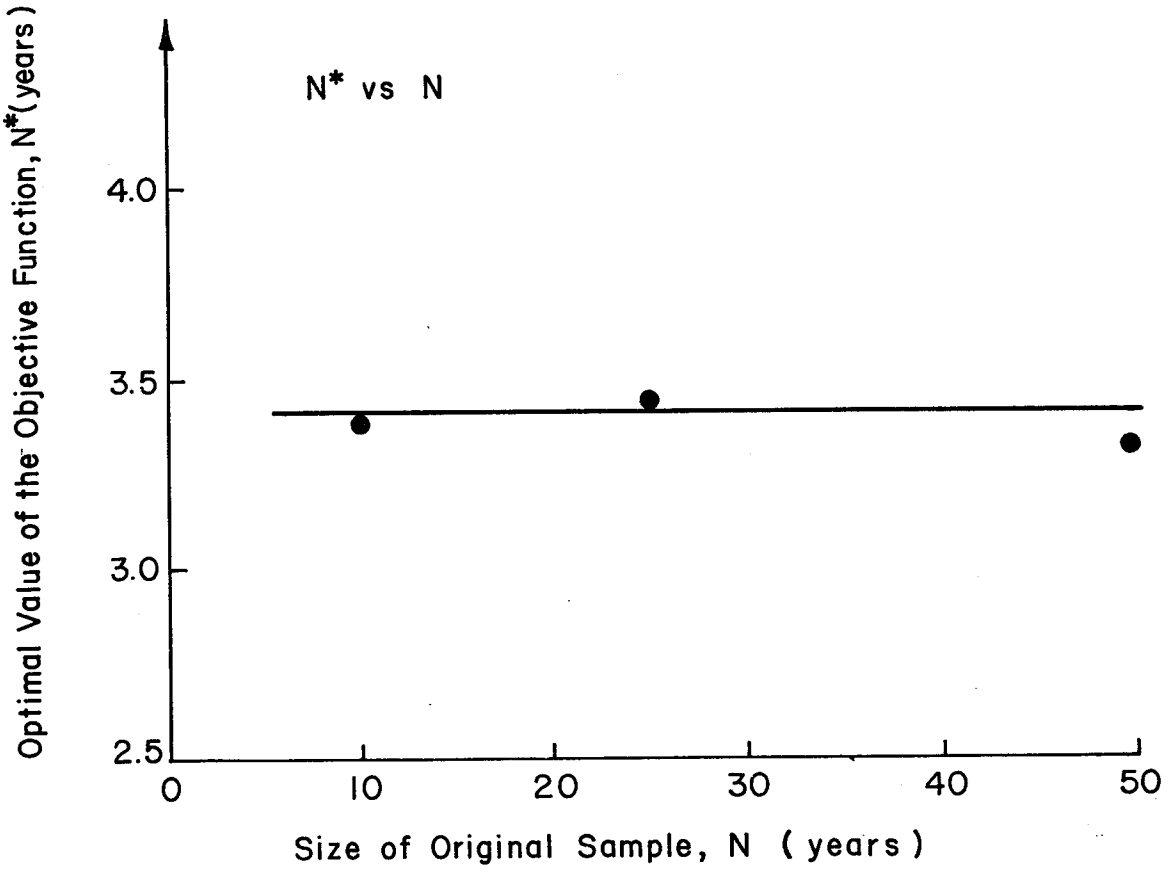


Figure 2. Mean of the Optimal Value of the Objective Function as Related to the Sample Length Used to Compute It.

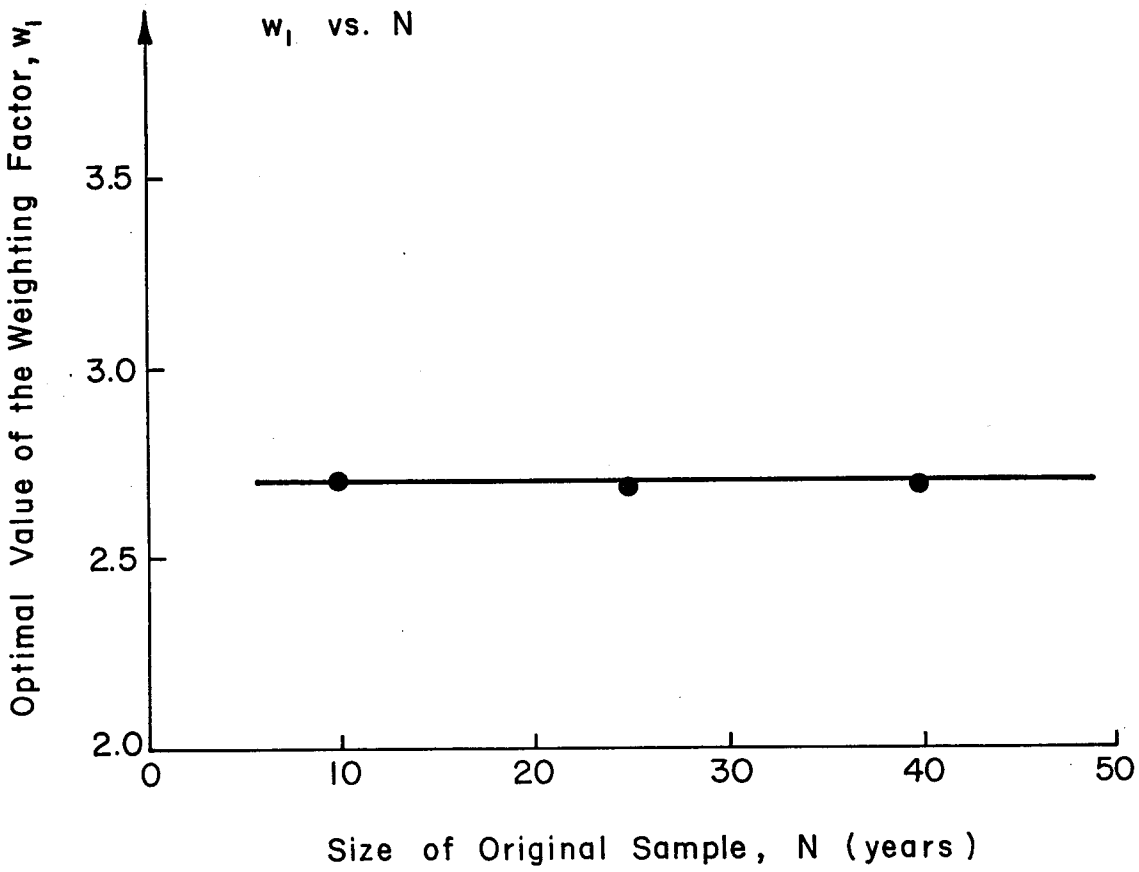


Figure 3. Mean of the Optimal Value of the Weighting Factor, w_1, as Related to the Sample Length Used to Compute It.

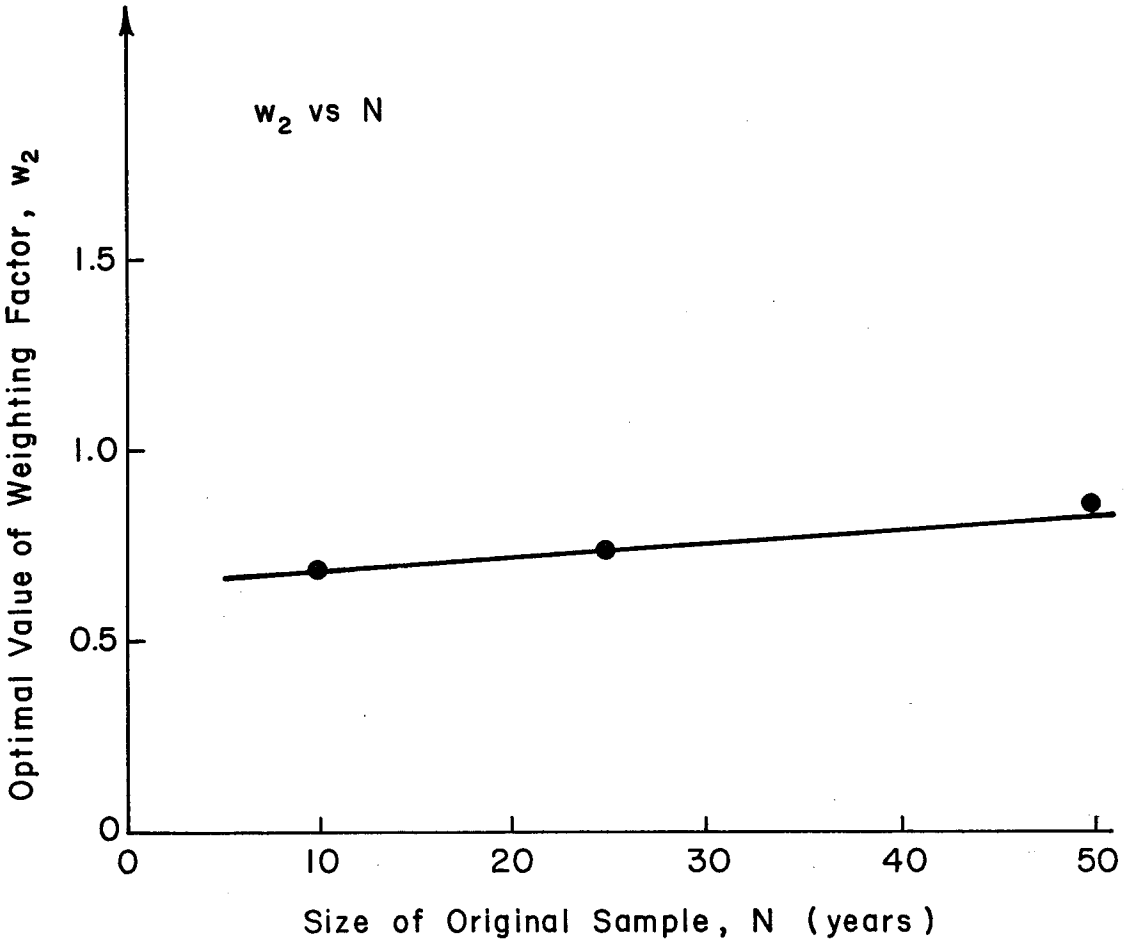


Figure 4. Mean of the Optimal Value of the Weighting Factor, w_2, as Related to the Sample Length Used to Compute It.

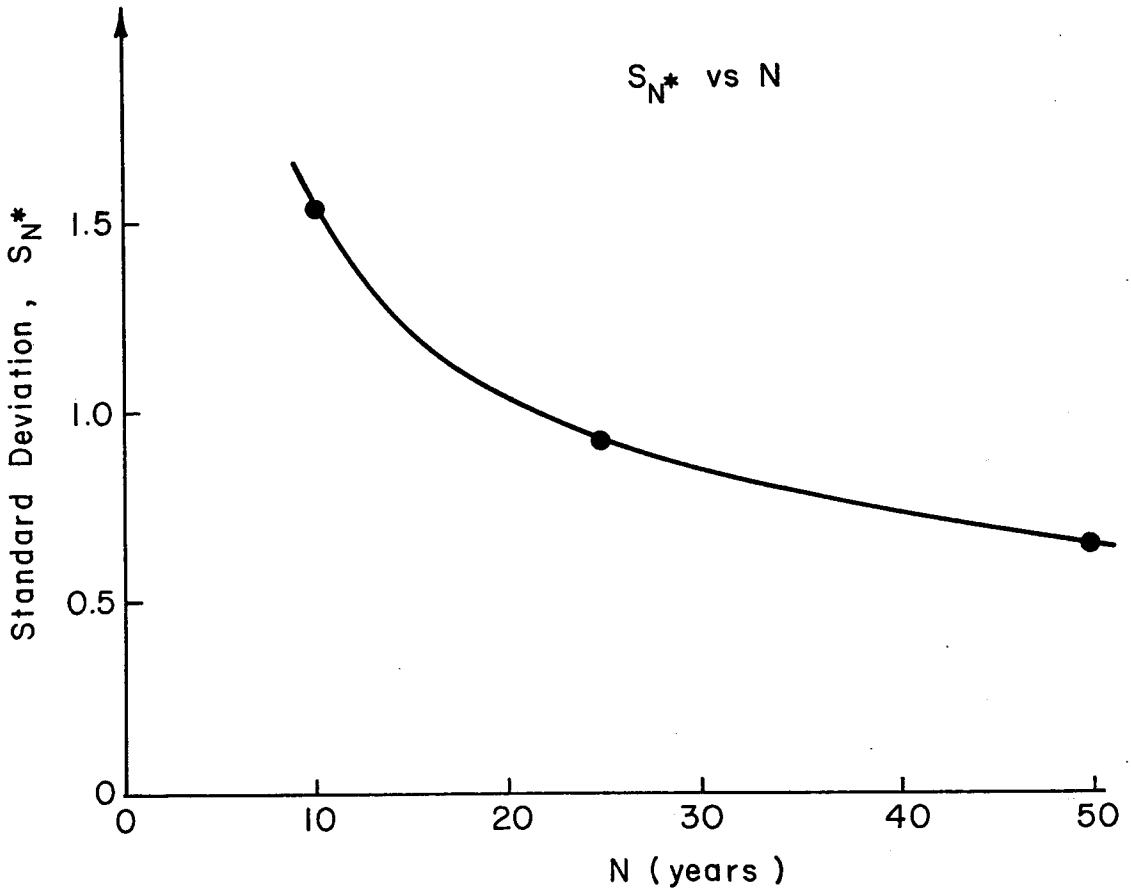


Figure 5. Standard Deviation of the Optimal Value of the Objective Function, N^*, as It Relates to the Sample Length Used to Compute the Statistics.

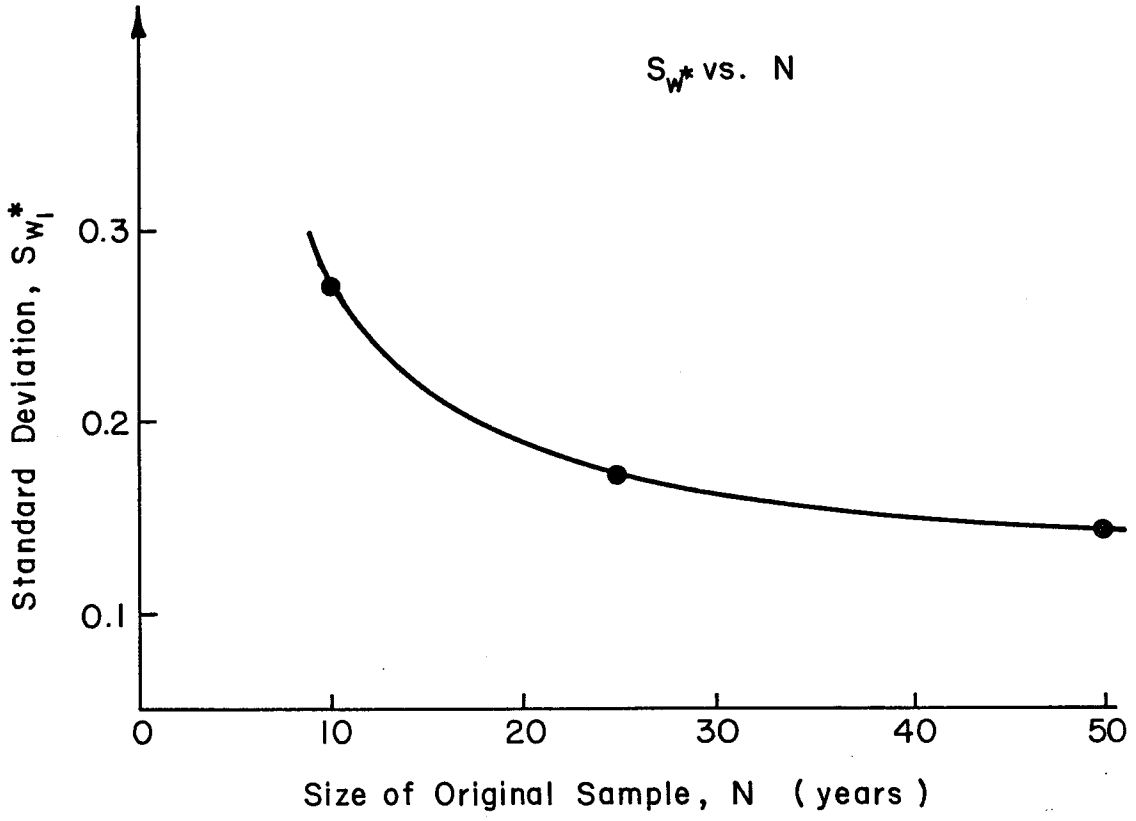


Figure 6. Standard Deviation of the Optimal Value of Weighting Factor, w_1, as Related to the Sample Length Used to Compute the Statistics.

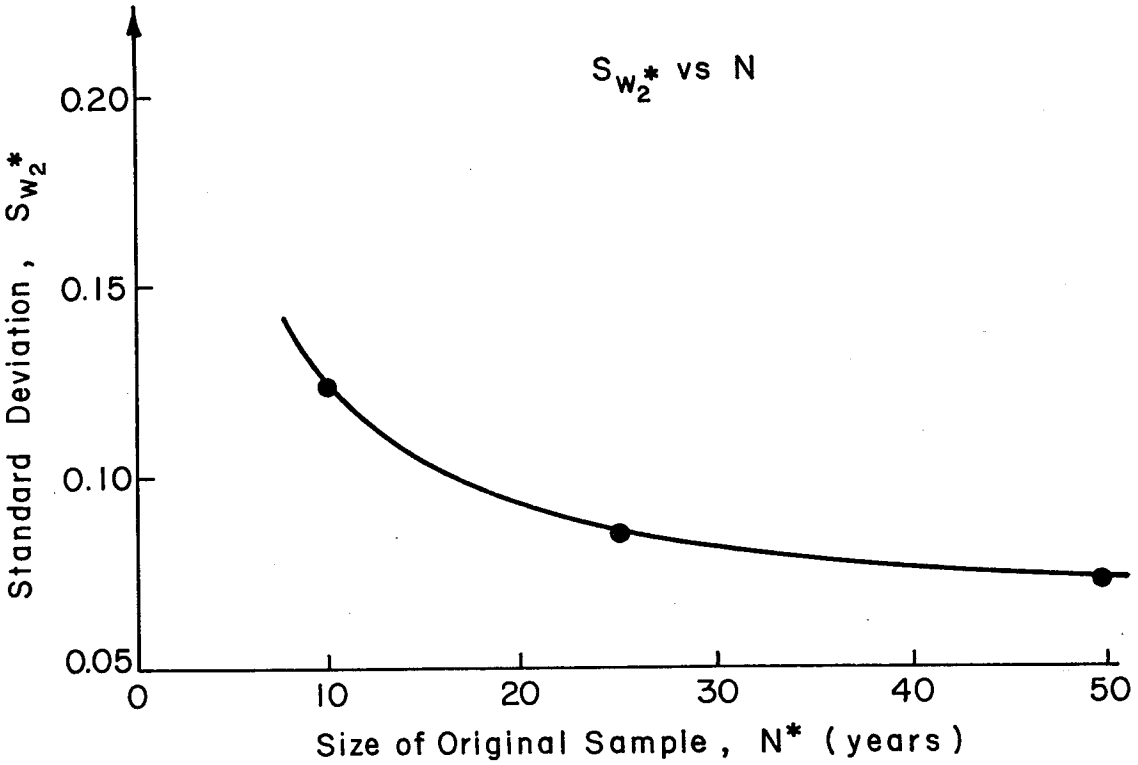


Figure 7. Standard Deviation of the Optimal Value of Weighting Factor, w_2, as Related to the Sample Length Used to Compute the Statistics.

The distribution of the optimal value of the objective function, N^*, is also related to the sample size used to compute the statistics. As noted from Table 15, all of the parameter estimates used to characterize the distribution with the exception of the mean, depend on the sample size. In particular, the standard deviation and skewness estimated from the 500 sets of generated data decrease as the sample size increases. Still, in each case, the skewness is sufficiently high to prevent assuming the objective function is normally distributed even for larger samples. The shape of the distribution of N^* can be seen in Figure 8, which represents a histogram of generated values for a sample size of 25 years. This distribution is, by both physical and mathematical considerations, bounded by zero and is skewed to the right thus it may be approximated by a lognormal or gamma distribution both of which have these properties and are relative easily applied.

Of particular note in this case, however, is the fact that the value of N^* is given by a piecewise function (Equation 21) due to the non-negativity conditions. Thus, depending on the values assumed by the weighting factors which are dictated by the sample statistics, N^* should be characterized by a piecewise rather than a single distribution. This would not be evident from the histogram plot and only comes to light from the analytical solution of the problem.

The distributions of the weighting factors, w_1 and w_2 are also bounded by zero due to the non-negativity conditions imposed in the problem. The larger weight, w_1, never assumes a zero value and thus has a continuous distribution. However, the smaller weight, w_2, frequently is given a zero value and thus, its distribution has a concentration of mass at the point $w_2 = 0$ and is therefore a mixed distribution as

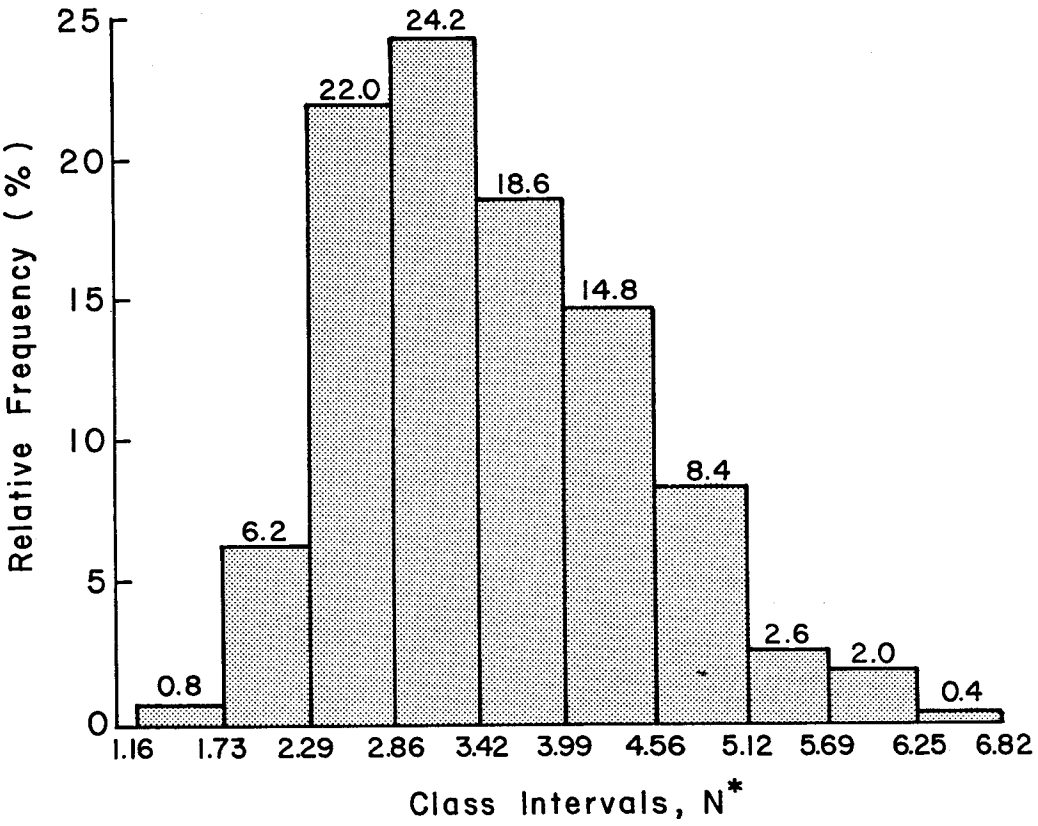


Figure 8. Histogram of N^* for a Sample Length of 25 Years and Non-Negativity Conditions.

described by Yevjevich (1972). Figure 9 and 10 show histograms of w_1 and w_2 respectively corresponding to a sample size of 25 values used to compute the weights.

The size and power of the test as estimated from the data generation study were also dependent on the sample size used to compute the weighting factors. This is due to the variability in the weights which cause the linear combinations to be random variable with a distribution other than Gaussian. As a result, the test does not always have the originally prescribed size and power. Table 16 presents a summary of the behavior of the size and power as sample size is increased. The change in the mean value of these variables is shown in Figures 11 and 12. For the small samples, the size of the test was far from the 0.05 value selected initially. With larger sample size, however, the size falls within the selected limits of 0.05 ± 0.02 with probability 95 percent. Thus, as would be expected, the actual size approaches the selected size but only for large sample sizes and so the probability of rejecting the null hypothesis when it is in fact true is much greater for small samples. The power is within the prescribed limits, 0.50 ± 0.05, in all cases and is not very sensitive to sample size, at least for a 10 percent change in the mean.

Imposing non-negativity conditions on the weighting factors in this problem was not necessary from either a physical or mathematical point of view. As such, it only removes the complete freedom of the variables to minimize the objective function. A simulation run was made to compare the effects of the non-negativity conditions on the problem and the results are summarized in Table 17. From this it is apparent that the mean value of the objective function is somewhat smaller in the free

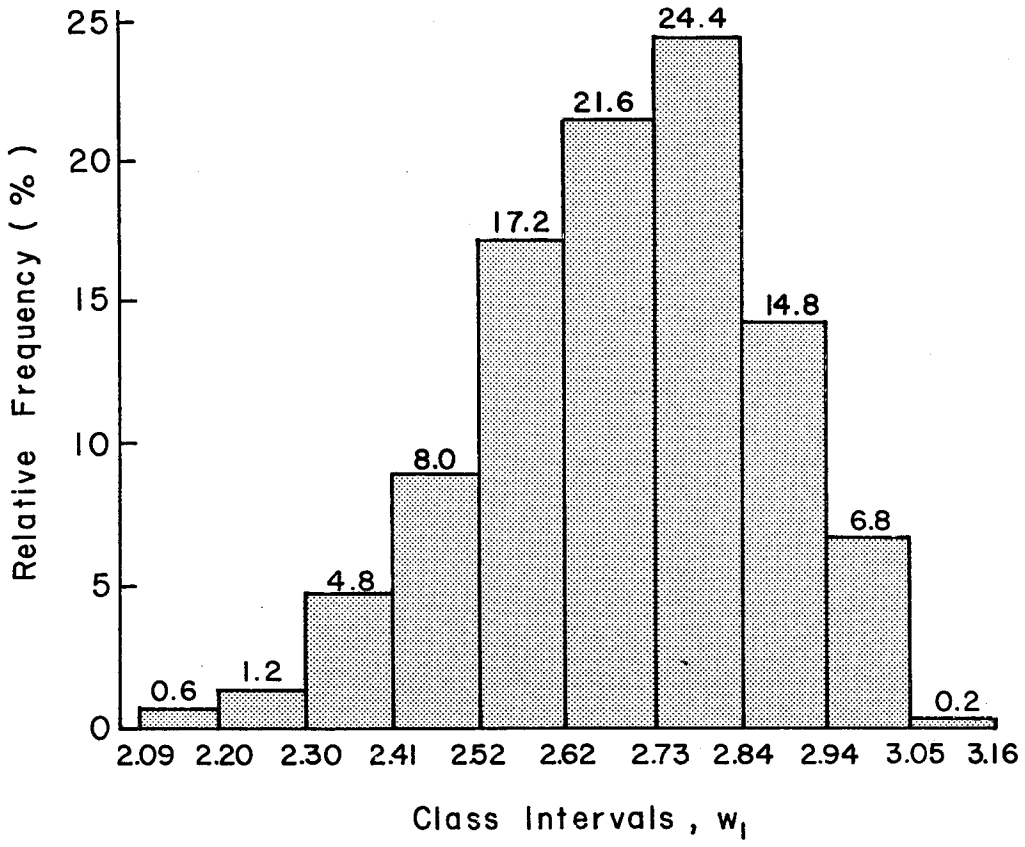


Figure 9. Histogram for Weighting Factor, w_1, from Simulation with Sample Length of 25 Years and Non-Negativity Conditions.

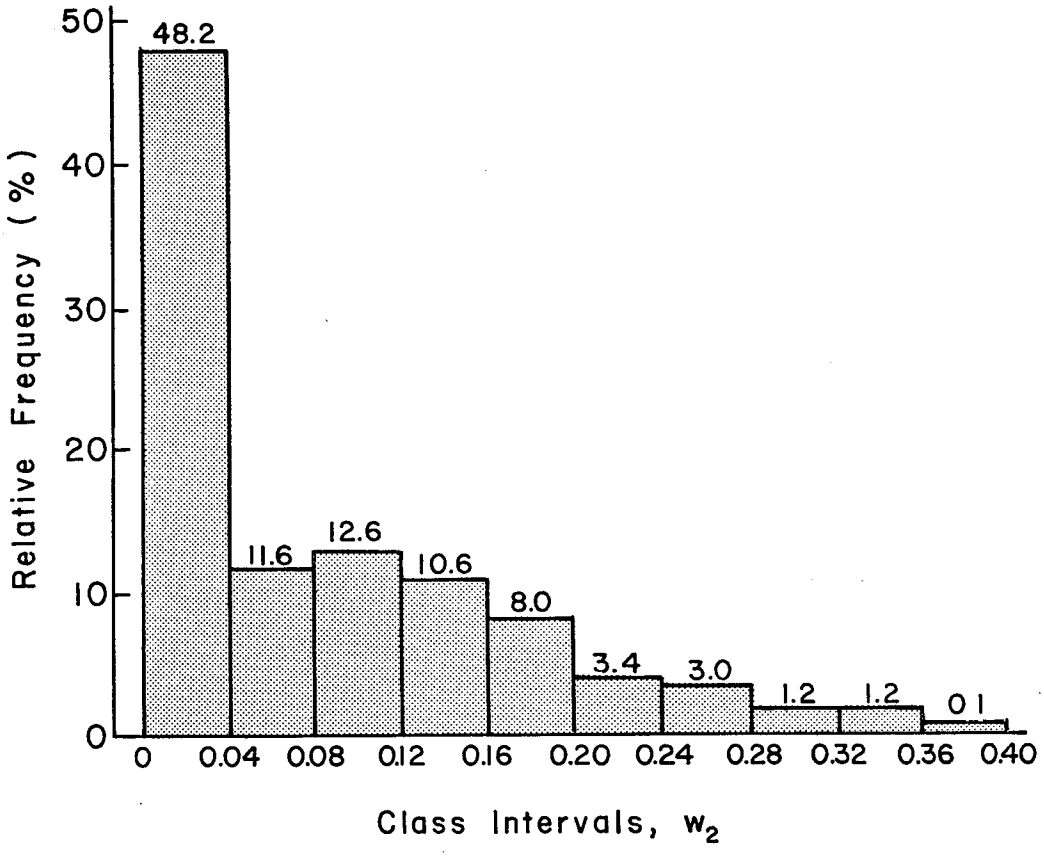


Figure 10. Histogram for Weighting Factor, w_2, from Simulation with Sample Length of 25 Years and Non-Negativity Conditions.

Table 16

Summary of the Size and Power of the Test as a Function of Original Sample Size

Run No.	Sample Size (years)	Variable	Mean	Standard Deviation	Skewness
2	10	$\hat{\alpha}$	0.101	0.101	1.898
		$\hat{1-\beta}$	0.531	0.225	0.020
3	25	$\hat{\alpha}$	0.068	0.057	1.441
		$\hat{1-\beta}$	0.545	0.165	0.004
4	50	$\hat{\alpha}$	0.060	0.038	1.000
		$\hat{1-\beta}$	0.542	0.125	-0.021

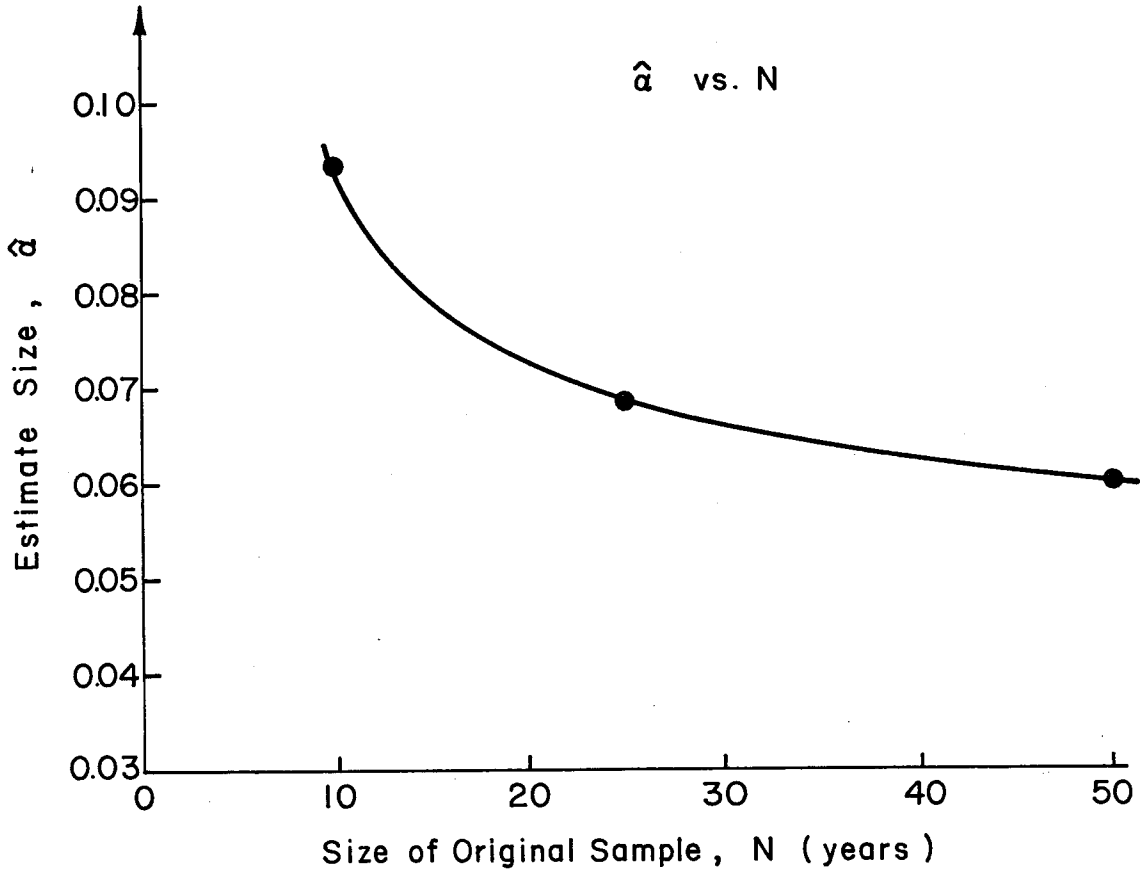


Figure 11. Estimated Size of the Test, $\hat{\alpha}$, as It Relates to the Original Sample Size, N.

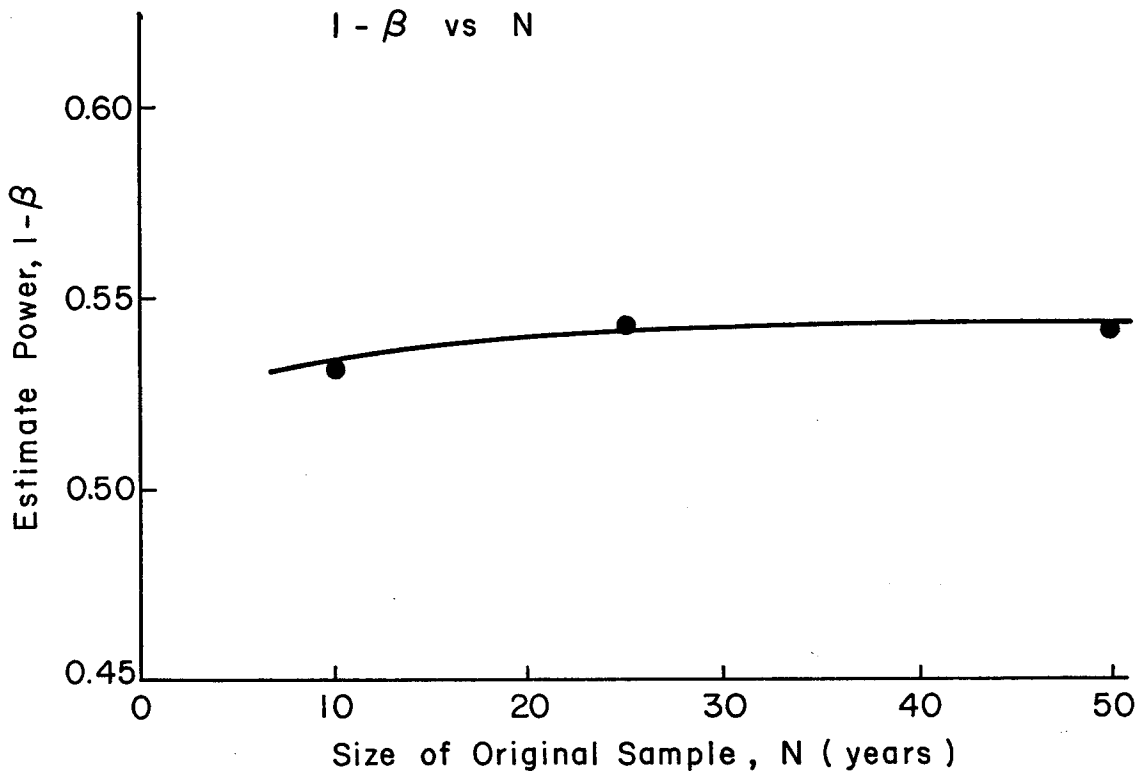


Figure 12. Estimated Power of the Test as It Relates to the Original Sample Size, N.

Table 17

Comparison of the Effects of Non-Negativity on the Objective Function and Optimization Variables

Run No.	Sample Size (years)	Variable	Mean	Standard Deviation	Skewness
1	25	N^*	3.268	0.954	0.499
		w_1	2.623	0.235	-0.264
		w_2	0.118	0.122	0.043
3	25	N^*	3.478	0.932	0.561
		w_1	2.696	0.173	-0.439
		w_2	0.075	0.085	1.035

variable case (Run No. 1) than in the more restrictive non-negative case (Run No. 3). In addition, in the free variable case, all of the variables have continuous, single function distributions rather than the piecewise and mixed distributions. Plots of the histograms for this case of N^*, w_1 and w_2 are presented in Figures 13, 14 and 15 for comparison. The distribution of N^* is somewhat skewed to the right whereas w_1 and w_2 are nearly symmetrically distributed as evidenced both by the histograms (Figures 14 and 15) and the skewness coefficients in Table 17.

Given the nature of the solutions as single functions in the free variable case, it is appropriate to explore the traditinal approach of fitting distributions to the objective function and weighting factors and evaluating the fit by the Chi-Square goodness-of-fit test. This will provide a basis for estimating confidence intervals on these variables to further quantify the reliability of the results. Since N^* is rather obviously skewed to the right and is bounded by zero, appropriate distributions to test are the log normal and gamma distribution. These both have the required properties, are well known and fairly easy to use. The weighting factors, w_1 and w_2, however, are not markedly shewed and it is appropriate to test the normal distribution for these variables. The results of the distribution fitting are summarized in Table 18. The gamma distribution was selected as the best fit for N^* while both weighting factors were best fit with the normal distribution. Knowing the distribution, it is then possible to compute confidence intervals of the variables. Of particular interest is the optimal value of the objective function, N^* which can be estimated using a gamma distribution with parameters taken from the results of

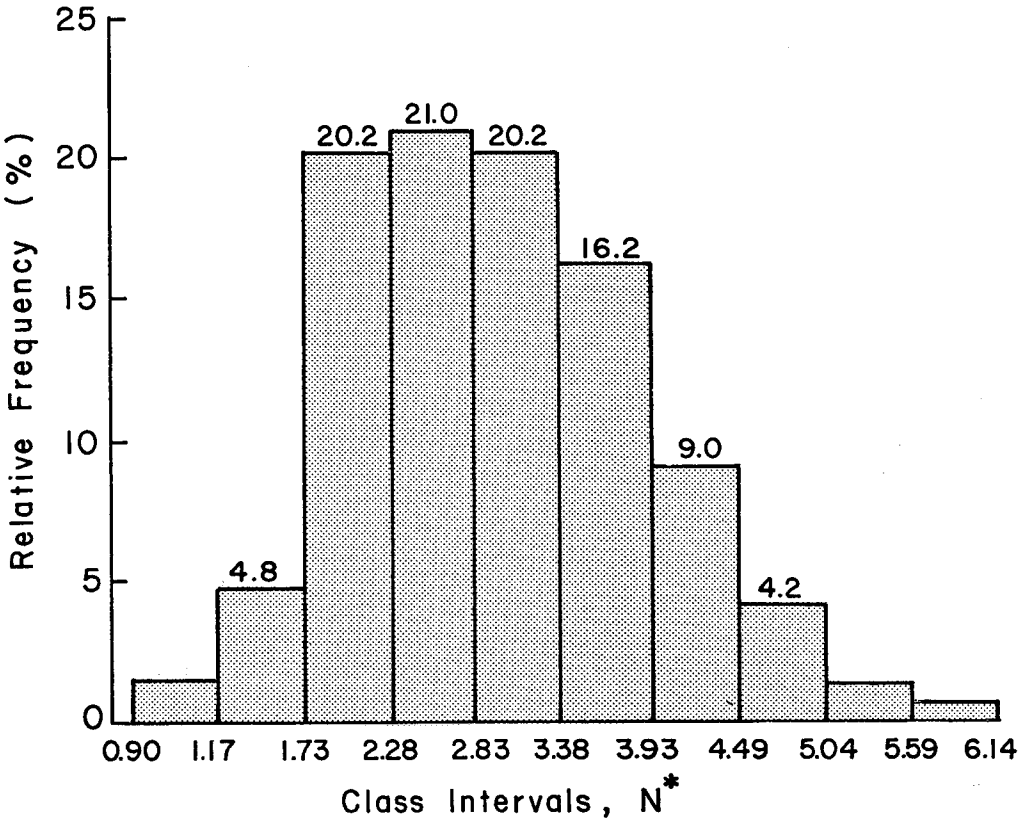


Figure 13. Histogram of the Optimal Value of the Objective Function, N^*, for the Free Variable Case with an Original Sample Length of 25 years.

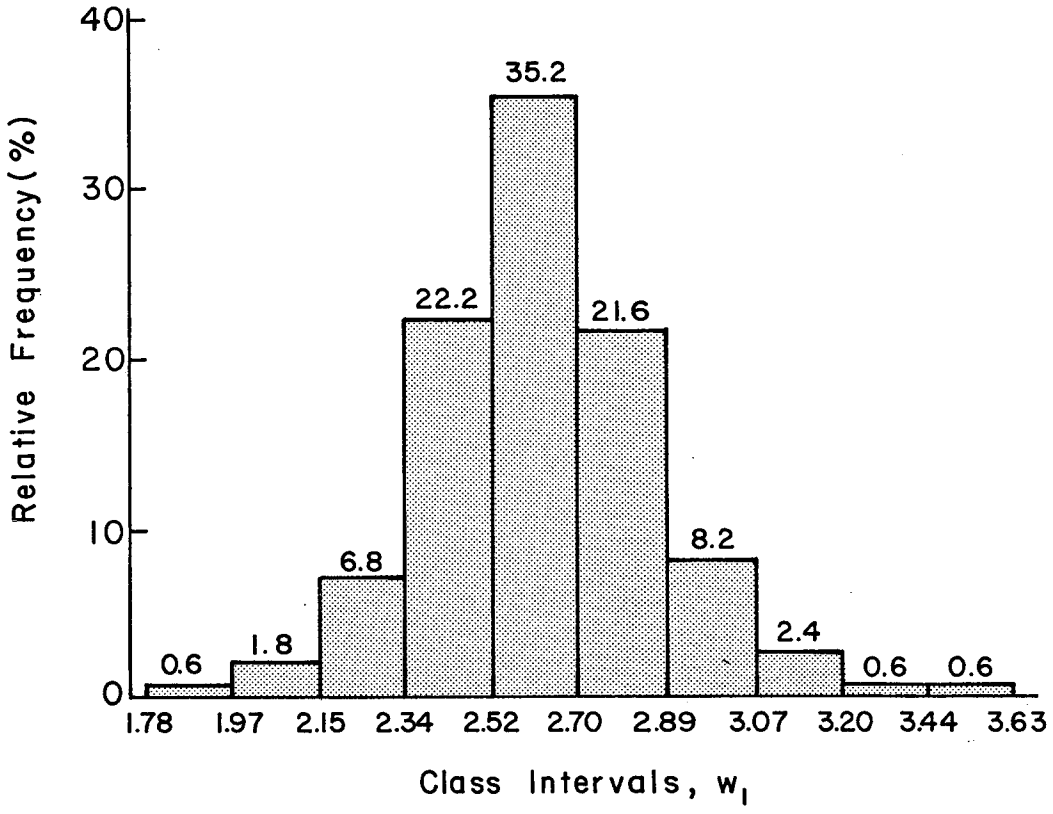


Figure 14. Histogram of Weighting Factor, w_1, in the Free Variable Case with an Original Sample Length of 25 years.

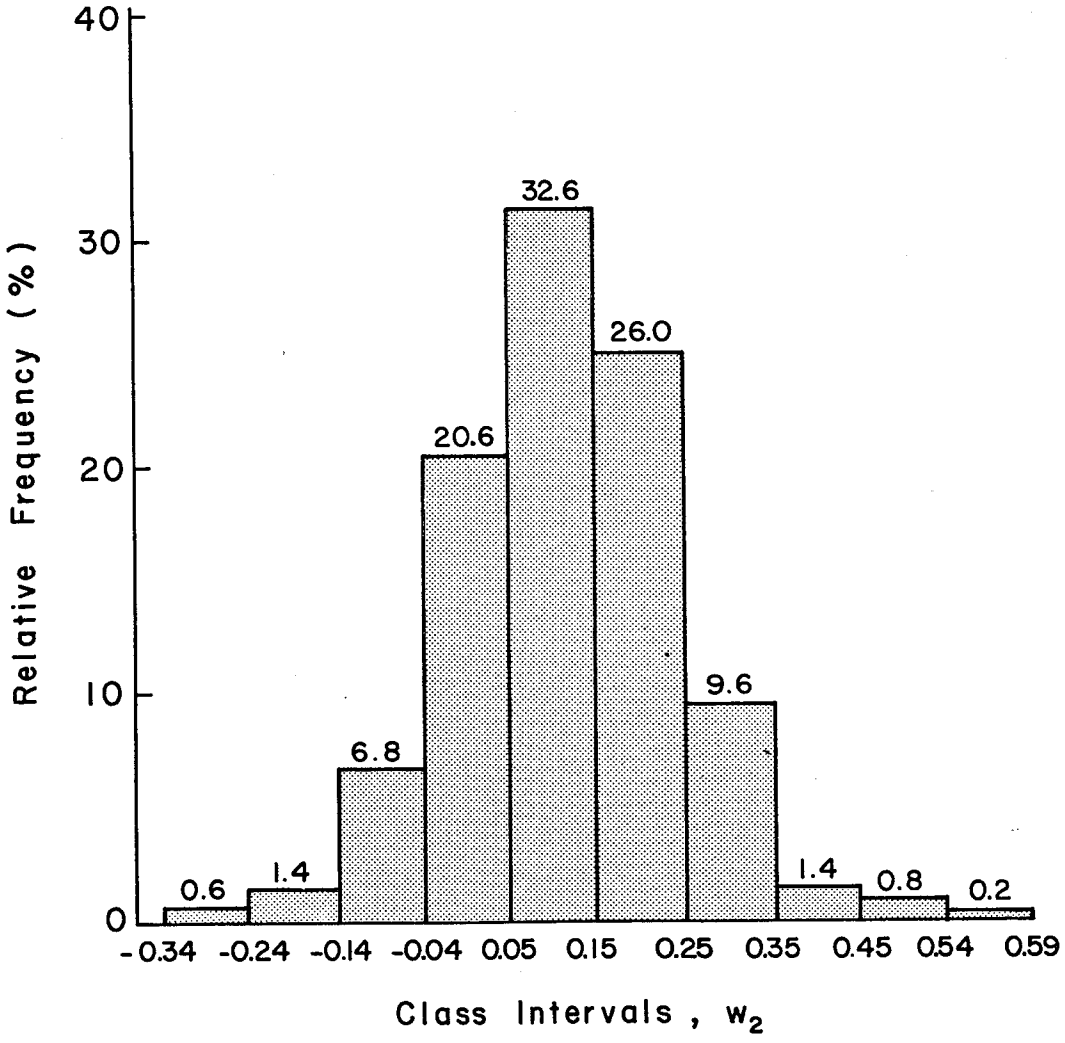


Figure 15. Histogram of Weighting Factor, w_2, in the Free Variable Case with an Original Sample Length of 25 years.

Table 18

Results of the Chi-Squared Goodness-of-Fit Tests[1/]

Variable	Distribution	Computed χ^2	Critical χ^2 ($\alpha = 0.05$)	Result
N^*	Gamma	9.80	14.07	Accept
	Log normal	45.40	14.07	Reject
w_1	Normal	9.56	14.07	Accept
	Log normal	46.2	14.07	Reject
w_2	Normal	9.80	14.07	Accept

[1/] All tests based on 7 degrees of freedom with $\alpha = 0.05$.

simulation Run No. 1. In this Case, the 95 percent equal tail area confidence interval is given by $(1.73 < N^{*} < 5.52)$. The width of the interval is then 3.79 years. If a conservative estimate of the sample size is desired, the upper confidence limit could be used rather than the value obtained from solving the optimization problem.

The size of the test also seems to be affected by the lack of non-negativity conditions while the power is relatively unaffected. A summary of the results of the data generation study regarding the affects of non-negativity condition on the size and power is presented in Table 19. The mean value of the size for the 500 simulated tests with the weights as free variables was somewhat greater than that resulting from the non-negative case. In fact, it can be said with some certainty that the size of the test based on a 25 year sample size is not within the 0.05 ± 0.02 limits. The power of the test was hardly affected in either case.

Table 19

Summary of the Size and Power of the Test in Relation to Non-Negativity

Run No.	Sample Size (years)	Variable	Mean	Standard Deviation	Skewness
1	25	$\hat{\alpha}$	0.077	0.060	1.196
		$\hat{1-\beta}$	0.550	0.159	-0.025
3	25	$\hat{\alpha}$	0.068	0.057	1.441
		$\hat{1-\beta}$	0.545	0.165	0.004

SUMMARY AND CONCLUSIONS

Detection Method

A method for detecting a change in a hydrologic variable originally proposed by Morel-Seytoux and Saheli (1973) has been adapted to water quality. The basis of the approach is the use of weighted linear combinations of variables together with a target-control (regression) test to decrease the time required to detect a specified level of change. The weights are selected using a mathematical programming approach which minimizes the time required for detection. The optimization routine developed for use in this study uses an iterative quadratic programming (QP) approach to solve the problem because of the highly nonlinear nature of the objective function. Tests of the approach on example problems indicated that the procedure reduces significantly the time required to detect a change especially when the data being considered was highly variable.

The behavior of the method depends on the variability of the water quality variables and their interrelationships. Within the target area, those variables with the lowest coefficients of variation receive the largest weights. High correlation between variables in the target area tends to give further weight to the variables having the lower coefficients of variation. Variables within the control area are weighted more highly if they have a low coefficient of variation, high correlation with stations in the target area and low correlation with other control stations. These rules can be used in the initial selection of stations for an application of the method.

Although unexplored in this study, the usefulness of this technique in problems other than a strict detection of change application can be

envisioned. In particular, this approach may be used in evaluating the effectiveness of various stations in a gaging network for use in detecting changes. Those stations with very low weights could be immediately seen to be ineffective for such a purpose. Further, the method is suitable for analyzing which variable or interrelationships of variables would be most useful for rapid detection of changes.

Reliability

If a complete knowledge of the behavior of the system were possible, the detection method would provide the complete answer to the problem. However, the information employed in deriving the optimal values of the objective function (number of years) and the weighting factors are sample estimates. Therefore the results of the optimization problem, being functions of these sample statistics, are also random variables. For a complete resolution of the problem, an assessment of the reliability, that is how the results can be expected to vary due to uncertainty in the sample statistics, is necessary.

A method has been presented for evaluation of the distributions of these optimized random variables based on an analytical solution of the optimization problem. From this approach, explicit expressions relating the objective function and variables of the optimization problem to the random coefficients and "constants" can be obtained. These relations can then be used to evaluate how the random variables combine to produce uncertainty in the optimization problem. The imposition of non-negativity conditions complicates this procedure by producing piecewise solutions depending on the values assumed by the random variables. This, however, is the case most often encountered in engineering and management applications. In theory, the explicit expressions are then

used to derive analytically the distributions of the optimized random variables. However, when there are many random variables in the expressions or if they are combined in a complicated manner, simulation may be required to estimate the uncertainty.

A simpler formulation of the detection problem is used to demonstrate this approach. Even in the simplest case, for two stations, the resulting analytical expressions become mathematically intractable. At that point, a few simulation runs for the free variable case are done to demonstrate, in this case, what information can be gained on the variability of the objective function and weighting factors. Only the free variable case was used in the simulation study since piecewise and mixed distributions resulted in the non-negative case.

An important result of the simulation study was the dependence of the number of observations (sample size) on the actual size and power of the test. For small samples, say 10 observations, the actual size of the test was significantly larger than that specified a priori. This is due to the use of sample statistics rather than population parameters to derive the weighting factors and results in a much larger Type I error than anticipated. To adjust for this problem, a smaller level of significance than actually required can be used in the analysis and the resulting test will very likely have a size closer to the desired value. For example, if a 0.05 level of significance is thought appropriate, the weights could be determined using a 0.03 value and the resulting test would have closer to the desired property. This will also effectively increase the sample size required to detect the specified change.

The major advantage of the method is in the analytical relationships produced between the variable of the optimization problem and the random variables used as coefficients and constants. At the

very least, these expressions allow a qualitative assessment of how the uncertainty in estimating these values affects the problem. For less complex formulations, these expressions can lead to analytical solution for the distribution of the optimized random variables and thus general conclusions can be drawn. For more complicated situations, simulation using these explicit expressions can be used to produce conclusions specific to the problem at hand. Further study is required to explore what formulations produce mathematically tractable solutions. In addition, dealing with the piecewise and mixed distributions resulting from the non-negativity conditions must be addressed.

REFERENCES

Beightler, C. S., D. T. Phillips and D. S. Wilde, "Foundations of Optimization," 2nd Editions, Prentice-Hall, Inc., 1979.

Haan, C. T., "Statistical Methods in Hydrology," Iowa State University Press, 1977.

Jønch-Clausen, T. and H. J. Morel-Seytoux, "User's Manual for QPTOR, a FORTRAN IV Quadratic Programming Routine," HYDROWAR Program, Colorado State University, Fort Collins, Colorado, CER76-77TJ-HJM48, 1978.

Koch, R. W., "User's Manual for IITQP, a FORTRAN IV Iterative Quadratic Programming Routine," Colorado State University, Fort Collins, Colorado, 1980.

Matalas, N. C. and W. B. Laugbein, Information Content of the Mean, Journal of Geophysical Research, 67(9): 3441-3448, 1962.

Mood, A. M., F. A. Graybill and D. C. Boes, "Introduction to the Theory of Statistics," 3rd Edition, McGraw-Hill Book Co., 1974.

Morel-Seytoux, H. J., Optimal Legal Conjunctive Operation of Surface and Ground Waters, Proceedings 2nd World Congress; International Water Resources Association, New Delhi, India, December, 1975.

Morel-Seytoux, H. J., "Foundations of Engineering Optimization," Lecture Notes, Department of Civil Engineering, Colorado State University, 1972, revised 1976, revised 1978.

Morel-Seytoux, H. J. and F. Saheli, "Test of Runoff Increase Due to Precipitation Management for the Colorado River Basin Pilot Project," Journal of Applied Meteorology, Vol. 12, No. 2, March, 1973.

Salas, J. D., J. W. Delleur, V. Yevjevich and W. L. Lane, "Applied Modeling of Hydrologic Time Series," Water Resources Publications, 1980.

Yevjevich, V., "Probability and Statistics in Hydrology," Water Resources Publications, 1972.

Yevjevich, V., "Stochastic Processes in Hydrology," Water Resources Publication, 1972.

APPENDIX A

MATHEMATICAL DEVELOPMENT OF EQUATIONS USED TO DETERMINE WEIGHTING COEFFICIENTS

The basic equation used as the objective function for the optimization problem is:

$$N = (z_{1-\beta} + z_{1-\alpha})^2 \left(\frac{\sigma}{k\mu_o}\right)^2 \qquad \text{(A-1)}$$

where N is the number of years required to detect the change in the mean

$z_{1-\beta}$ is the standard normal variate corresponding to the power of the test

$z_{1-\alpha}$ is the standard normal variate corresponding to the size of the test (level of significance)

μ_o is the mean of the variable specified in the null hypothesis

μ_1 is the mean of the variable specified in the alternate hypothesis

This equation applies to the detection of a given change in the mean value of a variable which is normally distributed, independent in time and whose variance does not change.

The derivation of this equation for a one sided test such as

$$H_o: \ \mu = \mu_o \quad \text{vs.} \quad H_1: \ \mu = \mu_1 > \mu_o$$

is as follows. First, define the level of significance of the test to be α; the probability of committing a Type I error (rejecting H_o when it should be accepted). Also define $1-\beta$ as the power of the test; the probability of accepting H_1 given that it is true. We can now proceed to derive the Equation (A-1) given the properties of the normal distribution.

Referring to Figure A-1 for a graphical representation of the following mathematics, let us first define the critical value $\bar{x}_c$ as:

α Size of the Test (Level of significance)

$1-\beta$ Power of the Test, P [Accept H_1 | H_1 True]

x_c Critical Value of the Variable x

μ_0 Mean of the Null Hypothesis, H_0

μ_1 Mean of the Alternate Hypothesis, H_1

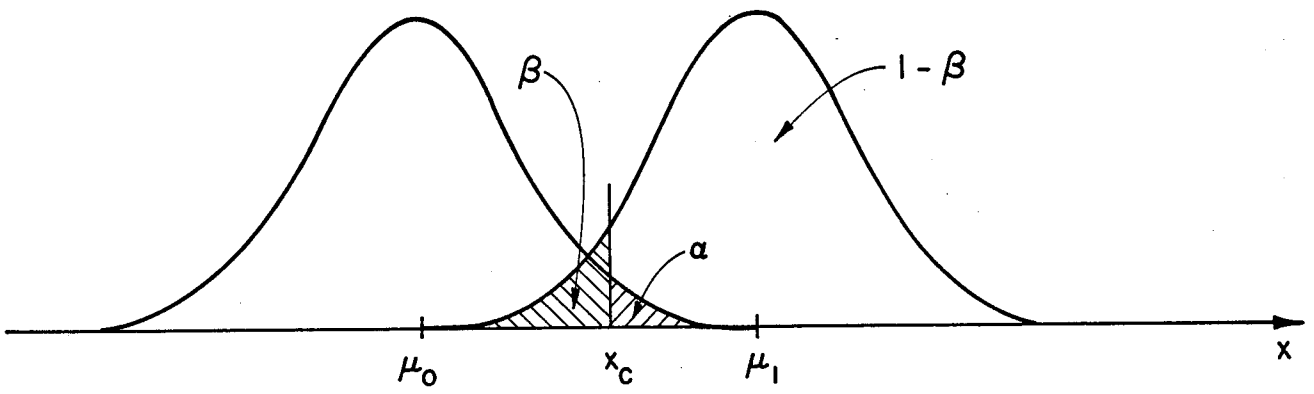


Figure A-1. Graphical Representation of the Statistical Theory.

$$\bar{x}_c = \mu_o + z_{1-\alpha}\ \sigma/\sqrt{n} \qquad \text{(A-2)}$$

Further, the power of the test is simply the area under the normal p.d.f. (centered at μ_1) to the right of $\bar{x}_c$ or the probability that $\bar{x} > \bar{x}_c$ given that $\mu = \mu_1$. This can be written as:

$$1 - \beta = \frac{1}{\sqrt{2\pi}\ \sigma/\sqrt{n}} \int_{\bar{x}_c}^{\infty} \exp\left[-\frac{1}{2}\left(\frac{\bar{x}-\mu_1}{\sigma/\sqrt{n}}\right)^2 dx\right]$$

or

$$1 - \beta = 1 - \frac{1}{\sqrt{2\pi}\ \sigma/\sqrt{n}} \int_{-\infty}^{x_c} \exp\left[-\frac{1}{2}\left(\frac{\bar{x}-\mu_1}{\sigma/\sqrt{n}}\right)^2 dx\right]$$

$$= 1 - \Phi\left(\frac{\bar{x}_c-\mu_1}{\sigma/\sqrt{n}}\right) \qquad \text{(A-3)}$$

where $\Phi(.)$ represents the normal cumulative distribution function. Given the expression for $\bar{x}_c$, Equation (A-2), we can replace $\bar{x}_c$ in Equation (A-3) to obtain:

$$1 - \beta = 1 - \Phi\left(\frac{\mu_o + z_{1-\alpha}\ \sigma/\sqrt{n} - \mu_1}{\sigma/\sqrt{n}}\right)$$

which simplifies to:

$$1 - \beta = 1 - \Phi\left(\frac{\mu_o-\mu_1}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \qquad \text{(A-4)}$$

Now, due to the symmetry of the normal distribution, $\Phi(z) = 1-\Phi(-z)$ so we can rewrite Equation (A-4) as:

$$1 - \beta = \Phi\left(-z_{1-\alpha} + \frac{\mu_1-\mu_o}{\sigma/\sqrt{n}}\right)$$

and since $-z_{1-\alpha} = z_{\alpha}$,

$$1 - \beta = \Phi\left(z_{\alpha} + \frac{\mu_1-\mu_o}{\sigma/\sqrt{n}}\right) \qquad \text{(A-5)}$$

If we view Φ a linear operator (with reference to Figure A-1), we see that operating on both sides of Equation (5) with Φ^{-1} will transform $1-\beta$ from a probability to a standard normal variate, $z_{1-\beta}$. Likewise, the right side will be transformed to give

$$z_{1-\beta} = z_{\alpha} + \frac{\mu_1-\mu_o}{\sigma/\sqrt{n}} \qquad \text{(A-6)}$$

or, rearranging we can obtain

$$n = \left(\frac{z_{1-\beta}-z_{1-\alpha}}{\mu_1-\mu_o}\right)^2 \sigma^2 \qquad \text{(A-7)}$$

This expression allows us to determine the number of observations, n, necessary to detect a change in the mean from μ_o to μ_1, with a level of significance, α, and a power, $1-\beta$. Noting that $z_{\alpha} = -z_{1-\alpha}$, we can make a final substitution to obtain

$$n = \left(\frac{z_{1-\beta}+z_{1-\alpha}}{\mu_1-\mu_o}\right)^2 \sigma^2 \qquad \text{(A-8)}$$

Note that $z_{1-\beta}$ and $z_{1-\alpha}$ are both positive for α and $\beta < 0.50$.

One further change of Equations (A-8) is useful. If we define the difference between μ_o and μ_1 as a change relative to μ_o, we can write

$$k = \frac{\mu_1-\mu_o}{\mu_o}$$

so that $\mu_1-\mu_o = k\mu_o$.

Substituting this expression into Equation (A-8) yields Equation (A-1)

$$n = \left(\frac{z_{1-\beta}+z_{1-\alpha}}{k\mu_o}\right)^2 \sigma^2 \qquad \text{(A-1)}$$

This relation expresses the number of observations necessary to detect a change in μ_o of $k\mu_o$ for the selected power and size of the test. As an example, a value of $k = 0.10$ indicates an increase in μ_o of 10 percent so the test is one which will detect a 10 percent increase in the mean.

A similar relation can be derived for a composite test (two tailed) with one minor assumption, to give

$$n \simeq \left(\frac{z_{1-\beta}+z_{1-\alpha}}{k\mu_o}\right)^2 \sigma^2 \qquad \text{(A-9)}$$

with the understanding that $\frac{k\mu_o}{\sigma/\sqrt{n}} > 0.5$. Otherwise a much more complex equation must be used requiring iterative solution techniques to determine "n".

In both Equations (A-1) and (A-9), the number of observations required to detect a given change in the mean is directly proportional to the variance of the random variable. Therefore a decrease in this parameter will lead to lesser number of required observations. One means of decreasing the variance is to find another variable which is (linearly) related to the first variable and is also normally distributed. This relationship can be exploited through the theory of the bivariate normal distribution to accomplish this objective.

The linear regression between two variables is based on the conditional, bivariate normal distribution, that is, the distribution of a random variable, Y, is a function of the value taken by the variable,

X. In this case Y is the dependent and X is the independent variable. The crux of this approach, is that, given a value of X, the variation in Y is reduced in proportion to the (linear) dependence between the two variables. The variance of this conditional distribution is

$$\sigma^2_{Y/X} = (1 - \rho^2_{XY})\sigma^2_Y$$

where $\sigma^2_{Y/X}$ is the variance of Y given X,

ρ_{XY} is the correlation coefficient between X and Y,

σ^2_Y is the variance of Y

So, if the degree of (linear) correlation as represented by ρ_{XY} is high, the conditional variance can be considerably smaller than the marginal variance.

This conditional variance was substituted for the variance of the original variable to give

$$n = (z_{1-\beta}+z_{1-\alpha})^2 \ (1-\rho^2_{XY}) \left(\frac{\sigma_Y}{k\mu_{oY}}\right)^2 \qquad \text{(A-10)}$$

This can be easily justified theoretically if the denominator is viewed as $\mu_1-\mu_o$ rather than $k\mu_o$. In this case $k\mu_{oY}$ also represents the change in the value of the conditional mean of Y given X. Equation (A-10) provides us with an expression for the number of observations required to detect a change of 100k percent in the mean of the random variable Y, given that it is correlated to another random variable, X, which does not change. This number of observations is obviously less than the number computed without using the conditional variance.

Thus far, the development of the methodology has been restricted to two variables. It was hypothesized and shown rather conclusively by Morel-Seytoux and Saheli (1973), that a weighted linear combination of

variables could be used effectively in further reducing the time required to detect a change in the random variable of interest. Thus, rather than a variable from a single location, a weighted linear combination of variables from several locations can be used as the random variable in the problem. We could define new variables X^* and Y^* to be linear combinations of the values of X and Y at various locations by:

$$Y^* = \sum_{i=1}^{NT} w_i Y_i \qquad \text{(A-11)}$$

$$X^* = \sum_{i=NT+1}^{NC} w_i X_i \qquad \text{(A-12)}$$

where w_i is the respective weighting factor,

NT is the number of stations in the target area, and

NC is the number of stations in the control area.

Then, writing Equation (A-1) for the linear combinations gives:

$$N = (z_{1-\beta}+z_{1-\alpha})^2 \ (1 - \rho^2_{Y^*X^*}) \left(\frac{\sigma_{Y^*}}{k\mu_{Y^*}}\right)^2 \qquad \text{(A-13)}$$

Now, the terms $\rho_{X^*Y^*}$, σ^{2*}_Y and μ_{Y^*} can be expanded as follows. For the correlation coefficient, $\rho_{X^*Y^*}$, the definition is:

$$\rho_{X^*Y^*} = \frac{\text{Cov}\ (Y^*X^*)}{\sigma_{X^*}\ \sigma_{Y^*}} \qquad \text{(A-14)}$$

Mood, Graybill and Boes (1974) give expressions for the covariance and variance and mean of linear combinations as:

$$\mathrm{Cov}[Y^*,X^*] = \mathrm{Cov}\left[\sum_{i=1}^{NT} w_i Y_i,\ \sum_{j=NT+1}^{NC} w_j X_j\right]$$

$$= \sigma_{i=1}^{NT} \sum_{j=NT+1}^{NT+NC} w_i w_j\ \mathrm{Cov}[Y_i,X_j] \qquad \text{(A-15)}$$

$$\sigma_Y^* = \mathrm{Var}\left[\sum_{i=1}^{NT} w_i Y_j\right] = \sum_{i=1}^{NT}\sum_{j=1}^{NT} w_i w_j\ \mathrm{Cov}[Y_i Y_j] \qquad \text{(A-16)}$$

$$\sigma_X^* = \mathrm{Var}\left[\sum_{i=NT+1}^{NT} w_i X_j\right] = \sum_{i=NT+1}^{NT+NC}\sum_{j=NT+1}^{NT+NC} w_i w_j\ \mathrm{Cov}[X_i X_j] \qquad \text{(A-17)}$$

$$\mu_Y^* = E\left[\sum_{i=1}^{NT} w_i Y_i\right] = \sum_{i=1}^{n} w_i \mu_{Y_i} \qquad \text{(A-18)}$$

Substituting these expressions into (A-13) gives

$$N = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{k \sum_{i=1}^{NT} w_i \mu_{Y_i}}\right)^2 \left[\sum_{i=1}^{NT}\sum_{j=1}^{NT} w_i w_j\ \mathrm{Cov}[Y_i Y_j]\right.$$

$$\left. -\left(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_j\ \mathrm{Cov}[Y_i X_k]\right)^2 \Big/ \sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NT+NC} w_k w_\ell\ \mathrm{Cov}(X_k X_\ell)\right] \qquad \text{(A-19)}$$

The original function, (A-1) is now expressed explicitly in terms of the weighting factors and the parameters of the variables.

APPENDIX B

QUADRATIC EXPANSION OF THE OBJECTIVE FUNCTION

Objective Function Expansion

The quadratic programming (QP) problem is generally posed in the following form

$$\underset{\underline{x}}{\text{Min}} \quad \{ \underline{c}'\underline{x} + \frac{1}{2} \underline{x}' Q \underline{x} \} \tag{B-1}$$

subject to: $\underline{x} \geq 0$

$A \underline{x} > \underline{r}$

Thus, a quadratic function is being minimized subject to a set of linear constraints. Non-negativity conditions may also be placed on the variables. Equation (B-1) can also be written in summation form as:

$$\underset{\underline{x}}{\text{Min}} \quad \sum_{i=1}^{N} c_i x_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j q_{ij} \tag{B-2}$$

where N is the number of variables.

The method selected to solve the nonlinear programming problem is to successively approximate the objective function at each feasible point by a quadratic function thus allowing the iterative use of a QP algorithm to solve the problem.

The objective function is approximated by expanding it in a Taylor series which is then truncated after the second order term. The multivariate form of the Taylor series expansion about the point $\underline{x}^\circ$ is given by

$$y(\underline{x}) = y(\underline{x}^\circ) + \sum_{i=1}^{N} (x_i - x_i^\circ) \left(\frac{\partial y}{\partial x_i}\right)^\circ$$

$$+ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (x_i - x_i^\circ)(x_j x_j^\circ) \left(\frac{\partial^2 y}{\partial x_i \partial x_j}\right)^\circ + 0(x^3) \tag{B-3}$$

where y is the dependent variable

$\underline{x}$ is the vector of independent variable

$0(x^3)$ represents the remaining terms of order x^3 and greater

The derivatives are also evaluated at the point, $\underline{x}^\circ$. If the terms of $0(x^3)$ are truncated, a quadratic approximation of the original function remains; the accuracy of the approximation depends on the magnitude of the truncated terms. Thus if the objective function, Equation (10) is approximated in this fashion, the following expression results:

$$N(w) \tilde{} N(\underline{w}^\circ) + \sum_{i=1}^{N} (w_i - w_i^\circ) \left(\frac{\partial N}{\partial w_i}\right)^\circ + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (w_i - w_i^\circ)(w_j - w_j^\circ) \left(\frac{\partial^2 N}{\partial w_i \partial w_j}\right)^\circ \qquad \text{(B-4)}$$

where $\underline{w}$ is the vector of unknown weighting coefficients, and

$\underline{w}^\circ$ is the value of $\underline{w}$ at the feasible point.

Inspecting this expression reveals that further simplification produces an equation with constant, linear and quadratic terms. Identifying the coefficients for this terms yields for the constant term

$$B = \sum_{i=1}^{N} (-w_i^\circ) \left(\frac{\partial N}{\partial w_i}\right)^\circ + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_i^\circ w_j^\circ \left(\frac{\partial^2 N}{\partial w_i \partial w_j}\right)^\circ \qquad \text{(B-5)}$$

for the linear term

$$c_i = \left(\frac{\partial N}{\partial w_i}\right)^\circ - \sum_{j=1}^{N} w_j^\circ \left(\frac{\partial N}{\partial w_i \partial w_j}\right)^\circ \qquad \text{(B-6)}$$

and, for the quadratic term

$$q_{ij} = \left(\frac{\partial^2 N}{\partial w_i \partial w_j}\right)^\circ \qquad \text{(B-7)}$$

Combining these terms produces the complete expression for the approximation of the objective function as:

$$N \tilde{\sim} B + \sum_{i=1}^{N} c_i w_i + \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j q_{ij} \qquad \text{(B-8)}$$

It should be noted that the constant term, B, does not affect the optimization problem. However, it is necessary to check the accuracy of the approximation vs. the actual objective function.

Evaluation of Derivatives

From Equations (B-5), (B-6) and (B-7), it is apparent that both the first and second derivatives of the objective function are necessary to obtain the quadratic approximation. Recalling that the objective function is given by

$$N = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{k \sum_{i=1}^{n} \mu_{y_i}} \right)^2 \left[\sum_{i=1}^{NT} \sum_{j=1}^{NT} w_i w_j \, \mathrm{Cov}(Y_i Y_j) - \left(\sum_{i=1}^{NT} \sum_{k=NT+1}^{NT+NC} w_i w_k \, \mathrm{Cov}(Y_i X_k) \right)^2 \Big/ \sum_{k=NT+1}^{NT+NC} \sum_{\ell=NT+1}^{NT+NC} c_k c_\ell \, \mathrm{Cov}(X_k X_\ell) \right] \qquad \text{(B-9)}$$

The form of the derivations will depend on whether the particular variable being considered, w_p, is in the target area, $1 \leq p \leq NT$, or the control area, $NT + 1 \leq p \leq NT + NC$. For the first derivatives there are two possible cases:

$$\frac{\partial N}{\partial w_p} \quad , \quad 1 \leq p \leq NT$$

$$\frac{\partial N}{\partial w_p} \quad , \quad NT + 1 \leq p \leq NT + NC.$$

However, for the second derivatives there are four combinations given as:

$$\frac{\partial}{\partial w_q}\left(\frac{\partial N}{\partial w_p}\right), \quad 1 \le p \le NT \quad \& \quad 1 \le q \le NT \tag{B-10}$$

$$\frac{\partial}{\partial w_q}\left(\frac{\partial N}{\partial w_p}\right), \quad 1 \le p \le NT \quad \& \quad NT+1 \le q \le NT+NC \tag{B-11}$$

$$\frac{\partial}{\partial w_q}\left(\frac{\partial N}{\partial w_p}\right), \quad N+1 \le p \le NT+NC \quad \& \quad 1 \le q \le NT \tag{B-12}$$

$$\frac{\partial}{\partial w_q}\left(\frac{\partial N}{\partial w_p}\right), \quad NT+1 \le p \le NT+NC \quad \& \quad NT+1 \le q \le NT+NC \tag{B-13}$$

Noting that (B-11) and (B-12) produce the same result, there are five derivations that must be computed. The results of these computations are presented below:

$$\frac{\partial N}{\partial w_p} = A\left\{ 2\sum_{j=1}^{NT} w_j \operatorname{Cov}(Y_p Y_j) - 2\left(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NM} w_i w_k \operatorname{Cov}(y_i x_k)\right)\left(\sum_{k=NT+1}^{NT+NM} w_k \operatorname{Cov}(y_p x_k)\right) \right.$$
$$\left. \Big/ \left(\sum_{k=NT+1}^{NT+NM}\sum_{\ell=NT+1}^{NT+NM} w_k w_\ell \operatorname{Cov}(X_k X_\ell)\right)\right\}, \quad 1 \le p \le NT \tag{B-14}$$

$$\frac{\partial N}{\partial w_p} = A\left\{ \left[2\left(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \operatorname{Cov}(Y_i X_k)\right)\left(\sum_{\ell=NT+1}^{NT+NC} w_\ell \operatorname{Cov}(X_p X_\ell)\right) \right.\right.$$
$$- 2\left(\sum_{k=NT+1}^{NT+NM}\sum_{\ell=NT+1}^{NT+NM} w_k w_\ell \operatorname{Cov}(X_k X_\ell)\right)$$
$$\left.\left. \left(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \operatorname{Cov}(Y_i X_k)\right)\left(\sum_{i=1}^{NT} w_i \operatorname{Cov}(Y_i X_p)\right)\right]\right.$$
$$\left. \Big/ \left(\sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NT+NC} w_k w_\ell \operatorname{Cov}(X_k X_\ell)\right)^2 \right\}, \quad NT+1 \le p \le NT+NM \tag{B-15}$$

$$\frac{\partial^2 N}{\partial w_q \partial w_p} = A \left\{ 2\ \mathrm{Cov}(Y_p Y_q) - 2\left(\sum_{k=NT+1}^{NT+NC} w_k\ \mathrm{Cov}(Y_q X_k)\right)\left(\sum_{k=NT+1}^{NT+NM} w_k\ \mathrm{Cov}(Y_p X_k)\right) \right.$$

$$\left. \Big/ \left(\sum_{k=NT+1}^{NT+NM}\ \sum_{\ell=NT+1}^{NT+NM} w_k w_\ell\ \mathrm{Cov}(X_k X_\ell)\right) \right\},$$

$$1 \le p \le NT, \quad 1 \le q \le NT \tag{B-16}$$

$$\frac{\partial^2 N}{\partial w_q \partial w_p} = A \left\{ \left[4\left(\sum_{i=1}^{NT}\ \sum_{k=NT+1}^{NT+NM} w_k w_\ell \mathrm{Cov}(Y_i X_k)\right)\left(\sum_{k=NT+1}^{NT+NM} w_k \mathrm{Cov}(X_p X_k)\right)\left(\sum_{\ell=NT+1}^{NT+NM} w_k \mathrm{Cov}(Y_p X_k)\right) \right.\right.$$

$$- 2\left(\sum_{k=NT+1}^{NT+NM}\ \sum_{\ell=NT+1}^{NT+NC} w_k w_\ell \mathrm{Cov}(X_k X_\ell)\right)\left[\left(\sum_{i=1}^{NT} w_i \mathrm{Cov}(Y_i X_q)\right)\left(\sum_{k=NT+1}^{NT+NC} w_k \mathrm{Cov}(Y_p X_k)\right) \right.$$

$$\left.\left. + \left(\sum_{i=1}^{NT}\ \sum_{k=NT+1}^{NT+NM} w_i w_k\ \mathrm{Cov}(Y_i X_k)\right)\left(\mathrm{Cov}(Y_p X_q)\right) \right] \right.$$

$$\left. \Big/ \left(\sum_{k=NT+1}^{NT+NM}\ \sum_{\ell=NT+1}^{NT+NM} w_k w_\ell\ \mathrm{Cov}(X_k X_\ell)\right)^2 \right\},$$

$$1 \le p \le NT \quad \& \quad NT + 1 \le q \le NT+NC \tag{B-17}$$

$$\frac{\partial^2 N}{\partial w_q \partial w_p} = A\Bigg\{2\Bigg(\sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NT+NC} w_i w_k \mathrm{Cov}(Y_i Y_k)\Bigg)^2\Bigg[\Bigg(\mathrm{Cov}(X_p X_q)\Bigg)\Bigg(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \mathrm{Cov}(Y_i X_k)\Bigg)^2$$

$$+ 2\Bigg(\sum_{\ell=NT+1}^{NT+NM} w_\ell \mathrm{Cov}(X_p X_\ell)\Bigg)\Bigg(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \mathrm{Cov}(Y_i X_k)\Bigg)\Bigg(\sum_{i=1}^{NT} w_i \mathrm{Cov}(Y_i X_q)\Bigg)$$

$$-\Bigg(\sum_{i=1}^{NT} w_i \mathrm{Cov}(Y_i X_p)\Bigg)\Bigg\{\Bigg(\sum_{i=1}^{NT} w_i \mathrm{Cov}(Y_i X_q)\Bigg)\Bigg(\sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NT+NC} w_k W_\ell \mathrm{Cov}(X_k X_\ell)\Bigg)$$

$$+ 2\Bigg(\sum_{\ell=NT+1}^{NT+NC} w_\ell \mathrm{Cov}(X_q X_\ell)\Bigg)\Bigg(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \mathrm{Cov}(Y_i X_k)\Bigg)\Bigg\}\Bigg]$$

$$- 6\Bigg(\sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NT+NC} w_k w_\ell \mathrm{Cov}(X_k X_\ell)\Bigg)\Bigg(\sum_{\ell=NT+1}^{NT+NC} w_\ell \mathrm{Cov}(X_q X_\ell)\Bigg)$$

$$\Bigg[\Bigg(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \mathrm{Cov}(Y_i X_k)\Bigg)2\Bigg(\sum_{\ell=NT+1}^{NT+NC} w_\ell \mathrm{Cov}(X_p X_\ell)\Bigg)$$

$$-\Bigg(\sum_{i=1}^{NT} w_i \mathrm{Cov}(Y_i X_p)\Bigg)\Bigg(\sum_{i=1}^{NT}\sum_{k=NT+1}^{NT+NC} w_i w_k \mathrm{Cov}(Y_i X_k)\Bigg)\Bigg(\sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NT+NC} w_k w_\ell \mathrm{Cov}(X_k X_\ell)\Bigg)\Bigg\}$$

$$/\Bigg(\sum_{k=NT+1}^{NT+NC}\sum_{\ell=NT+1}^{NT+NC} w_k w_\ell \mathrm{Cov}(X_k X_\ell)\Bigg)^4 ,$$

$$NT+1 \leq p \leq NT+NC \quad \& \quad NT+1 \leq q \leq NT+NC \qquad \text{(B-18)}$$

APPENDIX C

PROBABILITY PLOTS FOR EXAMPLE APPLICATION

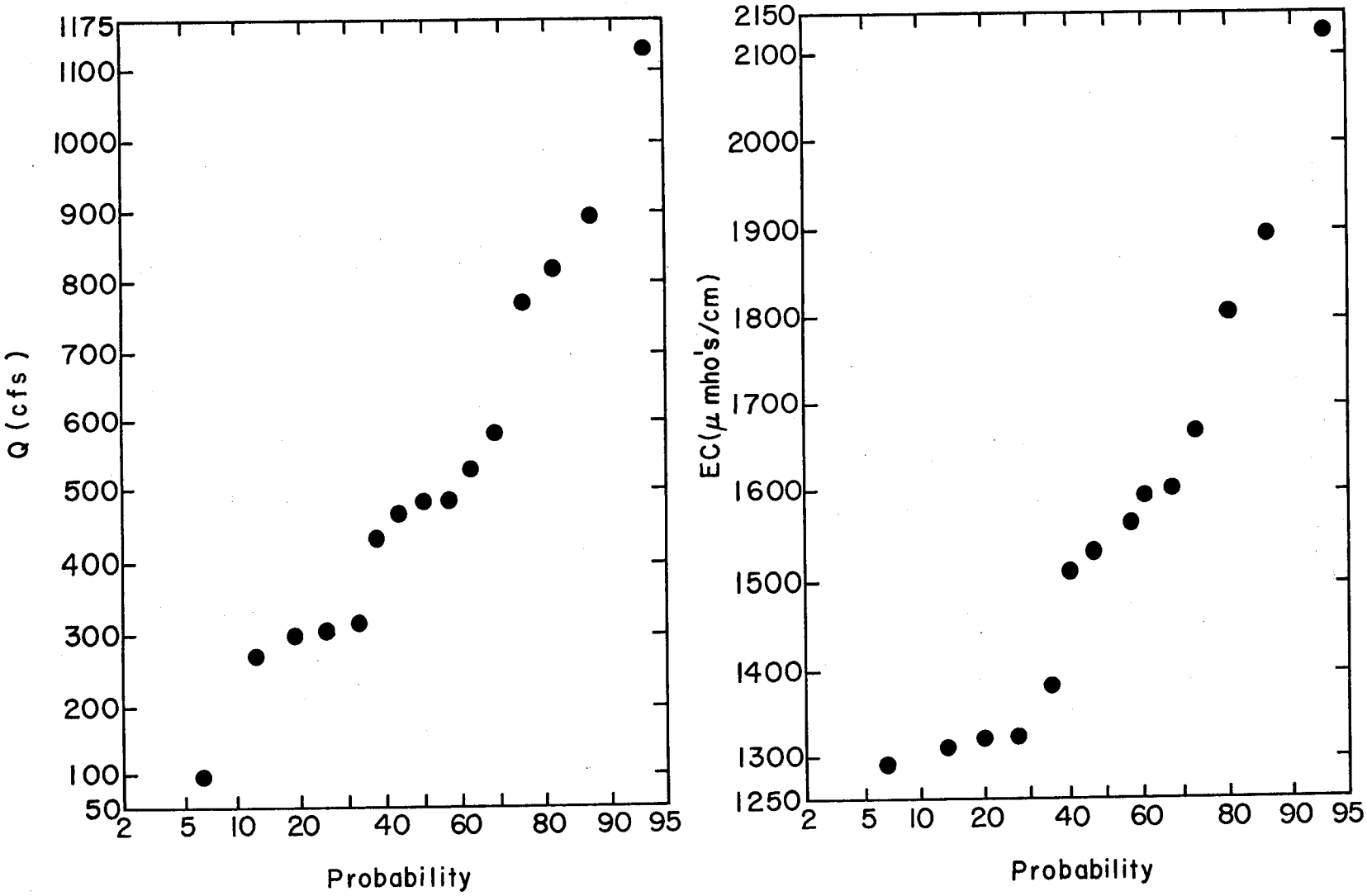


Figure C-1. Normal Probability Plots of Annual Flow and EC, Duchesne River.

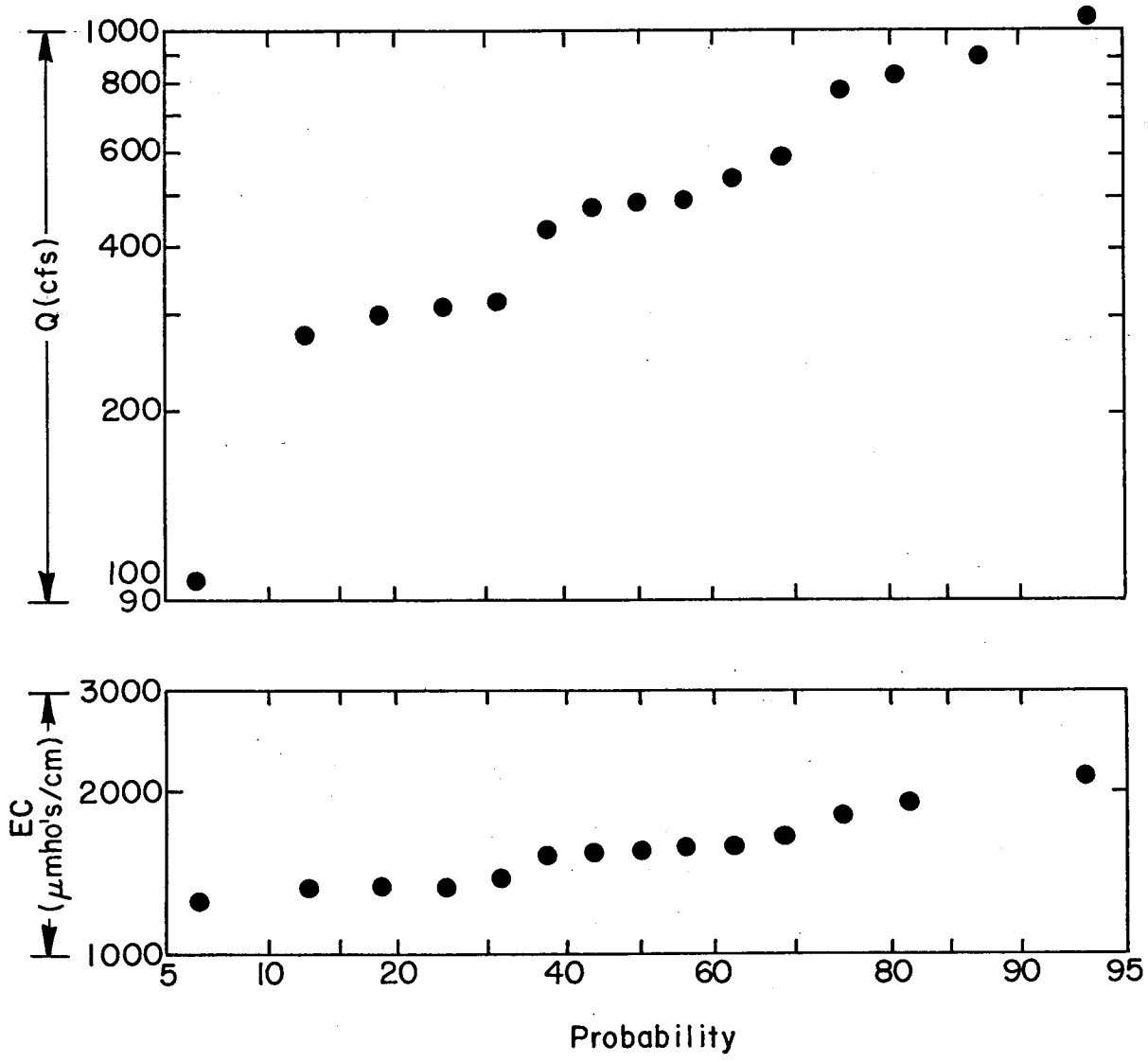


Figure C-2. Log-Normal Probability Plots of Annual Flow and EC, Duchesne River.

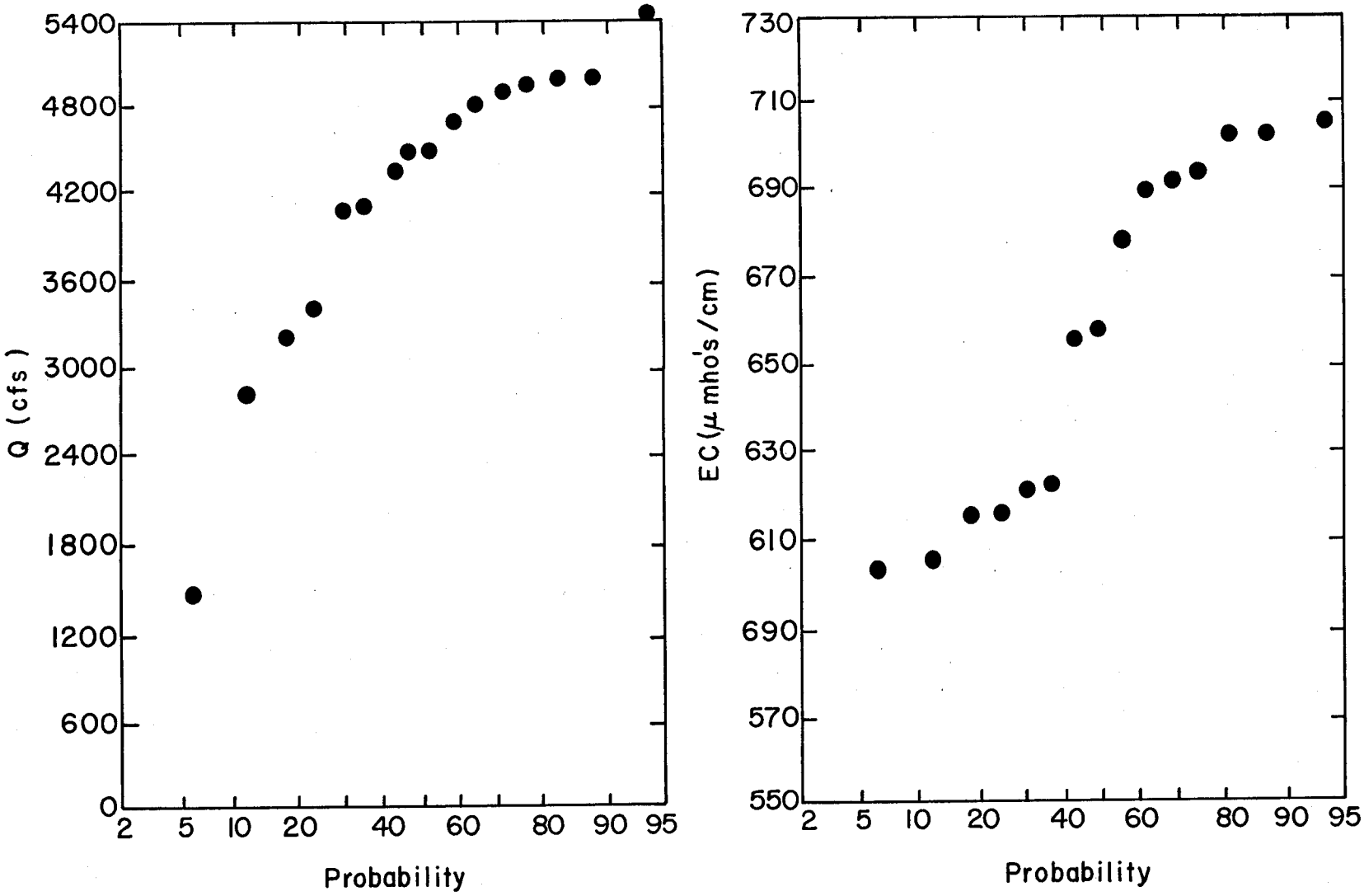


Figure C-3. Normal Probability Plots of Annual Flow and EC, Green River.

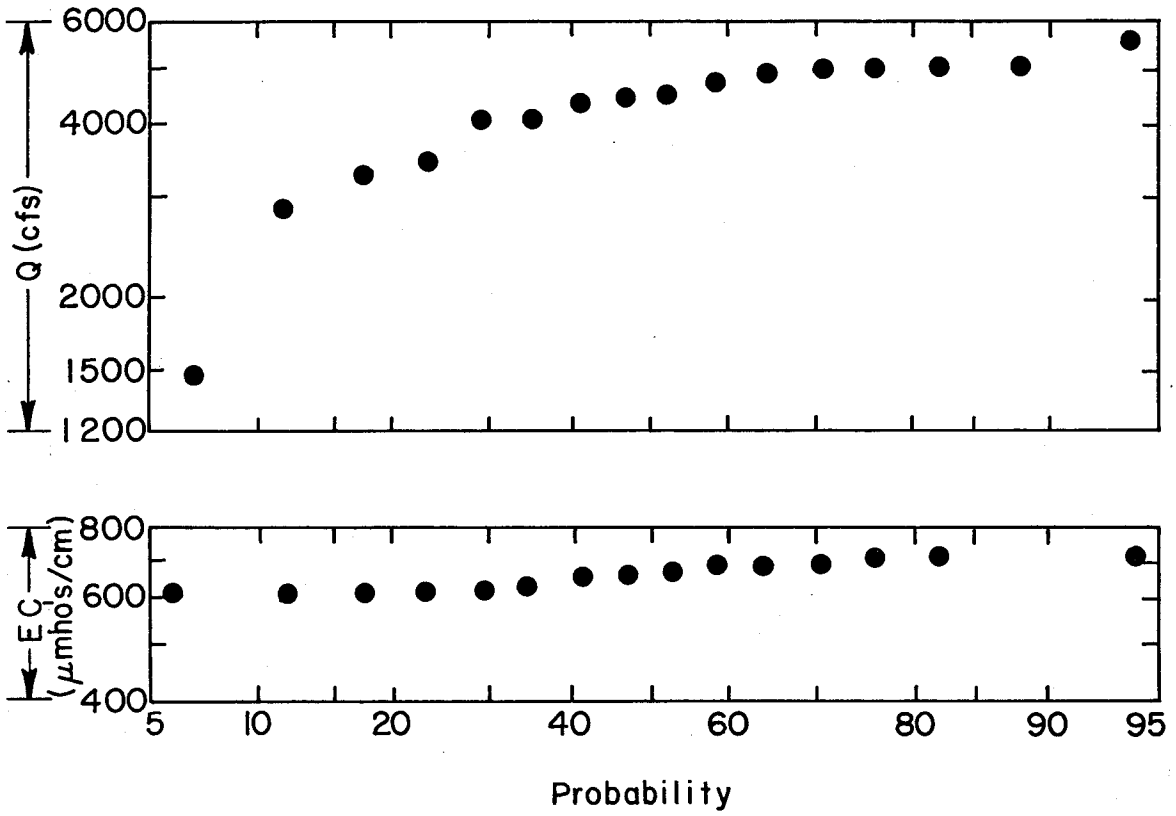


Figure C-4. Log-Normal Probability Plots of Annual Flow and EC, Green River.

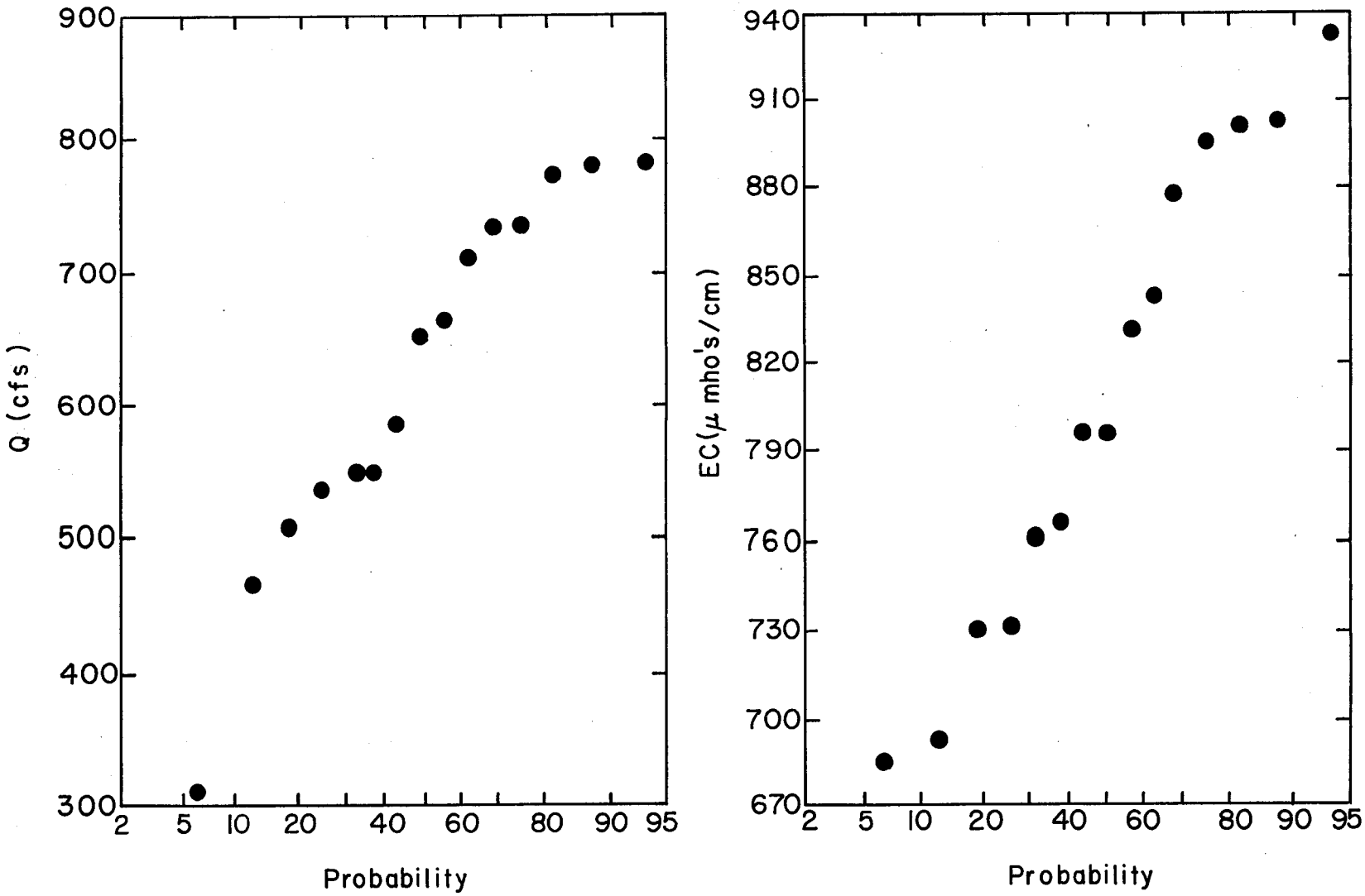


Figure C-5. Normal Probability Plots of Annual Flow and EC, White River.

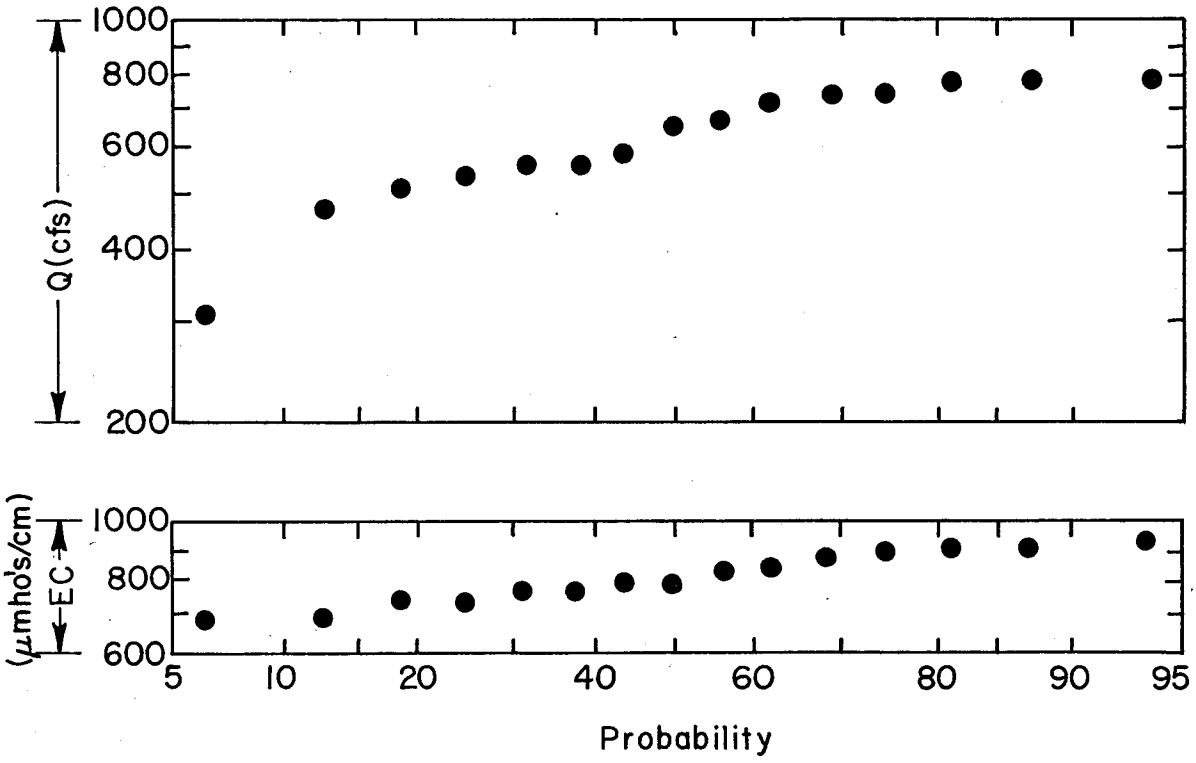


Figure C-6. Log-Normal Probability Plots of Annual Flow and EC, White River.

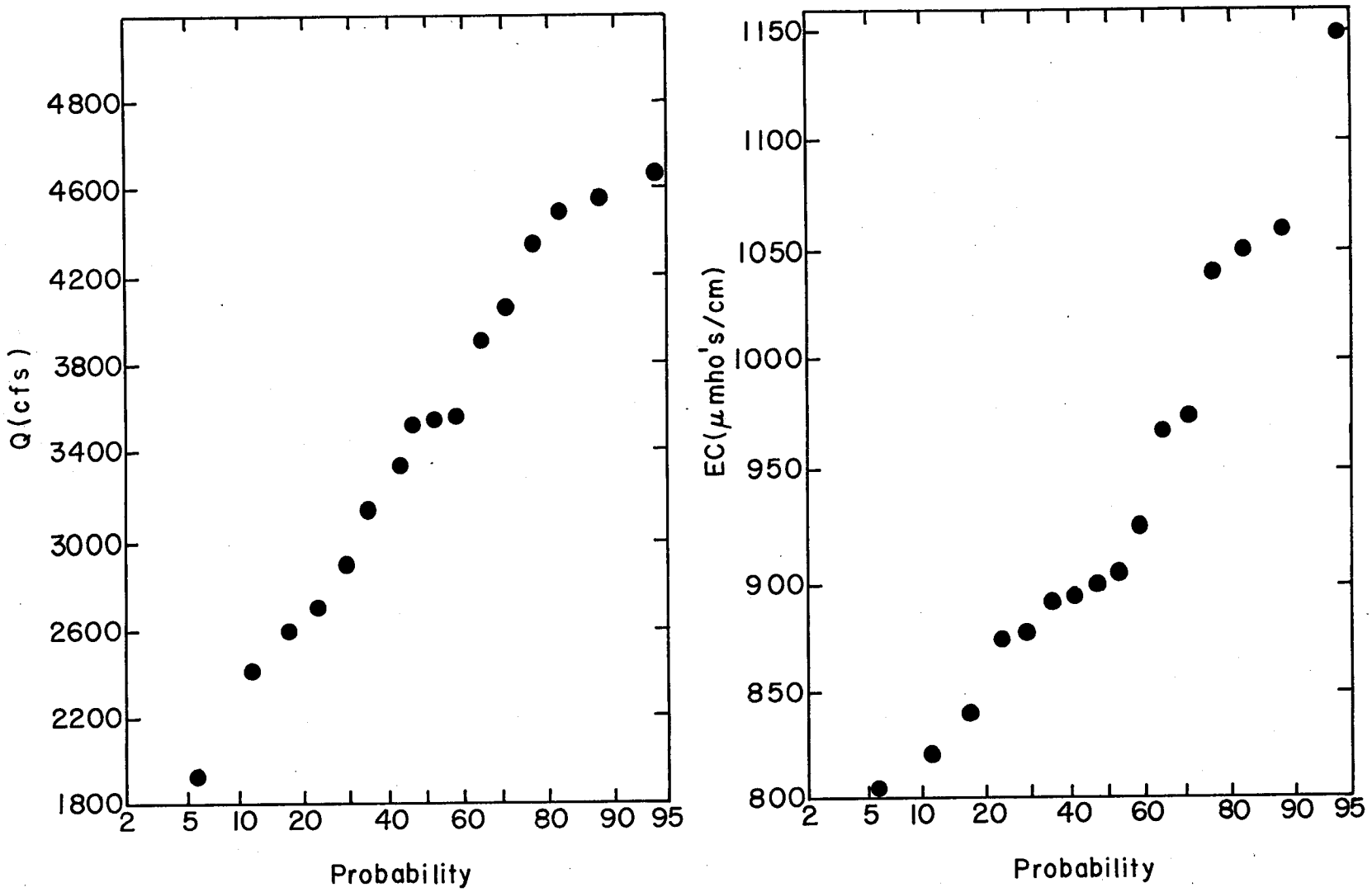


Figure C-7. Normal Probability Plots of Annual Flow and EC, Colorado River.

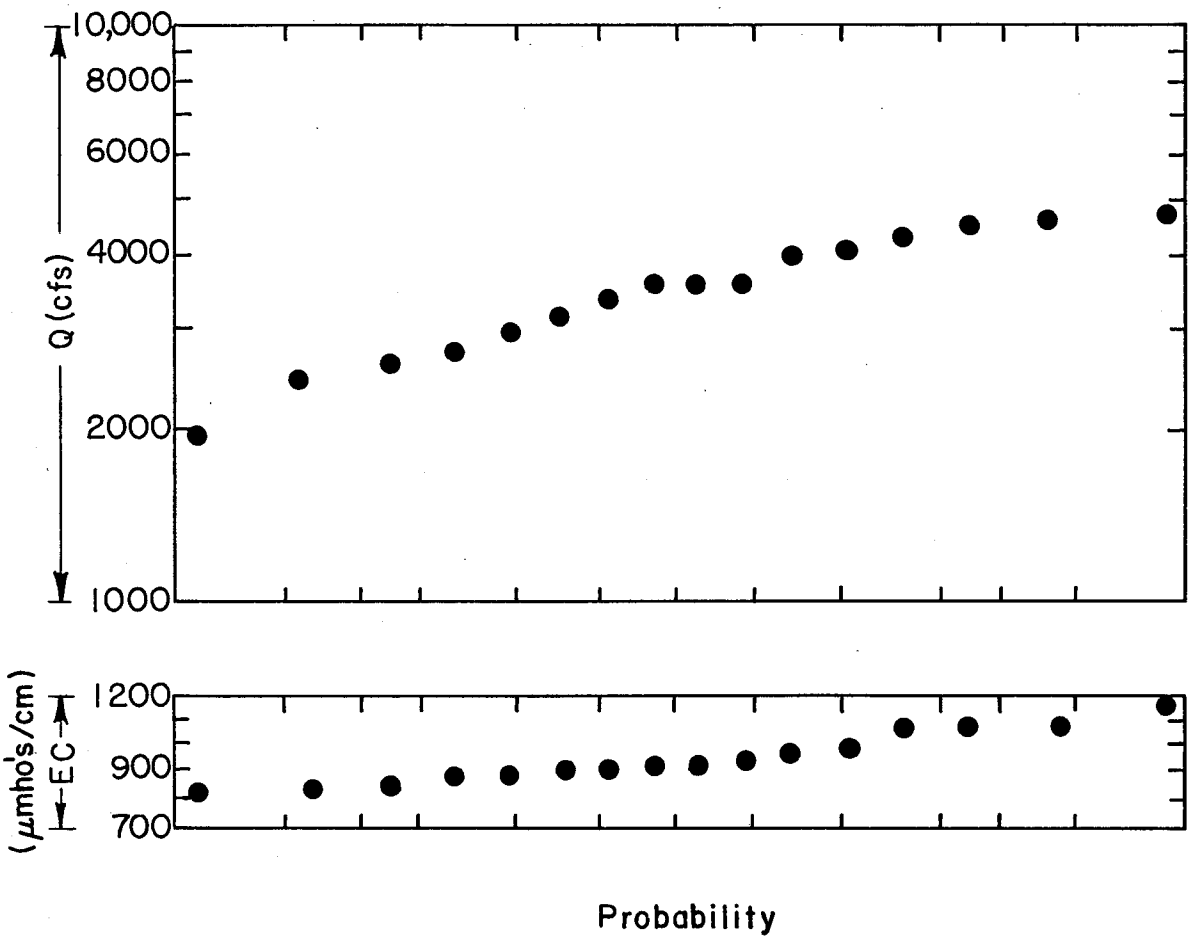


Figure C-8. Log-Normal Probability Plots of Annual Flow and EC, Colorado River.

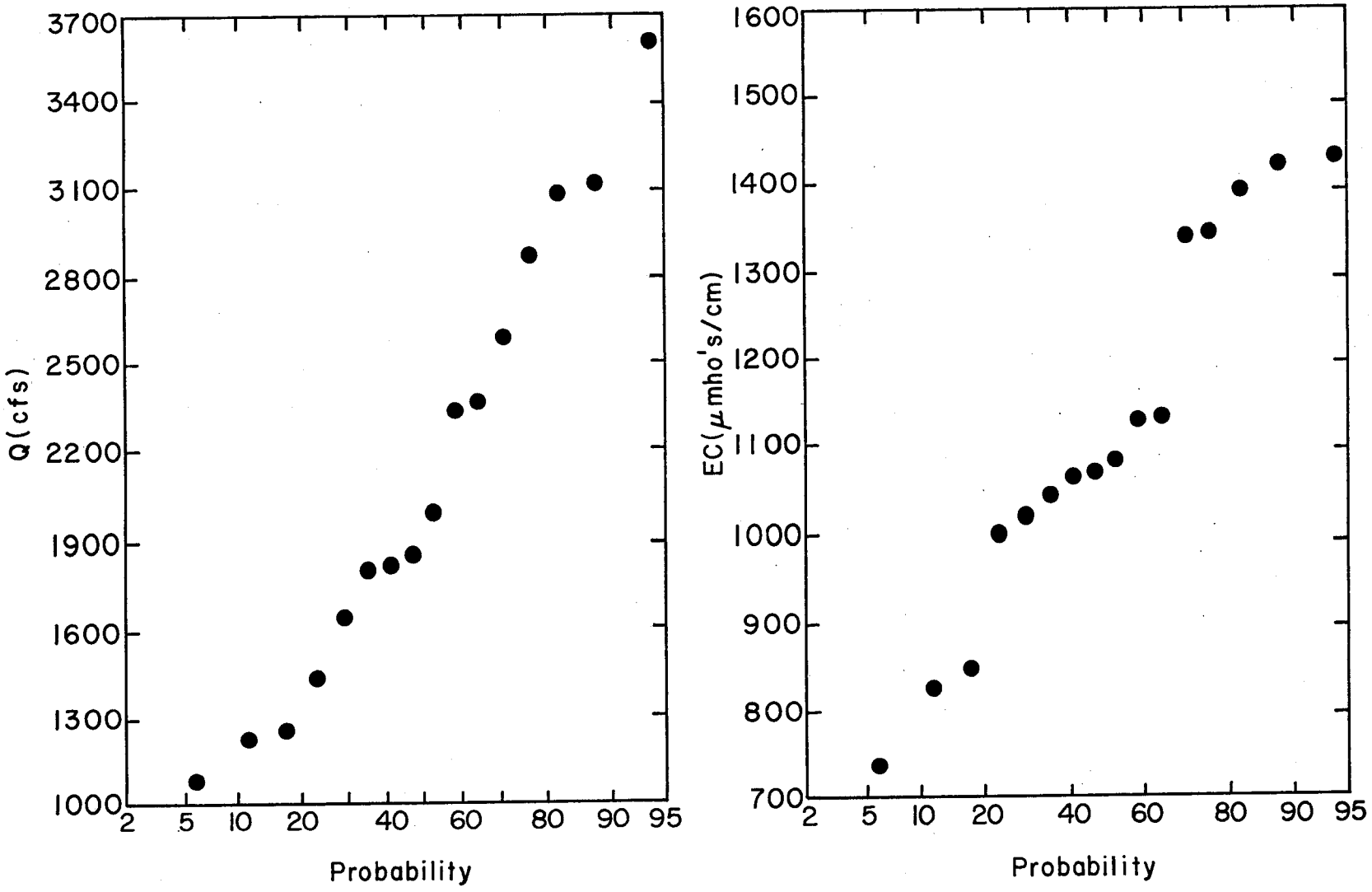


Figure C-9. Normal Probability Plots of Annual Flow and EC, Gunnison River.

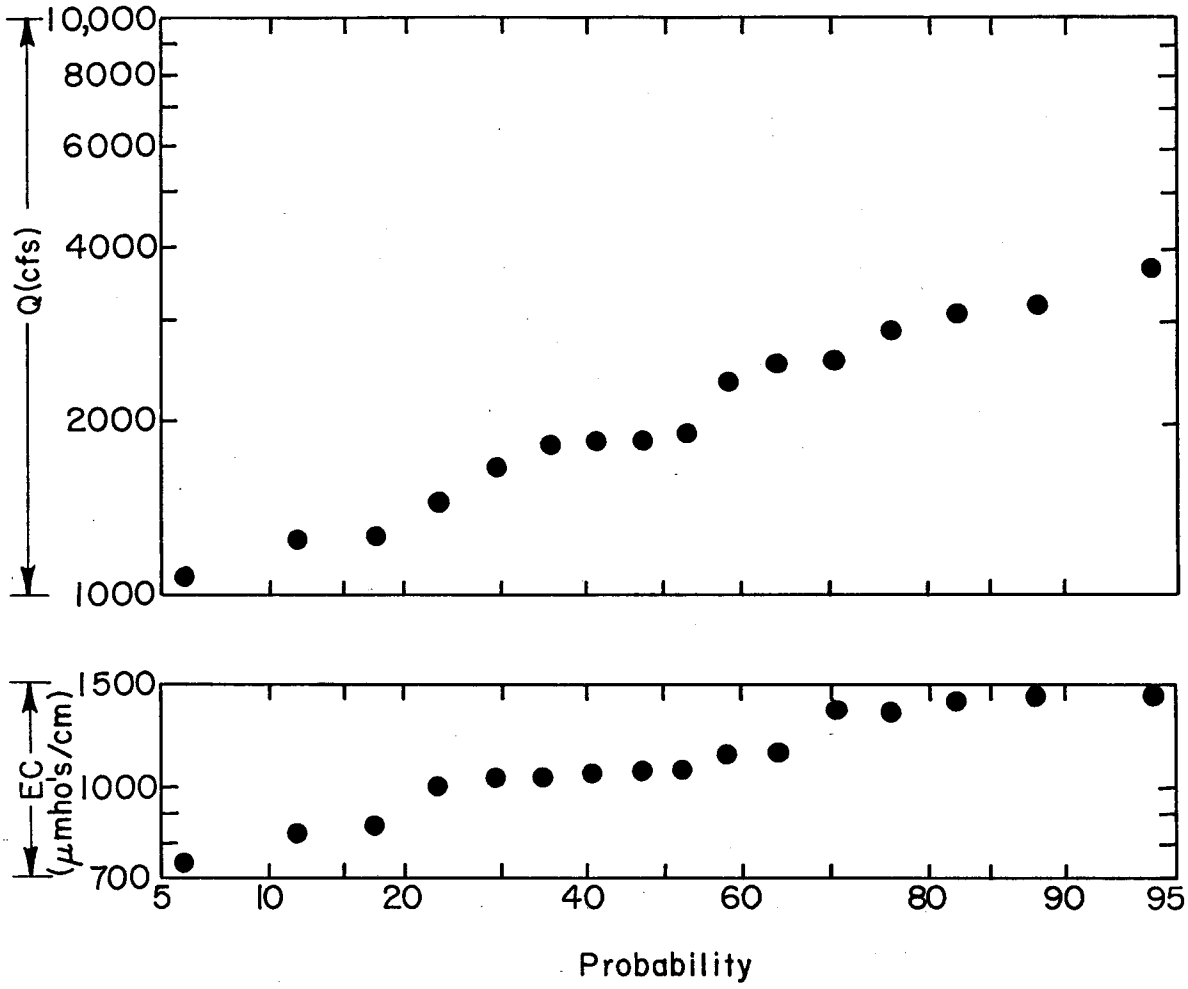


Figure C-10. Log-Normal Probability Plots of Annual Flow and EC, Gunnison River.

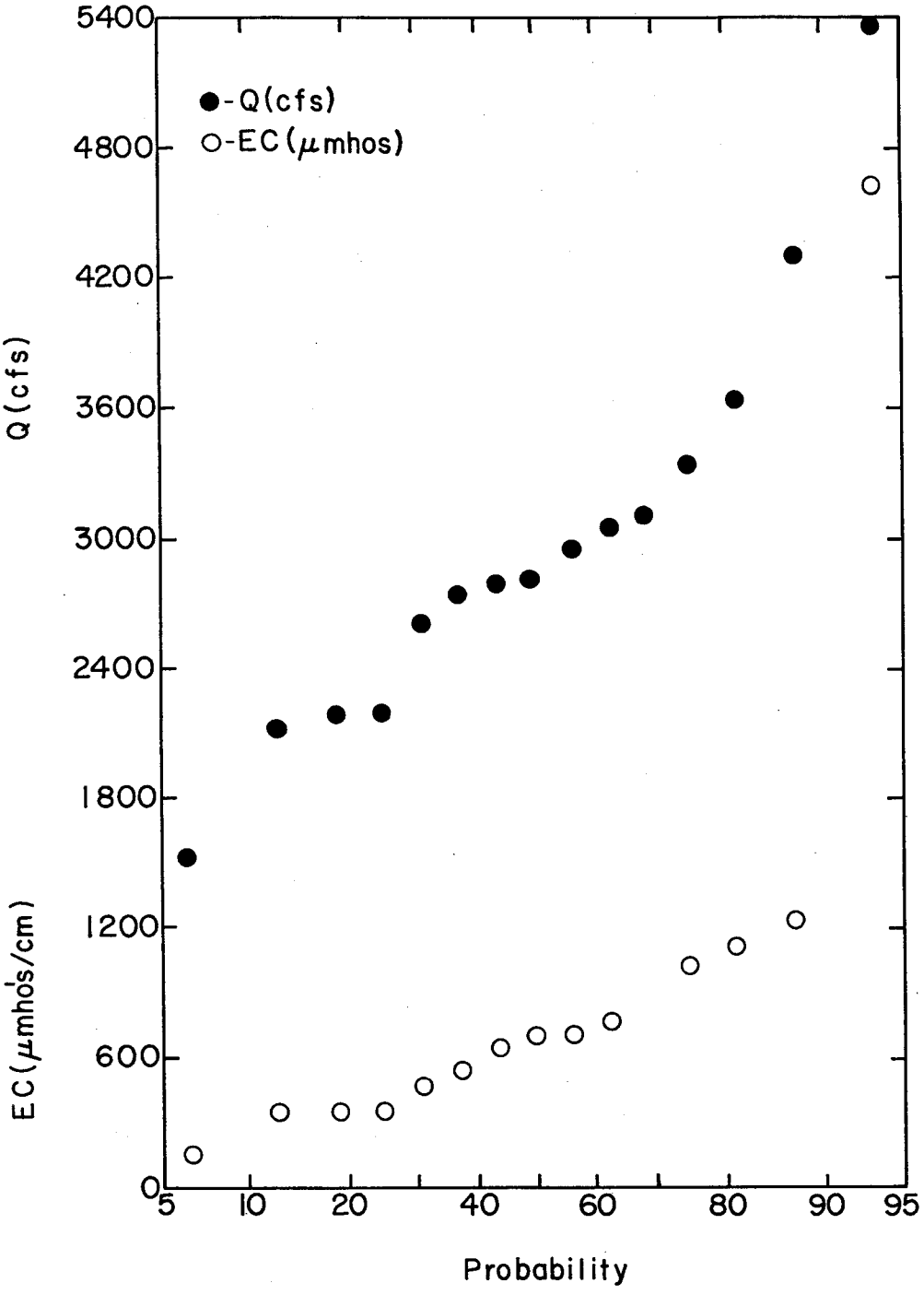


Figure C-11. Normal Probability Plots of Annual Flow and EC, Dolores River.

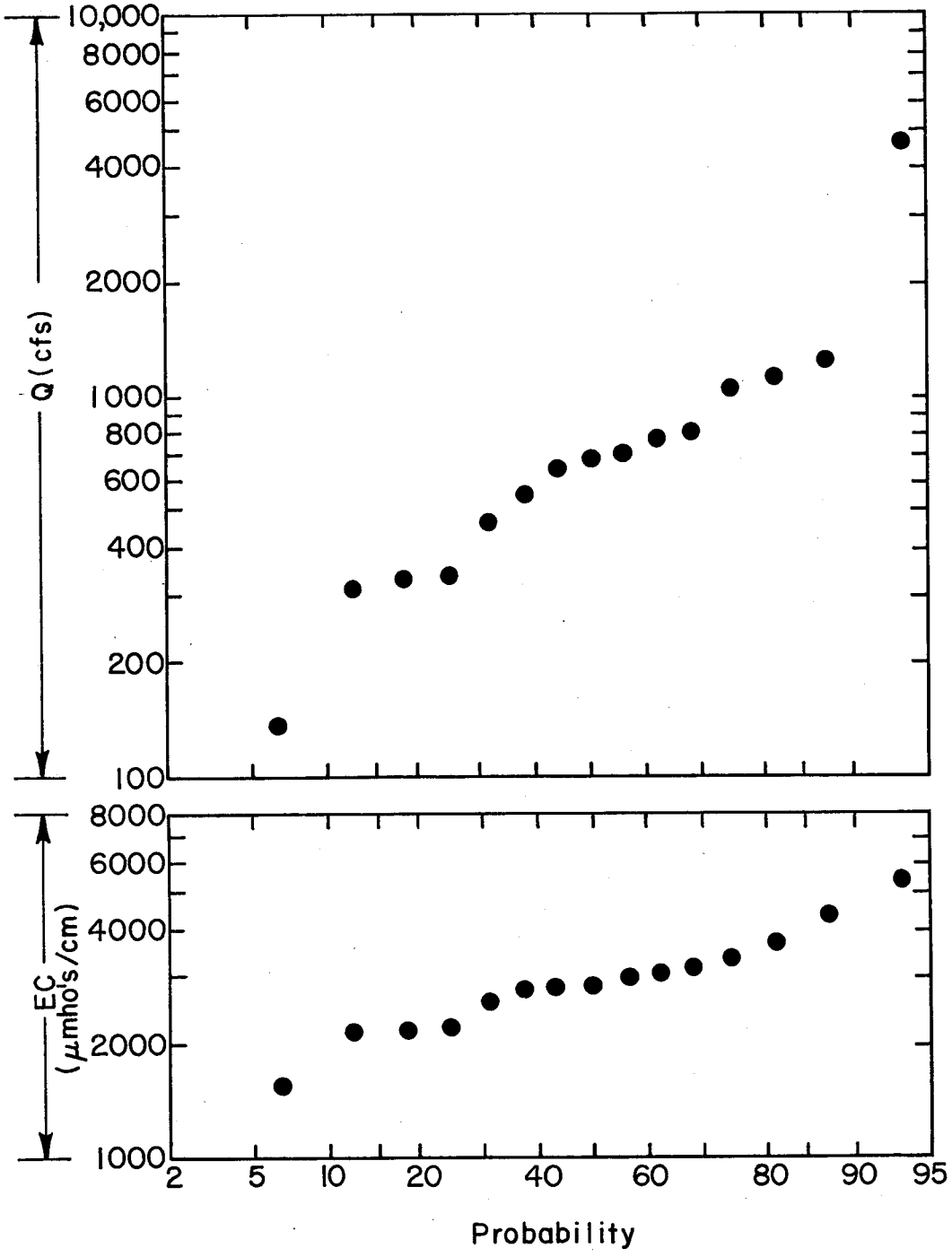


Figure C-12. Log-Normal Probability Plots of Annual Flow and EC, Dolores River.

APPENDIX D

UNDERLYING THEORY AND DERIVATIONS RELATED TO THE EVALUATION OF RELIABILITY

Stationarity Conditions

In any optimization problem, either constrained or unconstrained, location of a minimum (or maximum) is defined by the stationarity conditions. In the case of a univariate, unconstrained problem this is simply the point where the first derivative (or the slope of the curve) is zero. An analogous situation exists in the constrained case but, rather than the partial derivatives, the "constrained derivatives" are used to define the optimal solution (minimum or maximum). The constrained derivative (e.g., Beightler et al., 1979, or Morel-Seytoux, 1978) is defined as the partial derivative of the objective function while assuring that the constraints are always satisfied.

For a problem with N+K variables and K constraints, then K of the variables can be expressed in terms of the other N variable. These K variables $s_1, \ldots, s_k$ are termed state or solution variables while the remaining N variables $d_1, \ldots, d_n$ are called decision variables. At least in principle, then, these K state variables can be eliminated from the objective function and it can be expressed only in terms of the N decision variables. The constrained derivative, also called the decision derivative, is then defined as

$$\frac{\delta y}{\delta d_j} = \frac{\partial y}{\partial d_j} + \sum_{i=1}^{K} \frac{\partial y}{\partial s_i} \frac{\partial s_i}{\partial d_j} \quad , \quad j = 1, 2, \ldots, N \qquad \text{(D-1)}$$

where $\frac{\delta y}{\delta d_j}$ is the constrained derivative of the objective function, y, with respect to the decision variable, d_j;

$\frac{\partial y}{\partial d_j}$ is the usual partial derivative of the objective function with respect to the decision variable;

$\frac{\partial y}{\partial s_i}$ is the usual partial derivative of the objective function with respect to the state variable, s_i, and

$\frac{\partial s_i}{\partial d_j}$ is the partial derivative of the state variable s_i with respect to the decision variable, d_j.

The stationarity conditions are defined in terms of this constrained derivative.

The form of the stationarity conditions depends on whether the variables in the problem are confined to be greater than or equal to zero (non-negative variables) or whether they are allowed to assume any value (free variables). In the free variable case, the stationarity conditions are:

$$\frac{\delta y}{\delta d_j} = 0 \ , \quad j = 1, 2, \ldots, N \tag{D-2}$$

This can be seen to be completely analogous to the classical calculus situation. When non-negativity is imposed, however, the conditions become a bit more complex and are given by:

$$\frac{\delta y}{\delta d_j} \cdot d_j = 0 \qquad , \quad j = 1, 2, \ldots, N$$

and

$$\frac{\delta y}{\delta d_j} \geq 0 \ , \ d_j \geq 0 \ , \quad j = 1, 2, \ldots, N \tag{D-3}$$

Rather than merely requiring the constrained derivative to vanish at the stationary point, the product of the decision variable and constrained derivative must equal zero and both must also be non-negative.

The logic in these conditions can be outlined rather simply to provide an intuitive feeling for their meaning. If the constrained derivative were negative, an increase in the decision variable, d_j, would lead to a further decrease in the objective function so that the constrained derivative must be non-negative. Further, the decision

variable must be non-negative as required by the conditions originally imposed on the problem. If the constrained derivative has a zero value then either an increase or decrease in the variable, d_j, would cause an increase in the objective function (considering a minimization problem). So, if d_j is positive (not zero), $\delta y/\delta d_j$ must be zero. Finally, if the constrained derivative were positive and the variable, d_j, were positive, d_j could be decreased as far as zero which would decrease the objective function. Therefore, if the constrained derivative is positive, the decision variable must be at a zero value leading to the conditions that their product must be zero at the optimum.

These conditions, then, are necessary for a minimum to be achieved. It should be noted, however, that they are not sufficient for a global minimum. In addition to these conditions being fulfilled, the function must be positive definite for a minimum to always be the global minimum. As a final note, using the differential viewpoint to describe the stationarity conditions is not the only way this can be accomplished. The same results can be obtained by following through the generalized Kuhn-Tucker conditions using the classical Lagrange Multiplier technique (Beightler et al., 1978). However, it is felt that the approach presented above is a bit more intuitive by drawing an analogy to minimization as it is applied in classical calculus.

Solution for the Stationarity Conditions

To obtain the optimal solution in a minimization (or maximization) problem two approaches, the indirect and direct methods, are available. The indirect method, is based on the solution of a system of simultaneous equations to obtain the optimum while the direct method is based on successive improvement of the objective function through moving

progressively from one solution to a better one. It is the direct method that is the basis for mathematical programming as this is more efficient for large problems.

In the case when coefficients in the objective function and constraint equations are random in nature, however, the direct method may not provide the best method since the optimum will depend on the values assumed by these coefficients. To explore the variation of the "optimal" value of the objective function would require executing the mathematical programming problem a great number of times with different possible values for these coefficients. Another approach to the problem is to successively solve the stationarity conditions, (D-2) or (D-3) depending on the problem, for all possible combinations and finding those sets of conditions which are feasible in the context of the problem. This is a combinatorial problem which becomes larger as the number of variables in the problem increases. However, once the problem has been solved, explicit equations result relating the objective function and variables of the problem to the random coefficients. This technique will be demonstrated using the detection of change problem.

Direct Solution of Stationarity Conditions

For the general problem posed in the form of minimizing the conditional variance of a linear combination of stations, the objective function is highly nonlinear as are the derivatives presented in Appendix B. Thus, the constrained derivative, Equation (D-1), is also highly nonlinear making this a very difficult, possibly an impossible problem to solve analytically. Due to this and the fact that many engineering problems are posed as a simpler linear or quadratic objective function, a less complex problem is developed to demonstrate the

technique. The new problem is to minimize the number of years required to detect a change of 100k percent in a linear combination of stations, $Y^* = \sum_{i=1}^{N} w_i Y_i$, with a specified power, $1-\beta$, and size, α, of test. A constraint that the mean of the weighted linear combination must equal the mean of the unweighted sum of the stations is also imposed. This problem is written mathematically as:

$$\underset{\underline{w}}{\text{Min}} \left\{ N = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{k \bar{Y}^*} \right)^2 \sum_{i=1}^{N} \sum_{j=1}^{N} w_i w_j S_{Y_i Y_j} \right\}$$

subject to:

$$\sum_{i=1}^{N} w_i \bar{Y}_i = \sum_{i=1}^{N} \bar{Y}_i = \bar{Y}^* \qquad \text{(D-4)}$$

Non-negativity of the weights could also be imposed as a further condition.

This is a quadratic programming (QP) problem with one, linear equality constraint. The solution of this problem by direct application of the stationarity conditions will depend on whether non-negativity conditions are imposed. Both situations will be investigated with only two stations (N=2) being used for illustration.

Free Variable Case. In this situation, the only stationarity condition is:

$$\frac{\delta y}{\delta d_j} = 0 \quad , \quad j = 1, 2, \ldots, N$$

where N is the number of decision variables; the total number of variables in the problem minus the number of constraints (1 in this case).

From (D-1) the constrained derivative of (D-4) selecting w_1 as the decision variable is:

$$\frac{\delta N}{\delta w_1} = \frac{\partial N}{\partial w_1} + \frac{\partial N}{\partial w_2}\frac{\partial w_2}{\partial w_1} \tag{D-5}$$

In this specific case, where w_2 is expressed in terms of w_1 through the constraint:

$$\frac{\partial N}{\partial w_1} = 2(S_{Y_1}^2 w_1 + S_{Y_1Y_2} w_2)\left(\frac{z_{1-\beta} + z_{1-\alpha}}{k\bar{Y}^*}\right)^2 \tag{D-6}$$

$$\frac{\partial N}{\partial w_2} = 2(S_{Y_1Y_2} w_1 + S_{Y_2}^2 w_2)\left(\frac{z_{1-\beta} + z_{1-\alpha}}{k\bar{Y}^*}\right)^2 \tag{D-7}$$

$$\frac{\partial w_2}{\partial w_1} = -\frac{\bar{Y}_1}{\bar{Y}_2} \tag{D-8}$$

So, the constrained derivative is

$$\frac{\delta N}{\delta w_1} = 2\left(\frac{z_{1-\beta} + z_{1-\alpha}}{kY^*}\right)^2\left[(S_{Y_1}^2 w_1 + S_{Y_1Y_2} w_2) - \frac{\bar{Y}_1}{\bar{Y}_2}(S_{Y_1Y_2} w_1 + S_{Y_2}^2 w_2)\right] \tag{D-9}$$

Now, applying the stationarity condition, D-1 and gathering coefficients of w_1 and w_2 gives

$$\left(S_{Y_1}^2 - \frac{\bar{Y}_1}{\bar{Y}_2} S_{Y_1Y_2}\right) w_1 + \left(S_{Y_1Y_2} - \frac{\bar{Y}_1}{\bar{Y}_2} S_{Y_2}^2\right) w_2 = 0 \tag{D-10}$$

Along with the constraint equation there are now two linear equations and two unknowns, w_1 and w_2. Simultaneous solution of these two equations gives the following results:

$$w_1 = \frac{\bar{Y}^*\left(\frac{S_{Y_2}^2}{\bar{Y}_2} - \frac{S_{Y_1Y_2}}{\bar{Y}_1}\right)}{\left(S_{Y_1}^2\frac{\bar{Y}_2}{\bar{Y}_1} - 2\,S_{Y_1Y_2} + S_{Y_2}^2\frac{\bar{Y}_1}{\bar{Y}_2}\right)} \tag{D-11}$$

$$w_2 = \frac{\bar{Y}^* \left(\frac{S_{Y_1}^2}{\bar{Y}_1} - \frac{S_{Y_1Y_2}}{\bar{Y}_2} \right)}{\left(S_{Y_1}^2 \frac{\bar{Y}_2}{\bar{Y}_1} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \frac{\bar{Y}_1}{\bar{Y}_2} \right)} \qquad \text{(D-12)}$$

The optimal value of the objective function, N^*, can then be obtained by substituting w_1 and w_2 to give:

$$N^* = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{kY^*} \right)^2 \left\{ S_{Y_1}^2 \left(\frac{S_{Y_1}^2}{Y_2} - \frac{S_{Y_1Y_2}}{Y_1} \right)^2 + 2\, S_{Y_1Y_2} \left(\frac{S_{Y_2}^2}{Y_2} - \frac{S_{Y_1Y_2}}{Y_1} \right) \left(\frac{S_{Y_1}^2}{Y_1} - \frac{S_{Y_1Y_2}}{Y_2} \right) \right.$$

$$\left. + S_{Y_2}^2 \left(\frac{S_{Y_1}^2}{\bar{Y}_1} - \frac{S_{Y_1Y_2}}{\bar{Y}_2} \right)^2 \right\} \Big/ \left(S_Y^2 \frac{\bar{Y}_1}{\bar{Y}_2} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \frac{\bar{Y}_2}{\bar{Y}_1} \right)^2 \qquad \text{(D-13)}$$

Equations D-11, D-12 and D-13 express the optimal values of the variables (weighting factors in this case) and objective function as explicit functions of the statistics of the variables being investigated in the detection of change problem. The same expressions are obtained if w_2 is chosen as the decision variable. So, although the expressions are complex, they show how uncertainty in the estimation of the statistical parameters affects the results of the optimization.

Non-negative Variable Case. In this situation, the stationarity conditions are given by (D-3). Thus, there are 4 possible combinations which could result in this case. These are:

A) w_1 is the decision variable

1) $\frac{\delta N}{\delta w_1} \geq 0, \quad w_1 = 0$

2) $\frac{\delta y}{\delta w_1} = 0, \quad w_1 \geq 0$

B) w_2 is the decision variable

1) $\frac{\delta N}{\delta w_2} \geq 0, \quad w_2 = 0$

2) $\frac{\delta N}{\delta w_2} = 0, \quad w_2 \geq 0$

From the previous case the selection of the decision variable gave no difference in the value of w_1 and w_2, thus it can be seen that case A2 and B2 will produce results identical to the free variable case as long as w_1 and w_2 are non-negative. The non-negativity conditions require that both w_1 and w_2 in Equations (D-11) and (D-12) be non-negative. Inspection of these equations shows that the denominator is, if the population parameters are used, the variance of the weighted difference between Y_1 and Y_2. That is

$$\text{Var}\,[a\,Y_1 - b\,Y_2] = a^2\sigma_{Y_1}^2 - ab\,\sigma_{Y_1Y_2} + b^2\,\sigma_{Y_2}^2 \qquad \text{(D-14)}$$

Now, let $a = \sqrt{\mu_{Y_2}/\mu_{Y_1}}$ and $b = \sqrt{\mu_{Y_1}/\mu_{Y_2}}$ then

$$\text{Var}\left[\sqrt{\frac{\mu_{Y_2}}{\mu_{Y_1}}}\,Y_1 - \sqrt{\frac{\mu_{Y_1}}{\mu_{Y_2}}}\,Y_2\right] = \frac{\mu_{Y_2}}{\mu_{Y_1}}\,\sigma_{Y_1}^2 - 2\,\sigma_{Y_1Y_2} + \frac{\mu_{Y_1}}{\mu_{Y_2}}\,\sigma_{Y_2}^2 \qquad \text{(D-15)}$$

Replacing the population values by the sample values given the denominator in (D-11) and (D-12). And, since variances are always positive, only the non-negativity of the numerator need be considered. Thus, the non-negativity conditions are given by

$$\bar{Y}^*\left(\frac{S_{Y_2}^2}{\bar{Y}_2} - \frac{S_{Y_1Y_2}}{\bar{Y}_1}\right) \geq 0 \quad \text{for} \quad w_1 \geq 0 \qquad \text{(D-16)}$$

$$\bar{Y}^{*}\left(\frac{S_{Y_2}^2}{\bar{Y}_1} - \frac{S_{Y_1Y_2}}{\bar{Y}_2}\right) \geq 0 \quad \text{for} \quad w_2 \geq 0 \tag{D-17}$$

These conditions can be written as

$$\frac{\bar{Y}_1}{\bar{Y}_2} > \frac{S_{Y_1Y_2}}{S_{Y_2}^2} \quad \text{for} \quad w_1 > 0 \tag{D-18}$$

$$\frac{\bar{Y}_1}{\bar{Y}_2} < \frac{S_{Y_1}^2}{S_{Y_1Y_2}} \quad \text{for} \quad w_2 > 0 \tag{D-19}$$

It remains to investigate cases A1 and B1 to complete the evaluation of the non-negative case. In A1, w_1 equals zero and the following equations result from applying the stationarity conditions

$$w_1 = 0 \tag{D-20}$$

$$w_2 = \bar{Y}^{*}/\bar{Y}_2 \tag{D-21}$$

$$N^{*} = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{k}\right)^2 \left(\frac{S_{Y_2}}{\bar{Y}_2}\right)^2 \tag{D-22}$$

Similar results are obtained from B1 as

$$w_1 = \bar{Y}^{*}/\bar{Y}_1 \tag{D-23}$$

$$w_2 = 0 \tag{D-24}$$

$$N^{*} = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{k}\right)^2 \left(\frac{S_{Y_1}}{\bar{Y}_1}\right)^2 \tag{D-25}$$

To summarize the non-negative variable case, three different results are possible depending on the values of the statistics. These are

I) if $\dfrac{\bar{Y}_1}{\bar{Y}_2} \leq \dfrac{S_{Y_1Y_2}}{S_{Y_2}^2}$, then

$$w_1 = 0$$

$$w_2 = \bar{Y}^*/\bar{Y}_2$$

$$N^* = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{k}\right)^2 \left(\frac{S_{Y_2}}{\bar{Y}_2}\right)^2$$

II) if $\dfrac{\bar{Y}_1}{\bar{Y}_2} \geq \dfrac{S_{Y_1}^2}{S_{Y_1Y_2}}$, then

$$w_1 = \bar{Y}^*/\bar{Y}_1$$

$$w_2 = 0$$

$$N^* = \left(\frac{z_{1-\beta} + z_{1-\alpha}}{k}\right)^2 \left(\frac{S_{Y_1}}{\bar{Y}_1}\right)^2$$

III) if $\dfrac{S_{Y_1}^2}{S_{Y_1Y_2}} \geq \dfrac{\bar{Y}_1}{\bar{Y}_2} \leq \dfrac{S_{Y_1Y_2}}{S_{Y_2}^2}$, then

$$w_1 = \frac{\bar{Y}^* \left(\dfrac{S_{Y_1}^2}{\bar{Y}_1} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_2}\right)}{\left(S_{Y_1}^2 \dfrac{\bar{Y}_2}{\bar{Y}_1} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \dfrac{\bar{Y}_1}{\bar{Y}_2}\right)}$$

$$w_2 = \frac{\bar{Y}^* \left(\dfrac{S_{Y_2}^2}{\bar{Y}_2} - \dfrac{S_{Y_1Y_2}}{\bar{Y}_1}\right)}{\left(S_{Y_1}^2 \dfrac{\bar{Y}_1}{\bar{Y}_2} - 2\, S_{Y_1Y_2} + S_{Y_2}^2 \dfrac{\bar{Y}_1}{\bar{Y}_2}\right)}$$

$$N^{*} = \left(\frac{z_{1-\beta}+z_{1-\alpha}}{kY^{*}}\right)^{2}\left\{S_{Y_1}^{2}\left(\frac{S_{Y_2}^{2}}{\bar{Y}_2} - \frac{S_{Y_1Y_2}}{\bar{Y}_1}\right)^{2} + 2\, S_{Y_1Y_2}\left(\frac{S_{Y_2}^{2}}{\bar{Y}_2} - \frac{S_{Y_1Y_2}}{\bar{Y}_1}\right)\left(\frac{S_{Y_1}^{2}}{\bar{Y}_1} - \frac{S_{Y_1Y_2}}{\bar{Y}_2}\right)\right.$$

$$\left. + S_{Y_2}^{2}\left(\frac{S_{Y_1}^{2}}{\bar{Y}_1} - \frac{S_{Y_1Y_2}}{\bar{Y}_2}\right)^{2}\right\} \Big/ \left(S_{Y}^{2}\frac{\bar{Y}_1}{\bar{Y}_2} - 2\, S_{Y_1Y_2} + S_{Y_2}^{2}\frac{\bar{Y}_2}{\bar{Y}_1}\right)^{2}$$

Thus, although in a more complicated manner, the variables and objective function are again expressed explicitly as functions of the statistics (random variables) in the problem allowing direct evaluation of their affect. In this case, the additional feature is the regions in which various solutions apply. These are also expressed in terms of the statistics of the problem.

A final observations can be made in relation to the non-negativity criteria for determining the domain of each solution, Equations (D-18) and (D-19). These can further be rearranged to give the following:

$$\bar{Y}_1 - \bar{Y}_2\, r_{Y_1Y_2}\frac{S_{Y_1}}{S_{Y_2}} > 0 \qquad \text{(D-26)}$$

$$\bar{Y}_2 - \bar{Y}_1\, r_{Y_1Y_2}\frac{S_{Y_2}}{S_{Y_1}} > 0 \qquad \text{(D-27)}$$

where $r_{Y_1Y_2}$ is the sample correlation coefficient between Y_1 and Y_2.

These equations could be viewed as sample estimates of the regression constants in the linear regression between the two variables (Mood et al., 1974). For example, for the relation

$$Y_1 = a_1 + b_1\, Y_2 \qquad \text{(D-28)}$$

The constant, a_1, can be estimated by the left side of (D-26). Similarly for the equation

$$Y_2 = a_2 + b_2 Y_1 \qquad \text{(D-29)}$$

then, a_2, can be estimated by the left side of (D-27). Thus, if the intercept of the regression line is above the origin, the variables are greater than zero. Otherwise, they are assigned a zero value.

APPENDIX E

ESTIMATION OF SAMPLE SIZES REQUIRED FOR DETERMINATION OF SIZE AND POWER

Since data generation is an expensive proposition requiring much computer time, it is necessary to estimate the amount of data required for the purposes of the project. In this case, the interest is in estimating the actual size and power of the statistical test since all of the computations are based on sample estimates rather than population parameters. Viewing the size of the test as estimated from each simulation, as the outcome of an independent, Bernoulli trial, (i.e., rejection is a success) then the estimate of the size, α, or the parameter, p, in the Bernoulli distribution (Mood et al., 1974) is

$$E[\alpha] = p \qquad \text{(E-1)}$$

In addition, the variance of the variable is given as:

$$\text{Var}[\alpha] = pq \qquad \text{(E-2)}$$

where $q = 1 - p$

Now, the expression for the variance of the sample mean of a random sample is known to be (Mood et al., 1974)

$$\text{Var}[\bar{x}] = \frac{\text{var}[x]}{n} \qquad \text{(E-3)}$$

In this case, the variance of the sample estimate of α is:

$$\text{Var}[\bar{\alpha}] = \frac{pq}{n} \qquad \text{(E-4)}$$

Of interest in this analysis, is the ability to say, with some relatively high probability, that the estimate of α is, in fact, equal to the originally prescribed value within some small interval. Mathematically this is:

$$P_r[\alpha-0.5 < \varepsilon] = 0.95 \qquad \text{(E-5)}$$

where ε is some small quantity

This can also be written as

$$P_r[0.05-\varepsilon < \alpha < 0.05 + \varepsilon] = 0.95 \qquad \text{(E-6)}$$

Now, if it is assumed that this distribution can be approximated by the normal distribution since the parameter of interest is a mean, then (E-6) can be written as

$$\Phi\left(\frac{\varepsilon}{\sqrt{\frac{pq}{n}}}\right) - \Phi\left(\frac{\varepsilon}{\sqrt{\frac{pq}{n}}}\right) = 0.95 \qquad \text{(E-7)}$$

where $\Phi(\cdot)$ is the cumulative normal distribution

p is the success probability

q is the failute probability

$\sqrt{\frac{pq}{n}}$ is the standard deviation of the sample estimate of the parameter.

Setting $p = 0.05$, $q = 0.95$ and realizing that the value of the standard normal deviate for a two tailed, having a 95 percent confidence level is 1.96, the following equation results

$$\frac{\varepsilon}{\sqrt{\frac{0.0475}{n}}} = 1.96 \qquad \text{(E-8)}$$

Then, to solve for the sample size, n, the increment, ε, must be chosen. Table E-1 shows sample sizes for various values of ε. For this study, a sample size of 500 was selected which should provide a 95 percent assurance of the evaluation that the size is 0.05 ± 0.02.

A similar analysis can be done for the power. Using E-7 with $p = q = 0.50$, the following expression is obtained relating sample size and error

$$\frac{\varepsilon}{\sqrt{\frac{0.25}{n}}} = 1.96 \qquad \text{(E-9)}$$

Using this equation, Table E-2 was constructed. A sample size of 100 was selected which allows evaluation of the power as 0.50 ± 0.10 with 95 percent assurance.

Table E-1

Sample Size, n, Required to Estimate Whether the Size of the Test, α, is within the Stated Increment, ε

$\alpha = 0.05$

Sample Size n	Error ε
7,299	0.005
1,824	0.010
456	0.020
73	0.050

Table E-2

Sample Size, n, Required to Estimate Whether the Power of the Test, $1-\beta$, is within the Stated Increment, ε

$1-\beta = 0.50$

Sample Size n	Error ε
384	0.05
96	0.10
24	0.20