US Army Corps of Engineers

ENVIRONMENTAL & WATER QUALITY OPERATIONAL STUDIES

INSTRUCTION REPORT E-87-1

SAMPLING DESIGN FOR RESERVOIR WATER QUALITY INVESTIGATIONS

by

Robert F. Gaugush

Environmental Laboratory

DEPARTMENT OF THE ARMY
Waterways Experiment Station, Corps of Engineers
PO Box 631, Vicksburg, Mississippi 39180-0631

March 1987

Final Report

Approved For Public Release; Distribution Unlimited

Prepared for DEPARTMENT OF THE ARMY
US Army Corps of Engineers
Washington, DC 20314-1000

Under EWQOS Work Unit VIIA

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188
Exp. Date: Jun 30, 1986

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Instruction Report E-87-1	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION USAEWES Environmental Laboratory 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) PO Box 631 Vicksburg, MS 39180-0631	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION US Army Corps of Engineers 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code) Washington, DC 20314-1000	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO EWQOS VIIA

11. TITLE (Include Security Classification)

Sampling Design for Reservoir Water Quality Investigations

12. PERSONAL AUTHOR(S)
Gaugush, Robert F.

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final report	FROM _____ TO _____	March 1987	90

16. SUPPLEMENTARY NOTATION
Available from National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161.

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Reservoir, Statistics, Sampling, Water Quality, Sample Design

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This report introduces the statistical concepts involved in the development of sampling designs for reservoir water quality investigations. It provides a general overview of statistical inference and the assumptions involved in sampling. A discussion of basic statistics is included to familiarize the reader with the necessary statistical background for the material to follow. The presentation of the methods of sampling design is made with specific reference to reservoir water quality sampling and covers the determination of sample size, simple random sampling, stratified random sampling, and systematic sampling. The report also presents the use of decision matrices, tools to aid in the decisionmaking process involved in determining the number of samples, the parameters to be measured, and the frequency of sampling within funding constraints. Reservoir water quality patterns and their implications for sampling design are discussed with an emphasis on the influence

(Continued)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code) / 22c. OFFICE SYMBOL

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

19. ABSTRACT (Continued).

of parameter frequency distributions, spatial patterns, and temporal patterns on sampling design. From the consideration of reservoir water quality patterns, a generalized reservoir sampling design is developed. Methods for the evaluation of an implemented sampling design are also considered. The techniques of variance component analysis, cluster analysis, and error analysis can be used to determine the effectiveness of a given sampling design.

PREFACE

The material presented in this report is part of the Environmental and Water Quality Operational Studies (EWQOS) Program, Work Unit VIIA, Reservoir Field Studies, conducted by the US Army Engineer Waterways Experiment Station (WES) for the Office, Chief of Engineers (OCE), US Army. The OCE Technical Monitors were Dr. John Bushman, Mr. Earl Eiker, and Mr. James L. Gottesman.

The report was completed by the Aquatic Processes and Effects Group (APEG), Ecosystem Research and Simulation Division (ERSD), Environmental Laboratory (EL), WES. The report was prepared by Dr. Robert F. Gaugush, APEG. This report was prepared under the supervision of Dr. Thomas Hart, Chief, APEG; Mr. Donald L. Robey, Chief, ERSD; and Dr. John Harrison, Chief, EL. Dr. Jerome L. Mahloch was Program Manager of EWQOS. Editor for the report was Ms. Lee T. Byrne of the WES Information Technology Laboratory.

COL Allen F. Grum, USA, was the previous Director of WES. COL Dwayne G. Lee, CE, is the present Commander and Director. Dr. Robert W. Whalin is Technical Director.

This report should be cited as follows:

Gaugush, R. F. 1987. "Sampling Design for Reservoir Water Quality Investigations," Instruction Report E-87-1, US Army Engineer Waterways Experiment Station, Vicksburg, Miss.

CONTENTS

	Page
PREFACE	1
PART I: INTRODUCTION	3
Background	3
Purpose and Scope	4
PART II: STATISTICAL CONCERNS IN SAMPLING DESIGN	6
Statistical Inference and Sampling	6
Basic Statistics	7
PART III: SAMPLING DESIGN	12
Introduction	12
Determination of Sample Size	14
Simple Random Sampling	19
Stratified Random Sampling	23
Systematic Sampling	35
Decision Matrices	38
PART IV: RESERVOIR WATER QUALITY PATTERNS--IMPLICATIONS FOR SAMPLING DESIGN	46
Frequency Distributions	46
Spatial Patterns	49
Temporal Patterns	56
Interaction Between Spatial and Temporal Patterns	60
A Generalized Reservoir Sampling Design	62
PART VI: EVALUATION OF SAMPLING DESIGN	67
Introduction	67
Variance Component Analysis	68
Cluster Analysis	68
Error Analysis	72
REFERENCES	77
BIBLIOGRAPHY	79
APPENDIX A: SAMPLING DESIGN FORMULAE	A1

SAMPLING DESIGN FOR RESERVOIR WATER QUALITY INVESTIGATIONS

PART I: INTRODUCTION

Background

1. The US Army Corps of Engineers (CE) presently maintains and operates over 400 reservoir projects for the purposes of flood control, hydropower, navigation, water supply, and recreation. In an effort to better define reservoir water quality problems and to develop technological and managerial alternatives for ameliorating these problems in a manner consistent with project purposes, the CE initiated the Environmental and Water Quality Operational Studies (EWQOS) Program in 1978 and completed it in 1985. A major portion of this program consisted of limnological studies at four characteristic reservoirs. These reservoirs were Lake Red Rock, a flood control impoundment of the Des Moines River in central Iowa; DeGray Lake, a large hydropower project located in south-central Arkansas; West Point Lake, a large hydropower project created by the impoundment of the Chattahoochee River approximately 100 km downstream from Atlanta, Georgia; and Eau Galle Lake, a small eutrophic flood control reservoir in west-central Wisconsin.

2. Studies conducted at each of these sites since 1978 have been of two basic types: long-term monitoring and short-term, process-oriented studies. Long-term monitoring studies provided data needed for the calibration and evaluation of several numerical water quality models as well as information concerning general water quality characteristics and trends. Short-term studies, often conducted during specific seasons of the year or under a predetermined set of hydrometeorological conditions, were aimed at understanding processes impacting reservoir water quality. These types of studies provided a means for improving existing water quality models and have greatly increased the understanding of reservoir ecosystem processes.

3. The reservoir studies conducted under the EWQOS Program have also fostered the development of an understanding of the problems and concerns involved in reservoir water quality sampling. This manual summarizes the findings of the EWQOS Program as they pertain to the development of reservoir water quality sampling designs.

Purpose and Scope

4. The purpose of this report is to provide CE Division and District personnel a general introduction to the statistical concerns involved in the formulation and design of reservoir water quality sampling programs. The major and most common concepts and techniques involved in the development of water quality sampling designs will be discussed. This report is not intended to replace reference works on sampling design; rather it provides the necessary background for the effective and efficient use of those references. Also, this report does not provide specific guidance on the water quality variables to be measured, sample collection methods, or analytical techniques to be used in the analysis of the samples collected. The application of specific collection methods, the selection of parameter lists, and the analyses to be performed will be dependent on the nature of the perceived or expected water quality problem and site-specific features of the reservoir. Guidance on these topics can be found in the References and Bibliography.

5. This report is intended for use by all personnel involved in the design, implementation, and data interpretation of water quality monitoring programs in CE reservoirs. Most of the information contained in this report is discussed in a more detailed manner in other sources. A number of introductory statistics textbooks are identified in the References and Bibliography, and it is suggested that at least one of these be on hand as the report is read. These textbooks also contain mathematical and statistical tables that will be required for the implementation of the techniques presented in this report. This report

is also intended to serve as a companion manual to "Statistical Methods for Reservoir Water Quality Investigations" (Gaugush 1986).

6. The presentation of material in this report is divided into several parts. Part II presents the general statistical background necessary for an understanding of the more specific material to follow. Part III, which presents the core of the statistical treatment on sampling design, deals with the determination of sample size and the methods for allocating those samples with respect to space or time. General spatial and temporal patterns in reservoir water quality and their implications for sampling design are discussed in Part IV. Part V describes three methods for the evaluation and possible modification of various sampling designs. Appendix A lists all of the calculation formulae presented in the report and gives the page number where more detailed information can be found.

PART II: STATISTICAL CONCERNS IN SAMPLING DESIGN

Statistical Inference and Sampling

7. Sampling and statistical data analysis make it possible, under certain assumptions, to infer the characteristics of the whole from the characteristics of a limited number of its parts. Sampling consists of taking a very limited number of observations from the much larger set of possible observations. The set of possible observations is referred to as the target population, and the limited subset of observations is called the sample population or, simply, the sample. Examples of target populations may include all the reservoirs in a specified geographical region, all of the coves in a given reservoir, or the water in a single cove. Corresponding sample populations may be three of the reservoirs from the specified region, five selected coves, or 10 ℓ of water from the single cove. Sampling design is concerned with the methods for determining the number of observations that comprise the sample and the selection of those observations from the target population.

8. Sampling design and the statistical analysis of the data collected make two very important assumptions: (a) the values of the target population are normally distributed and (b) the values of the sample population are independent. The assumption of a normal distribution of the target population is required because most statistical methods have been developed for normally distributed populations. However, most statistical tests are robust with respect to deviations from normality. The term "robust" implies that the tests may still be applied with confidence if the distribution of the target population is suspected of minor deviations from the normal distribution. Methods for the analysis of nonnormal populations, called distribution-free or nonparametric statistics, exist, and the reader can find an introduction to the subject in Gaugush (1986). The assumption of independence is also a requirement for statistical data analysis. Independence of the values of a sample is assured if every possible observation of the target population has an equal chance of being selected for the sample.

The lack of independence will usually result in estimates that are seriously biased.

9. The major objective of sampling design is to provide a means for obtaining an accurate and precise estimate of the target population within the constraints of funding and time available for the study. Accuracy refers to the relationship between the estimate and the actual target population value. An accurate estimate will, on the average, be centered on the population value, whereas an inaccurate estimate will consistently tend to overestimate or underestimate the population value. An inaccurate estimate is said to be biased. Precision refers to the reliability of the estimate and to the variability between repeated measures of the same quantity. An estimate that varies widely between repeated measures is said to lack precision.

10. Accuracy is essentially a function of the analytical techniques used to ascertain the value of a given observation. For example, the phosphorus concentration of a given water sample is determined by a colorimetric test, and the accuracy of the estimate is dependent on the detection limit of the technique, the technician's competence, and the quality of the analytical equipment. An estimate's accuracy can be assured through quality control, technician training, and careful selection of the analytical technique. Precision, on the other hand, is a function of the inherent variability of the target population and the number of observations used to produce the estimate.

11. A discussion of basic statistics is presented in the following section with the objective of providing the background required for an understanding of the sampling design concepts to be presented in Part III.

Basic Statistics

12. Sampling will most often be conducted to generate estimates of the target population parameters. Sample estimates are referred to as statistics. The two most common parameters of interest are the population mean μ , which is a measure of central tendency, and the

variance σ^2, which is a measure of the dispersion about the mean. The mean is the sum of all the individual observations y_i in the target population divided by the total number of observations or units N in the population

$$\mu = N^{-1} \sum_{i=1}^{N} y_i \tag{1}$$

while the population variance is

$$\sigma^2 = N^{-1} \sum_{i=1}^{N} \left(y_i - \mu\right)^2 \tag{2}$$

13. The population parameters cannot be determined directly (unless every y is known); rather they are estimated by their corresponding statistics, the sample mean and variance. A sample, in its simplest form, consists of n units selected at random from the N units of the target population. The sample mean is

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i \tag{3}$$

and the sample variance is

$$s^2 = \left(n - 1\right)^{-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \tag{4}$$

where

$\bar{y}$ = sample mean

n = sample size

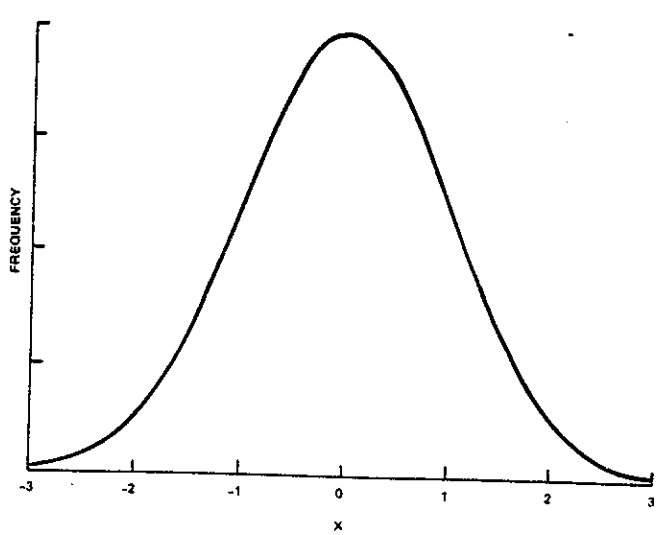
s^2 = sample variance

Using the $(n - 1)$ term rather than n corrects for the bias introduced by using $\bar{y}$ to estimate the population mean.

14. The standard deviation, the square root of the variance, describes the variability of the target population values about the population mean. If the population values are normally distributed, 95 percent of the values lie between the bounds $\mu - 1.96\ (\sigma)$ and $\mu + 1.96\ (\sigma)$. Figure 1 presents the normal distribution for a population with $\mu = 0$ and $\sigma^2 = 1$. Since $\sqrt{\sigma^2} = \sigma$ and $\sqrt{1} = 1$, it can easily be seen that the majority of the values (in fact, greater than 95 percent) lie in the interval bounded by -2 and +2. In other words, greater than 95 percent of the values fall within two standard deviations of the mean.

15. One additional statistical concept, the distribution of the sample mean, must be considered before proceeding to a discussion of sampling design. A distribution of sample means would result if one were able to take a number of independent samples (where a single sample is a set of observations selected from the population) from the population with each sample producing its respective sample mean. The set of sample means would then have a frequency distribution describing the variability of the sample means. If the observations in the target population are normally distributed, then the distribution of sample means

Figure 1. Standardized normal



will approximate the normal distribution. Even if the target population is not normally distributed, the distribution of the sample means will approximate the normal distribution, if the sample size is sufficiently large.

16. If it is impossible to collect a sufficient number of samples to determine directly the variability of the sample means, there must be a method for estimating this variability. The variance of the sample means can be estimated by

$$s^2_{\bar{y}} = \frac{s^2}{n} \tag{5}$$

where $s^2_{\bar{y}}$ = variance of sample means .

17. What is the purpose behind an estimate of the variance of the sample mean? To answer this question, another statistic, the standard error, defined as

$$s_{\bar{y}} = \left(s^2_{\bar{y}}\right)^{1/2} = \left(\frac{s^2}{n}\right)^{1/2} \tag{6}$$

must be considered. The standard error describes the variability of the sample means in the same manner that the standard deviation describes the variability of the observations in the target population. It was shown earlier that in a normal distribution 95 percent of the population values are contained in the interval defined by approximately two standard deviations above and below the population mean. Knowledge of the variance of the sample means and, in turn, the standard error allows the construction of an interval about the population mean that contains 95 percent of all possible sample means. In other words, the standard error determines an interval that, with high probability, 0.95 for example, will contain the population mean.

18. The interval defined by the standard error is referred to as a confidence interval. The confidence interval provides a measure of the precision of the sample mean. A relatively small confidence interval implies a very precise estimate of the population mean. The confidence interval is defined by the confidence limits

$$L_L = \bar{y} - ts_{\bar{y}} \tag{7}$$

$$L_U = \bar{y} - ts_{\bar{y}} \tag{8}$$

where

L_L = lower confidence limit

t = Student's t value for $n - 1$ degrees of freedom and a specified probability

L_U = upper confidence limit

19. The basic statistics presented above should provide sufficient background for an understanding of the material to follow. Additional and more detailed information can be found in Gaugush (1986) or any of the introductory statistical texts listed in the Bibliography.

PART III: SAMPLING DESIGN

Introduction

20. The purpose of sampling theory is to make sampling more efficient. Given that both dollars and manpower are in limited supply, a sampling program must provide the most information possible per unit cost and per man-hour expended. Sampling theory has developed methods that, at the lowest possible cost, provide estimates that are sufficiently precise to meet the objectives of the study.

21. Five distinct steps are involved in the design, implementation, and analysis of any sampling program:

a. Problem identification.

b. Statement of the objective.

c. Formulation of the sampling design.

d. Implementation of the sampling design.

e. Data analysis.

This report will consider only the third step and will not address collection techniques, analytical techniques, or the statistical analysis of the data collected. These topics are adequately treated in the following sources:

a. Collection and analytical techniques:

(1) "Reservoir Water Quality" (Office, Chief of Engineers (OCE) 1985).

(2) "General Guidelines for Monitoring Contaminants in Reservoirs" (Waide 1986).

(3) "Handbook for Sampling and Sample Preservation of Water and Wastewater" (US Environmental Protection Agency 1982).

(4) Methods for Collection and Analysis of Aquatic Biological and Microbiological Samples (Greeson et al. 1977).

(5) Methods for Collection and Analysis of Water Samples for Dissolved Minerals and Gases (Brown, Skougstad, and Fishman 1970).

(6) Methods for Determination of Inorganic Substances in Water and Fluvial Sediments (Skougstad et al. 1979).

(7) National Handbook of Recommended Methods for Water-Data Acquisition (US Geological Survey (USGS) 1977).

(8) "Procedures for Handling and Chemical Analysis of Sediment and Water Samples" (Plumb 1981).

(9) Recommended Methods for Water Data Acquisition (USGS 1972).

(10) Standard Methods for the Examination of Water and Wastewater (American Public Health Association 1980).

b. Statistical analysis:

(1) "Reservoir Water Quality" (OCE 1985).

(2) "Statistical Methods for Reservoir Water Quality Investigations" (Gaugush 1986).

A general treatment of statistical data analyses can also be found in any of the introductory statistics textbooks listed in the Bibliography.

22. The identification of an existing or expected water quality problem is the first step in the development of a sampling design. A problem statement serves to identify the specific water quality concern, the possible causes of the problem, and the related water quality responses to the problem situation. For example, if the occurrence of algal blooms is identified as the water quality problem, then one possible cause may be excessive nutrient loading. Algal blooms may, in turn, be the cause of other water quality problems, such as impaired recreational use of the reservoir or taste and odor problems in a water supply reservoir.

23. Once a problem is identified, the objectives of the sampling program can and should be explicitly stated. For the example above, the objectives may be to:

a. Determine the nutrient load to the reservoir.

b. Analyze the relationship between nutrient load and resultant algal biomass.

c. Determine the frequency of occurrence of algal blooms.

d. Relate the observed frequency of algal blooms and their magnitude to the impairment of recreational usage.

The explicit statement of objectives allows for their prioritization. Prioritizing objectives makes it possible to decide which objectives

must be met and which objectives may be dropped as a result of funding or manpower constraints. With a problem identified and the objectives explicitly stated, sampling design can be considered.

24. The sampling design must provide the answers to four fundamental questions:

a. What to sample?

b. How many samples?

c. Where to sample?

d. When to sample?

The answer to the first question is a list of the parameters to be measured. The formulation of this list is so problem and site specific that little general information can be provided. If the problem identification and the objective statement phases are carried out properly, the parameters of interest should be obvious. The Environmental Engineer Manual "Reservoir Water Quality" (OCE 1985) and "General Guidelines for Monitoring Contaminants in Reservoirs" (Waide 1986) can be consulted for information on important water quality parameters. Answers for the second, third, and fourth questions (i.e., sample size, sample placement, and sampling frequency, respectively) are much more amenable to a general treatment.

25. After determining the sample size, sampling design is essentially a statement of where and when to sample. There are three fundamental methods for allocating samples in space or time: (a) simple random sampling, (b) stratified random sampling, and (c) systematic sampling. All of the designs that one will normally encounter in reservoir sampling are either one of these designs or some combination of the above. The discussion of the three basic designs that follows applies equally well to the allocation of samples in space (sample placement) and to the allocation of samples in time (sampling frequency).

Determination of Sample Size

26. One of the most important steps in the design of a sampling program is the determination of sample size, or the number of samples

that will be collected during the course of the study. If too many samples are taken, then the program will not be cost-effective; and if too few samples are collected, then the sampling program may not meet the stated objectives of the study. How is the decision concerning the number of samples to be made? Sampling theory provides the means to address this problem.

27. An example will help to clarify the steps involved. Suppose that an investigator wants to estimate the trophic state of a particular reservoir and to do so he has decided to determine the mean summer chlorophyll *a* concentration. He has also decided to take a simple random sample from the reservoir's epilimnion during the months of May through September. How many samples should he collect? Before this question can be answered, the investigator must be able to specify the precision with which he desires to measure the chlorophyll *a* concentration. Assume that the objectives of the study can be met if the final estimate of the chlorophyll concentration is correct within ±10 percent. That is, if the mean of the sample is 15 μg/ℓ, then the value for the reservoir is expected to be between 13.5 and 16.5 μg/ℓ (i.e., 15 ± 1.5 μg/ℓ).

28. It is also necessary to specify the probability of error that the investigator is willing to accept. No matter how large the sample is, there is a chance that, once the samples are collected, the parameter estimated is in error by more than the desired 10 percent. Only by sampling the entire target population can the probability of error be ignored. Most often a 5 percent probability of error is chosen, but there may be cases in which a larger probability of error is acceptable.

29. Given the desired precision and the acceptable level of error, the determination of sample size can be considered. Sample size for a simple random sample can be calculated from the following equation:

$$n = \frac{t^2 s^2}{(r\bar{y})^2} \qquad (9)$$

where r = the desired precision. In this formula there is a very basic dilemma. In order to calculate the number of samples to be taken, some very specific information (i.e., the mean and variance) about the population to be sampled is required. Since this information is unknown (if it were known, sampling would be pointless), the investigator must use his background and expertise to provide these values.

30. There are essentially three methods for estimating the target population's mean and variance: (a) conducting a pilot study, (b) using the results from previous studies of the same population or a similar population, and (c) making an educated assessment about the nature of the population. A pilot study may be feasible if a large or long-term sampling program is being planned, but in many cases the use of a pilot study will be unavailable because of funding or time constraints. The best methods for estimating the mean and variance rely on existing data. Data from previous studies, even if extremely limited in scope, will provide acceptable estimates if one very important assumption can be made: that is, the target population has not undergone any serious changes since the study was conducted. For example, this assumption can not be made if the reservoir's watershed has been subjected to a major increase in development since the time of the last data set. The limitation of this assumption is even more significant if data from a similar, as opposed to the same, population are used. Rarely will the final method, making an educated assessment of the population parameters, be necessary.

31. In the example above, it can be assumed that the investigator has appropriate data and that in previous years the chlorophyll *a* mean was 15 μg/ℓ with a standard deviation of 7 μg/ℓ. The determination of sample size is an iterative process because the t value is dependent on the sample size. As a first approximation, a Student's t value of 2.0 is chosen because the value of t is approximately 2.0 at all but the smallest sample sizes. So, the first approximation is

$$n = \frac{(2.0)^2(7)^2}{[(0.10)(15.0)]^2}$$

$$n = \frac{(4.0)(49)}{(1.50)^2}$$

$$n = \frac{196.0}{2.25}$$

$$n = 87.11 \simeq 87 \tag{10}$$

32. Given this first approximation of sample size, the number of samples can be recalculated using a t value with $86(n - 1)$ degrees of freedom (DF). The use of a t table shows that a DF of 86 and a 5 percent probability of error result in a Student's t of 1.987. Recalculating n gives

$$n = \frac{(1.987)^2(7)^2}{[(0.10)(15.0)]^2}$$

$$n = \frac{(3.95)(49)}{(1.50)^2}$$

$$n = \frac{193.55}{2.25}$$

$$n = 86.02 \simeq 86 \tag{11}$$

A sample size of 86 will provide the desired 10 percent precision about the mean with a 5 percent probability of error.

33. Rarely will a sampling program be designed to sample for a single variable. How then is the sample size determined for a multivariable study? In determining the overall sample size, it is necessary to prioritize the parameter list. Certain of the parameters to be

sampled will be critical to the objectives of the study, whereas others will be of secondary importance. Using the critical variables, an estimation of the sample size can be made separately for each of the variables. It may happen that all of the separate n's are reasonably similar. If the largest of these n's can be handled within budgetary and time constraints, then it can be used for all of the variables. More often, there will be considerable disparity between the values of the separate sample sizes. In such a case if the largest n is chosen for all the variables, the sampling program may exceed the project funding level, and some of the variables will be measured with a higher precision than required. At this point two options are available: (a) each variable can be measured using its own sample size, or (b) a smaller overall n can be chosen with the knowledge that the precision of some of the variables will be reduced. The first option will rarely be feasible because of the demands it places on the personnel conducting the study. To randomly sample for a number of variables, each with its own sample size, requires a separate sample design for each of the variables. Each of the variables must have its own set of sampling stations and/or times of sampling. The second option results in much more tractable designs. Selecting a smaller overall n will reduce the precision of those variables where the original value of n is larger. The precision that will result from using the smaller sample size can be determined by rearranging the equation for sample size. Given

$$n = \frac{t^2 s^2}{(r\bar{y})^2} \tag{12}$$

then

$$r^2 = \frac{t^2 s^2}{n\bar{y}^2} \tag{13}$$

and

$$r = \frac{ts}{\sqrt{n}\,\bar{y}} \tag{14}$$

If the value of the precision is within acceptable limits, the smaller value of n can be chosen.

34. The use of decision matrices to determine overall sample size in a multivariable sampling program will be discussed in a later section after the presentation of the use of simple random, stratified random, and systematic sampling.

Simple Random Sampling

35. Simple random sampling is a procedure for selecting n units out of the N possible units of the target population so that all possible samples have an equal chance of being selected. This selection process is carried out by selecting units one by one. The units in the target population are numbered from 1 to N ; then a series of random numbers between 1 and N is selected from either a table of random numbers or from a computer program that produces such a series. Any selection during the process must give an equal chance of selection to any unit in the target population not already selected. The set of units that bears these n numbers comprises the simple random sample. Since a number that has been selected is removed from the target population, this method of sample selection is also referred to as random sampling without replacement.

36. Selection of a simple random sample most often makes use of a table of random numbers, which is a table of the digits 0-9 with each digit having an equal probability of selection. Tables of random numbers usually appear in standard statistical texts. Table 1 presents a table of 1,000 random digits.

37. To illustrate the use of a random number table to select a simple random sample, consider the data presented in Table 2. The first

step in the selection process is to number the units of the target population from 1 to N; the numbers in Table 2 represent a target population where N = 100. Assume that a sample size of 10 (i.e., n = 10) is wanted. Select two adjacent columns from Table 1, perhaps columns 10 and 11, and select the first 10 distinct numbers between 1 and 100. If the number 00 is encountered, select unit 100 from the target population because the number 100 cannot be selected from two-digit numbers. The selection process results in the numbers 30, 01, 00, 73, 38, 24, 07, 61, 62, and 65. Referring to the target population, the values 17.65, 13.09, 2.44, 18.08, 5.42, 12.04, 9.36, 18.22, 9.35, and 10.66 make up the random sample. If another random sample is to be selected, it will be necessary to choose two different columns in the table of random numbers.

Table 1

One Thousand Random Digits

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	84224	73865	30215	16205	37120	42108	70666	85264	61135	10612
01	06863	21047	01207	73612	44320	14862	04105	38665	13366	54756
02	04254	06221	00473	64110	32157	77377	84184	50283	73037	73326
03	88805	82344	73303	67081	76278	72646	38005	53704	72371	70532
04	28468	38604	38542	81650	81537	46088	83220	41650	24404	37740
05	67521	43813	24205	51611	75740	22304	42776	37700	50706	84413
06	76275	38554	07081	30322	15314	14405	73811	40456	21115	10242
07	81745	64828	61274	28860	28231	33543	54836	70361	73104	51138
08	86485	76066	62183	33286	65101	11833	37552	62068	76344	54352
09	01270	26336	65848	10814	80625	84144	64148	33164	60273	07147
10	44113	81582	75013	77542	41872	17434	30302	71678	56468	56420
11	04688	85443	40645	12111	27885	15301	11828	04705	15736	51780
12	61534	43205	75207	87186	86340	24224	05309	33345	12467	48466
13	40751	43168	75714	41700	88212	44251	03641	61041	51830	25384
14	16285	74357	08705	28733	04486	73328	37074	12360	86075	65636
15	45521	66638	86411	61501	70852	88812	50215	33415	45145	22211
16	81326	38380	66026	22453	65554	44188	74738	83872	85848	20735
17	56037	05256	86205	08832	53844	46280	80112	26226	85741	74458
18	67252	01512	60821	26176	15528	21636	81180	16837	33352	50643
19	80625	40062	21582	52621	24154	57132	25133	53778	88311	65280

Table 2

Target Population Values

	01-20	21-40	41-60	61-80	81-100
1	13.09	18.17	5.50	18.22	6.50
2	5.13	11.92	10.13	9.35	6.75
3	6.08	17.10	7.53	10.62	17.24
4	4.84	12.04	5.15	9.08	6.04
5	2.29	2.59	8.64	10.66	13.35
6	17.94	7.24	15.43	2.97	12.92
7	9.36	15.86	9.53	6.41	12.62
8	3.94	0.31	6.78	17.62	8.90
9	4.42	11.91	6.96	13.76	10.25
10	12.27	17.65	10.75	11.68	13.22
11	8.39	9.47	3.40	16.17	4.89
12	10.11	12.44	8.50	11.19	6.45
13	1.81	10.96	10.98	18.08	5.74
14	3.96	6.01	17.44	12.76	8.77
15	3.01	14.44	20.17	15.11	9.77
16	5.01	18.47	12.37	17.02	0.83
17	11.67	12.87	9.59	1.53	8.23
18	5.13	5.42	6.75	7.96	16.03
19	6.35	9.69	9.44	3.02	4.11
20	11.93	9.54	1.98	10.58	2.44

38. A simple random sampling can be undertaken for a number of objectives, and most often the purpose or interest centers on the determination of the mean, variance, and the standard error or precision. The sample mean is

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i \tag{15}$$

where y_i = the ith value of the sample . The sample variance is

$$s^2 = (n - 1)^{-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \tag{16}$$

and the standard error is

$$s_{\bar{y}} = \left(\frac{s^2}{n}\right)^{1/2} \tag{17}$$

39. An example will help to illustrate the concepts presented above. Figure 2 represents the frequency distribution of a target population with a mean of 50.70 and a standard deviation of 16.16. Assume that the objective is to randomly sample this population and produce an estimate of the mean within an error of ±10 percent. Assuming a Student's t value of 2.0, the necessary sample size can be calculated from

$$n = \frac{(2.0)^2(16.16)^2}{[(0.10)(50.70)]^2}$$

$$n = \frac{(4.0)(261.15)}{(5.07)^2}$$

$$n = \frac{1{,}044.60}{25.71}$$

$$n = 40.63 \simeq 41 \tag{18}$$

To illustrate the relationship between sample size and the resulting precision, sample sizes of 90, 45, and 21 are used. For each of these sample sizes, four different random samples from the target population are selected. The results of this series of samplings are presented in Table 3. For the mean to be within the desired precision of 10 percent,

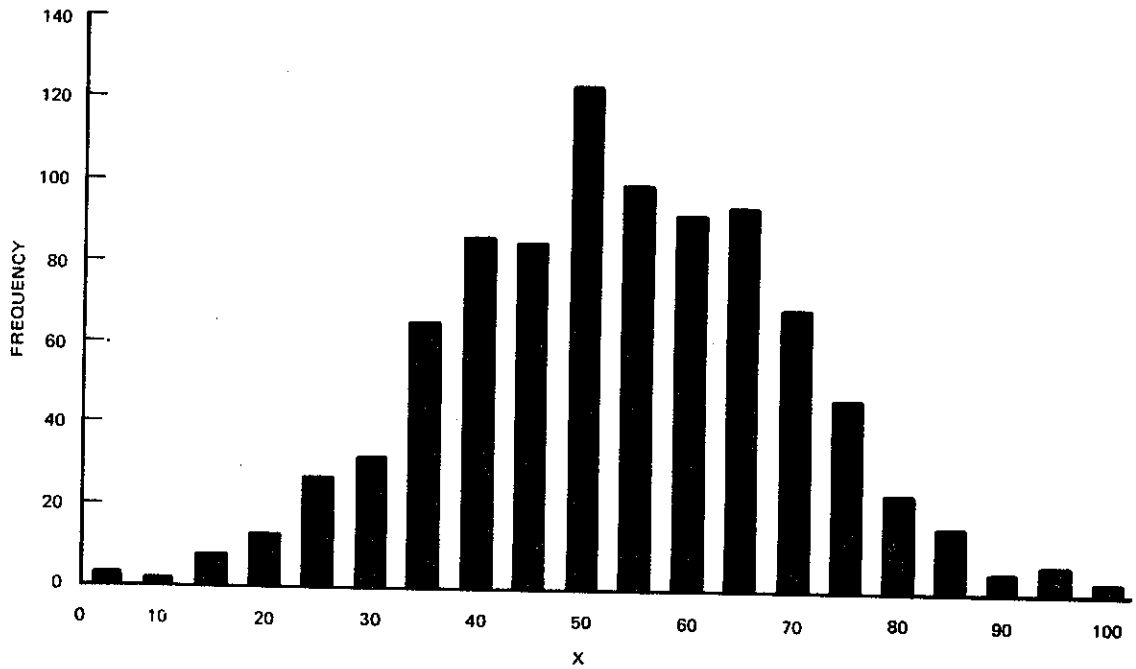


Figure 2. Target population frequency distribution

the standard error of the mean must be less than or equal to 5.07 (0.10 × 50.70). Note that all three sample sizes result in standard errors below 5.07. The occurrence of precision greater than the desired precision will often result when the variance of the target population is overestimated.

Stratified Random Sampling

40. Stratified random sampling consists of dividing the target population into a distinct set of subpopulations. The target population with N units is divided into subpopulations of N_1, N_2, N_3, ..., N_L units where the subpopulations do not overlap, and together they make up the entire target population or

$$N_1 + N_2 + \ldots + N_L = N$$

The subpopulations are referred to as strata. If a simple random sample

is drawn from each stratum, then the sampling design is referred to as stratified random sampling.

41. Stratified random sampling has two important advantages over simple random sampling. First, it can be advantageous to have data on separate subsets of the target population. For example, in a reservoir sampling program with the purpose of determining the annual water quality, it would also be useful to have data on the seasonal differences in

Table 3

Target Population Estimates for Simple Random Samples

Sample Size	Sample Mean	Sample Variance	Standard Error
90	51.22	227.59	1.59
	51.32	195.48	1.47
	49.84	290.92	1.80
	54.55	219.81	1.56
45	50.28	225.47	2.24
	51.50	333.77	2.72
	49.51	256.52	2.39
	53.23	318.49	2.66
21	48.50	263.97	3.55
	50.75	232.38	3.33
	51.22	214.23	3.19
	52.71	317.63	3.89

water quality. In this situation, the year could be divided into seasonal strata. Second, stratified sampling may produce an increase in the precision (i.e., reduce the error) of the estimates for the entire population. The concept behind stratification involves dividing a heterogeneous population into more homogeneous subpopulations. If the measurements within a stratum vary little from one unit to another, a precise estimate of the stratum mean can be obtained with relatively few samples. The estimates for the strata can then be combined into a precise estimate for the target population. The total number of samples used in such a design will often be less than would be required by a simple random sampling design.

42. The mean, variance, and standard error (or precision) are not calculated in the same manner, as those estimates were derived from a simple random sample. The following notation will be used throughout the remainder of this section. The subscript h refers to the stratum and the subscript i refers to a unit within that stratum.

N = total number of units in the target population

N_h = number of units in stratum h

n_h = sample size for stratum h

y_{hi} = value of the ith unit in stratum h

$w_h = N_h/N$, stratum weight

$f_n = n_h/N_h$, sampling fraction in stratum h

$$\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi} \text{ , stratum mean}$$

$$s_h^2 = \left(n_h - 1\right)^{-1} \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2 \text{ , stratum variance}$$

L = number of strata

43. The concept of the stratum weight needs some explanation before proceeding to the calculation of the estimates. Since the target population is divided into strata that may differ in size, it is necessary to weight the estimates in the calculation of the characteristics of the entire target population.

44. The target population mean derived from a stratified sampling program can be calculated from

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h \tag{19}$$

where the subscript st refers to stratified. The mean $\bar{y}_{st}$ is not, in general, the same as the sample mean calculated from all of the y_i values. The sample mean would be

$$y = n^{-1} \sum_{h=1}^{L} n_h \bar{y}_h \tag{20}$$

where

$$n = n_1 + n_2 + \ldots + n_L$$

45. The variance of the target population is the weighted sum of the strata variances, or

$$s^2_{st} = \sum_{h=1}^{L} \left(w_h^2\right)\left(s_h^2\right) \tag{21}$$

The precision of the target population mean (i.e., the standard error) can be calculated from

$$s^2_{(\bar{y}_{st})} = \sum_{h=1}^{L} \frac{\left(w_h^2\right)\left(s_h^2\right)}{n_h} \tag{22}$$

and

$$s_{(\bar{y}_{st})} = \left[\sum_{h=1}^{L} \frac{\left(w_h^2\right)\left(s_h^2\right)}{n_h}\right]^{1/2} \tag{23}$$

46. Developing a stratified sampling program proceeds through three steps: (a) identifying and weighting the strata, (b) determining the total sample size n_{st}, and (c) allocating the samples to the strata or determining the sampling fraction f_h. A complete discussion of strata identification will be deferred until the discussion of reservoir water quality patterns in Part IV, but the basics can be presented here. For stratification to lead to an increase in the precision of the final estimates, relatively homogeneous subsets of the target population must be identified. With a reservoir as the target population, the epilimnion, metalimnion, and the hypolimnion can be defined as the strata of interest. The weights of these strata are calculated from the proportion of the total reservoir volume represented by each stratum. For temporal sampling, such as sampling the inflow concentration of a nutrient in order to calculate nutrient loading, the duration of the stratum with respect to total duration of the sampling program defines the stratum weight. For example, because it is known that the nutrient load varies with the hydrologic flow, the strata can be defined on the basis of flow. One approach is to define a high flow stratum to cover that portion of the year, usually early spring, when flows are elevated and a low flow stratum for the rest of the year when flows are at or near base flow conditions. The weights will then be calculated by dividing the number of days in each stratum by 365. To reiterate, strata definition consists of identifying nonoverlapping subsets of the target population with variances less than the target population as a whole. These strata can then be weighted by dividing the number of units in each stratum by the number of units in the target population.

47. With the strata defined, the total number of samples can be estimated in a manner analogous to that used for simple random sampling. The sample size for a simple random sample was calculated from

$$n = \frac{t^2 s^2}{(r\bar{y})^2} \tag{24}$$

The total sample size for a stratified random sample is

$$n_{st} = \frac{t^2 s_{st}^2}{(r\bar{y}_{st})^2} \tag{25}$$

which can be expanded to

$$n_{st} = \frac{t^2 \sum_{h=1}^{L} w_h^2 s_h^2}{\left(r \sum_{h=1}^{L} w_h \bar{y}_h\right)^2} \tag{26}$$

From the expression above, it can be seen that the determination of total sample size requires estimates of the mean and variance for each of the strata. These can be obtained in the same manner as were the estimates for simple random sampling. It should be apparent that the increase in precision that can result from stratified sampling is not without cost. Whereas simple random sampling required only a single mean and variance, stratified sampling requires an estimate of the mean and variance for each stratum.

48. Given the total sample size, the final step in the development of the stratified random design requires the allocation of the samples to the strata. The three options available for the allocation of

samples are: (a) equal, (b) proportional, and (c) optimal. For equal allocation

$$n_h = \frac{n_{st}}{L} \tag{27}$$

where L = number of strata . In proportional allocation, the samples are assigned to the strata relative to the strata weights

$$n_h = w_h n_{st} \tag{28}$$

Optimal sample allocation considers not only the size of the strata (i.e., the strata weights), but also the variances of the strata and the cost of sampling each stratum. The cost of sampling can be estimated by a linear cost function

$$C = C_0 + \sum_{h=1}^{L} C_h n_h \tag{29}$$

where

C = total cost
C_0 = fixed cost or overhead
C_h = cost per sample in stratum h

The objective of including cost estimates in the sample design formulation is to allocate samples in such a way as to minimize costs. Using the cost function described in Equation 29, samples can be allocated optimally in the following:

$$n_h = n_{st} \frac{\frac{w_h s_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \left(\frac{w_h s_h}{\sqrt{C_h}} \right)} \tag{30}$$

In most cases, the cost per sample will not vary among the strata, and samples can be allocated optimally using a simplification of Equation 30:

$$n_h = n_{st} \frac{w_h s_h}{\sum_{h=1}^{L} (w_h s_h)} \tag{31}$$

From Equations 30 and 31, it can be seen that more samples will be allocated to a stratum if:

a. It is larger (w_h term).

b. It is more variable (s_h term).

c. It costs less to sample ($1/\sqrt{C_h}$ term).

49. In order to demonstrate the usefulness of stratification, reconsider the example that was presented in the section on simple random sampling. Figure 2 presented the population's frequency distribution, and it was determined that 41 samples would be required to estimate the mean with an error of 10 percent. Assume that three strata with the frequency distributions shown in Figure 3 can be defined. Each of these strata has a considerably lower variance than the target population. Table 4 provides the mean and variance for each stratum.

50. With these data, the number of samples required to estimate the mean with a precision of 10 percent can be determined. The total number of samples in a stratified design is given by

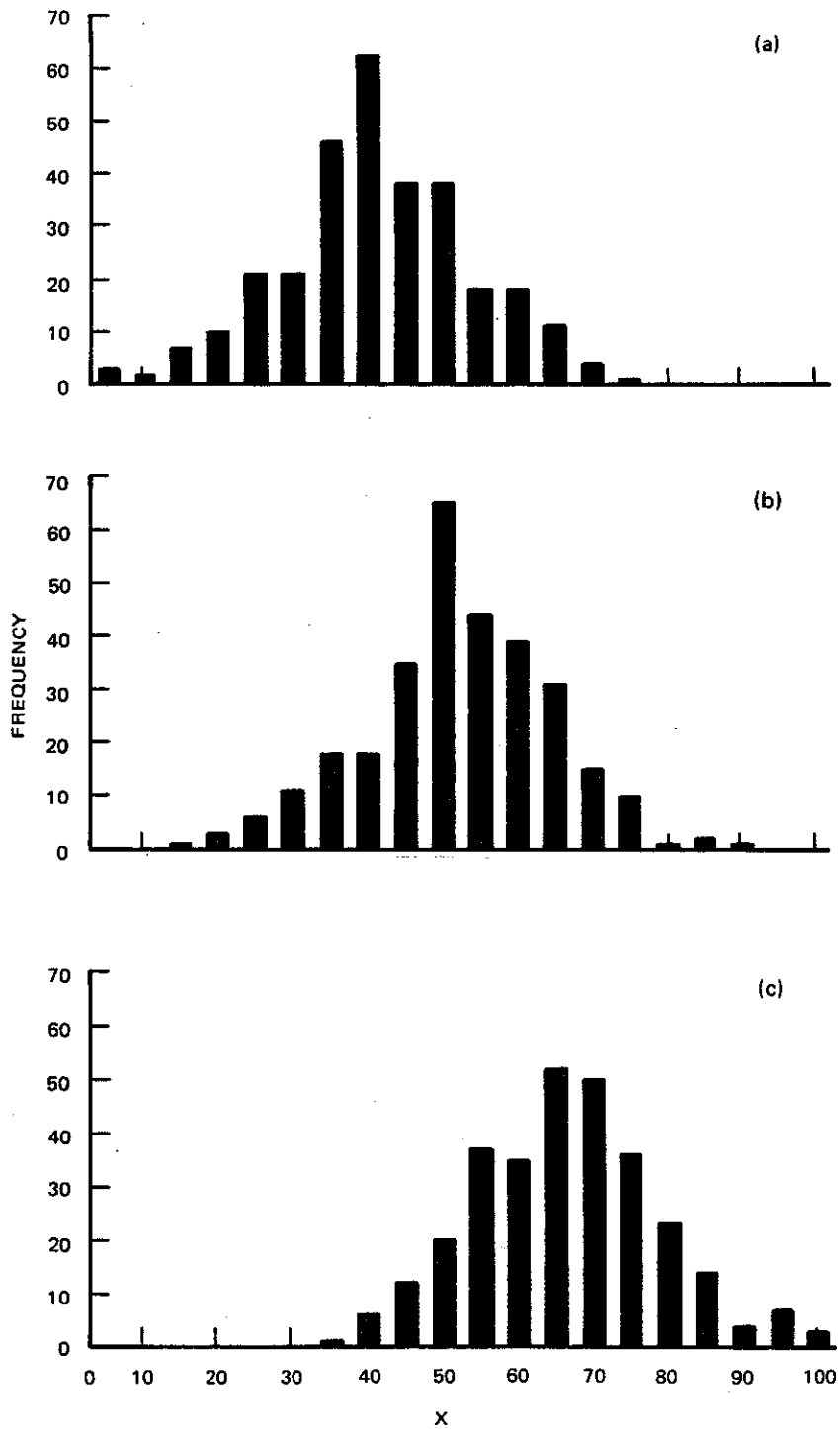


Figure 3. Strata frequency distributions for the sub-populations with means of 38.61 (a), 49.71 (b), and 63.79 (c)

Table 4
Estimated Means and Variances for the Defined Strata

Stratum	Mean	Variance	Weight
1	38.61	158.51	0.25
2	49.71	154.71	0.50
3	63.79	152.42	0.25

$$n_{st} = \frac{t^2 \sum_{h=1}^{L} w_h^2 s_h^2}{\left(r \sum_{h=1}^{L} w_h \bar{y}_h \right)^2} \tag{32}$$

For the data in Table 4

$$\sum_{h=1}^{L} w_h^2 s_h^2 = (0.25)^2(158.51) + (0.50)^2(154.71) + (0.25)^2(152.42)$$

$$= 58.11 \tag{33}$$

and

$$\sum_{h=1}^{L} w_h y_h = (0.25)(38.61) + (0.50)(49.71) + (0.25)(63.79)$$

$$= 50.46 \tag{34}$$

As a first step, assume an initial t value of 2.0; then n_{st} is

$$n_{st} = \frac{(2.0)^2(58.11)}{[(0.10)(50.46)]^2}$$

$$= 9.13 \simeq 9 \qquad (35)$$

Given that the t value with a DF of 8 $(n - 1)$ is 2.306, recalculate n_{st} as follows:

$$n_{st} = \frac{(2.306)^2(58.11)}{[(0.10)(50.46)]^2}$$

$$= 12.14 \simeq 12 \qquad (36)$$

Since the value of n has not converged yet, calculate it again using the t value for 11 DF (i.e., 2.201); thus

$$n_{st} = \frac{(2.201)^2(58.11)}{[(0.10)(50.46)]^2}$$

$$= 11.06 \simeq 11 \qquad (37)$$

One further iteration with 10 DF and a t value of 2.228 gives

$$n_{st} = \frac{(2.228)^2(58.11)}{[(0.10)(50.46)]^2}$$

$$= 11.33 \simeq 11 \qquad (38)$$

and sample size converges on 11 samples. The total sample size for the stratified design is almost one fourth the sample size for simple random sampling. Stratification in this case would lead to a considerable reduction in effort while still providing the same level of precision.

51. The example above can also be used to demonstrate the allocation of samples to the strata. For a program using equal allocation, it would be necessary to increase sample size by one and allocate four samples to each of the strata. Using proportional allocation, the stratum sample sizes would be

$$n_1 = w_1 n_{st} = (0.25)(11) = 2.75 \simeq 3$$

$$n_2 = w_2 n_{st} = (0.50)(11) = 5.50 \simeq 6$$

$$n_3 = w_3 n_{st} = (0.25)(11) = 2.75 \simeq 3 \qquad (39)$$

Here again it would be necessary to add one sample to the total sample size. Stratum sample sizes using optimal allocation and assuming that costs are constant among strata would be

$$n_h = n_{st} \frac{w_h s_h}{\sum_{h=1}^{L} (w_h s_h)} \qquad (40)$$

with

$$\sum_{h=1}^{L} (w_h s_h) = (0.25)(12.59) + (0.50)(12.44) + (0.25)(12.35)$$

$$= 12.46$$

$$n_1 = (11)\frac{(0.25)(12.59)}{12.46}$$

$$= 2.87 \simeq 3$$

$$n_2 = (11)\frac{(0.50)(12.44)}{12.46}$$

$$= 5.49 \simeq 5$$

$$n_3 = (11)\frac{(0.25)(12.35)}{12.46}$$

$$= 2.73 \simeq 3 \qquad (41)$$

In this example the use of optimal allocation does not produce a design that is any different from the design resulting from proportional allocation. This lack of a difference between the two designs is the result of the nearly equal variances of the three strata (see Table 4).

Systematic Sampling

52. Systematic sampling is considerably different from simple random sampling. Given that the N units of the target population are numbered from 1 to N in some order, then to obtain a sample of n units, a unit is selected at random from the first k units and every k^{th} unit thereafter. The value k is defined as the closest integer to N/n. For example, if N is 500 and a sample of 20 units is wanted, then

$$k = \frac{N}{n} = \frac{500}{20} = 25 \qquad (42)$$

To obtain a systematic sample of 20 units with a k of 25, a number between 1 and 25 is selected at random. If this random number is 12, then the systematic sample consists of the units

12, 37, 62, 87,..., 437, 462, 487

The selection of the first unit determines the entire sample. This type of sample is called an every k^{th} systematic sample.

53. Systematic sampling is a very common sampling method, particularly considering temporal sampling. For instance, a biweekly sampling program is an every k^{th} systematic sample with $k = 14$ days. The first sampling date can be chosen at random, and all the remaining samples are taken at 14 day intervals. Cochran (1977) has identified two apparent advantages of systematic sampling over simple random sampling. First, systematic sampling is an advantageous design because the development of the sampling program is relatively simple and easy to execute. It is for this reason that systematic sampling is so popular. For example, if an investigator wishes to sample the epilimnion of a 40-km-long reservoir, it will be much easier to select 20 samples spaced at 2-km intervals than to select 20 samples randomly scattered along the length of the reservoir. Second, it at first appears that systematic sampling is probably more precise than simple random sampling. Systematic sampling can be considered to stratify the target population into n strata, with the first stratum made up of the first k units, the second stratum made up of the second k units, and so on. Given this view of systematic sampling, it may be expected that the systematic sample will be nearly as precise as the corresponding stratified random sample with one unit sampled per stratum. But there is a difference. In the systematic sample, the sampled units occur at the same relative position in the stratum, whereas in the stratified random sample, the position of the sampled unit in the stratum is randomly selected. As a result, the systematic sample is spread more evenly throughout the population (a function of the k value), and this more even distribution sometimes leads to systematic sampling being considerably more precise than stratified random sampling.

54. With these advantages, why would anyone choose to conduct more difficult sampling programs using either simple random sampling or stratified random sampling? There are very obvious disadvantages to the use of a systematic design. Whereas simple random and stratified random samplings are primarily directed at providing estimates of the target

population as a whole (i.e., developing an estimate of the population mean and variance), providing these estimates can be very difficult from the data collected from a systematic design. The sample mean from a systematic sample can be calculated in a straightforward fashion from

$$\bar{y} = n^{-1} \sum_{k=1}^{n} y_k \tag{43}$$

but it is much more difficult to determine the sample variance and the standard error or the precision of the mean. If the sample values can be assumed to be random with respect to their order, the variance and standard error can be calculated in the same manner as a simple random sample

$$s^2 = (n - 1)^{-1} \sum_{k=1}^{n} \left(y_k - \bar{y}\right)^2 \tag{44}$$

and the standard error is

$$s_{\bar{y}} = \left(\frac{s^2}{n}\right)^{1/2} \tag{45}$$

Complications arise if the sample values display a trend or periodicity as a function of their order. If a trend is evident, such as a nearly linear decrease in chlorophyll *a* concentration from the headwaters to the dam of a reservoir, the calculation of an unbiased sample variance becomes very difficult. The problem arises because the successive values in the sample are no longer independent and the lack of independence violates one of the major sampling assumptions. Formulae for the calculation of the sample variance from systematic designs exist, but

they are beyond the scope of this manual. The reader can consult Cochran (1977) to pursue this subject.

55. With the above limitations in mind, a systematic sampling design can be effectively used to meet three different objectives. First, if the assumption of the independence of the observations can be made (i.e., samples are effectively random with respect to their order and no trends exist), then systematic sampling can represent an effective approach with execution being easier than simple random sampling. Second, systematic sampling can be used in a pilot study to define or identify stratum boundaries and provide the initial estimates for the determination of the total sample size in a stratified design. Third, the disadvantage of a stratified design can be turned to an advantage if the objective of the study is to detect trends or patterns. Whereas the random sampling designs are well suited to estimating the overall mean, they are poorly suited to providing a "picture" of the changes over time or space. A systematic design is much better suited to detect patterns in the behavior of the variable.

Decision Matrices

56. A decision matrix is an aid to the determination of sample size for multivariable sampling programs and can be used for either simple random or stratified random sampling designs. The decision matrix is simply a table that incorporates the necessary factors that determine sample size: (a) cost, (b) desired precision, (c) probability of error, (d) an estimate of the mean, and (e) an estimate of the variance.

57. Consider a sampling program with the objective of providing the annual means of three variables X, Y, and Z using a simple random sampling design. The information required to produce the decision matrix can be extracted from previous studies and is presented in Table 5. From these data, a decision matrix can be calculated using the equation for the sample size of a simple random sample

Table 5

Estimated Means, Variances, and Unit Costs

Variable	Mean	Variance	Unit Cost
X	12	100	15.00
Y	38	500	3.00
Z	72	650	20.00

$$n = \frac{t^2 s^2}{(r\bar{y})^2} \tag{46}$$

The decision matrix is presented in Table 6.

58. The decision matrix for the example presents a number of important features concerning the determination of the sample size in a multivariable sampling design. Note the differences in sample sizes as a function of precision and probability of error for any given variable. For example, the sample size for variable X ranges from 269 to 30 samples depending upon the desired precision and the probability of error. These data also help to illustrate the relationship between sample size, precision, and the probability of error. Accepting a reduced precision (i.e., accepting greater variance about the mean) has a much greater effect on sample size than allowing for a greater probability of error. For variable X , with an error probability of 5 percent, a desired precision of ±10 percent results in a sample size of 269, whereas if a precision of ±20 percent is acceptable, the sample size is reduced to 69. At a precision of ±10 percent, increasing the error of probability from 5 to 10 percent only reduces the sample size from 269 to 190. The effects of changes in precision and error probability are more apparent if the equation for sample size is considered. Sample size for a simple random sample is given in Equation 46 and can be rearranged to

$$n = \frac{1}{r^2}\left(\frac{t^2 s^2}{\bar{y}^2}\right) \tag{47}$$

Table 6

Decision Matrix for Variables X, Y, and Z

	Variable X					
Precision	10%			20%		
Probability	5%	10%	20%	5%	10%	20%
Number of samples	269	190	115	69	49	30
Total cost	$4,035	2,850	1,725	1,035	735	450
	Variable Y					
Precision	10%			20%		
Probability	5%	10%	20%	5%	10%	20%
Number of samples	135	96	58	36	25	16
Total cost	$ 405	288	174	108	75	48
	Variable Z					
Precision	10%			20%		
Probability	5%	10%	20%	5%	10%	20%
Number of samples	51	36	22	15	10	7
Total cost	$1,020	720	440	300	200	140

to account for effect of changes in precision or to

$$n = t^2 \left[\frac{s^2}{(r\bar{y})^2} \right] \tag{48}$$

to account for the changes in the probability of error. The limited range of t values (t ranges between 0.675 and 2.57 for error probabilities between 0.05 and 0.50 with DF > 5) restricts the effect of the error probability on sample size.

59. The decision matrix also displays the disparity of sample sizes among variables at a given precision and error probability. In the example matrix at a ±10 percent precision and a 5 percent probability of error, sample size ranges from 269 for variable X to 51 for variable Z . The occurrence of considerably different sample sizes for different variables is quite common and serves to complicate the decision to be made concerning sample size in a multivariable sampling program.

60. The decision matrix can also be used to examine the consequences of accepting any given sampling design. Cost estimates for the maximum and the minimum designs can be evaluated to provide a range of costs for the possible designs. The maximum design results from using the largest sample sizes from the decision matrix, and the minimum sample design results from using the smallest sample sizes. In the example decision matrix, the largest sample sizes are those for variable X , and the smallest sample sizes are from variable Z . Cost estimates for the maximum and minimum designs are presented in Table 7.

61. The maximum and minimum designs can also be examined for their influence on the precisions that will result from adopting either of these designs. The resulting precision can be calculated given the sample size, estimates of the mean and variance, and the probability of error, as follows:

Table 7

Maximum and Minimum Designs Based on the Decision Matrix of Table 6

	Maximum Design					
Precision*	10%			20%		
Probability	5%	10%	20%	5%	10%	20%
Number of samples	269	190	115	69	49	39
Total cost	$10,222	7,220	4,370	2,622	1,862	1,140
	Minimum Design					
Precision*	10%			20%		
Probability	5%	10%	20%	5%	10%	20%
Number of samples	51	36	22	15	10	7
Total cost	$ 1,938	1,368	836	570	380	266

* The level of precision refers only to that variable on which the design is based. For the maximum design only the precision of variable X is specified, and in the minimum design only the precision of Z is specified. The precisions of the remaining variables would have to be recalculated using the given number of samples.

$$r = \left(\frac{t^2 s^2}{n\bar{y}^2} \right)^{1/2} \tag{49}$$

Using the data in Tables 5 and 7, the resultant precisions can be calculated and are presented in Table 8.

62. The maximum design oversamples variables Y and Z , and as a result the precisions for these variables are greater (lower error about the mean) than the desired precisions specified in Table 6. If the maximum design is acceptable in terms of cost, this consequence does not in any way impair the design. On the other hand, the consequence of accepting the minimum design can seriously impair the sampling program. The minimum design undersamples variables X and Y , and the precisions are lower (greater error about the mean) than the desired precisions given in Table 6. If the mean of variable Z is the primary concern of the study, then the minimum design may be acceptable, but if X or Y is the variable of importance, then the minimum design will not provide the means of these variables within the desired bounds.

63. Decision matrices can also be produced for stratified random designs. If the cost of sampling each stratum is constant, the total sample size for variable Y is

$$n_{st} = \frac{t^2 \sum_{h=1}^{L} w_h^2 s_h^2}{\left(r \sum_{h=1}^{L} w_h \bar{y}_h \right)^2} \tag{50}$$

and the samples are allocated to the strata (assuming an optimal allocation) by

Table 8

Precisions of Variables X , Y , and Z Based on the Maximum and Minimum Designs

	Maximum Design					
Probability	5%	10%	20%	5%	10%	20%
Number of samples	269	190	115	69	49	29

Variable	Precision, %	
X	10	20
Y	7	14
Z	4	9

	Minimum Design					
Probability	5%	10%	20%	5%	10%	20%
Number of samples	51	36	22	15	10	7

Variable	Precision, %	
X	23	46
Y	17	34
Z	10	20

$$n_h = \frac{w_h s_h}{\sum_{h=1}^{L} w_h s_h} \tag{51}$$

The maximum and minimum designs, and their associated costs, can be calculated in the same manner as those for a simple random sample. The precisions resulting from either the maximum or minimum design can be calculated from

$$r = \left[\frac{t^2 \sum_{h=1}^{L} w_h^2 s_h^2}{\left(n \sum_{h=1}^{L} w_h y_h \right)^2} \right]^{1/2} \tag{52}$$

64. Decision matrices will not, in themselves, solve the problem of determining sample size in multivariable sampling programs. However, the matrices can be used to examine the consequences, in terms of cost and precision, of using any particular design. They are an aid in the development of a sampling design and can be an integral part of the decisionmaking process.

PART IV: RESERVOIR WATER QUALITY PATTERNS-- IMPLICATIONS FOR SAMPLING DESIGN

Frequency Distributions

65. One of the basic assumptions of sampling design is that the values of the target population are normally distributed. A problem arises in that a number of variables typically collected in a reservoir sampling program will frequently deviate from a normal distribution. Figures 4 and 5 present the percent frequency distributions for total phosphorus (TP), total nitrogen (TN), and chlorophyll *a* for samples collected from DeGray Lake, a flood control, hydropower reservoir in south-central Arkansas, and Eau Galle Lake, a flood control reservoir in west-central Wisconsin, respectively.

66. In DeGray Lake (Figure 4) the TP and TN distributions are positively skewed, whereas the chlorophyll *a* distribution is platykurtic. A skewed distribution has a concentration of values on one side of the distribution and a "tail" of less frequent values on the other. If the tail is on the right, the distribution is said to be positively skewed, and if the tail is on the left, the distribution is negatively skewed. As a result of the positive skew for TP and TN, the sample mean would be "pulled" away from the most frequently observed values of the sample. Therefore, the sample mean may not adequately characterize the target population. A platykurtic distribution, such as that of chlorophyll *a*, is defined as a distribution with a greater frequency of values between the mean and the tails than in a normal distribution. This results in a flatter distribution when compared with the bell-shaped normal distribution. All of the variables for Eau Galle Lake exhibit considerable positive skew (Figure 5).

67. A number of alternatives are possible for the treatment of deviations from normality, depending on whether or not the data have already been collected. If the data have been collected and deviations from normality are evident, the problem is one of data analysis. With nonnormally distributed data, data analysis can proceed by one of three

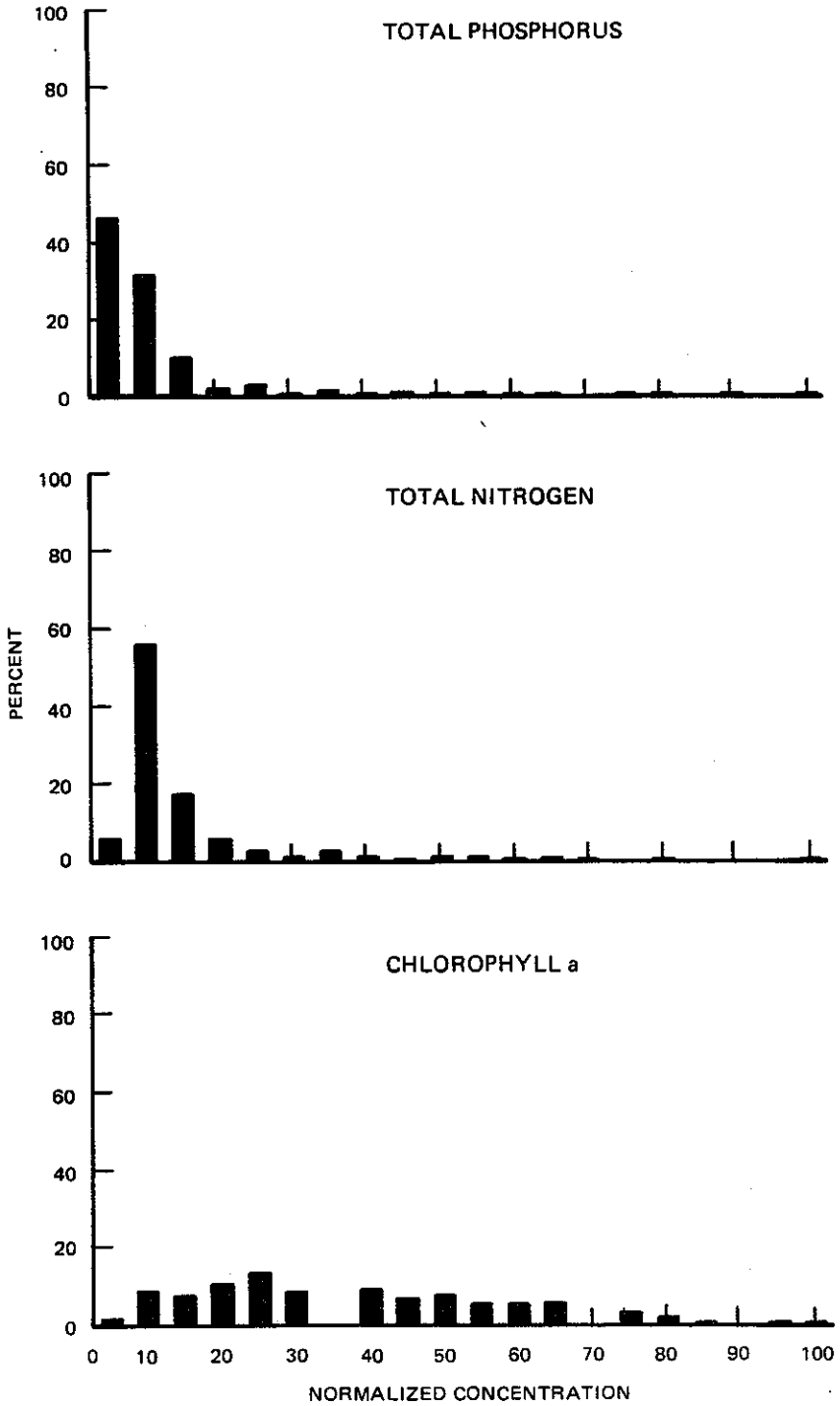


Figure 4. Normalized frequency distributions for total phosphorus, total nitrogen, and chlorophyll *a* from DeGray Lake. Normalized concentration is defined as (concentration/maximum concentration) × 100

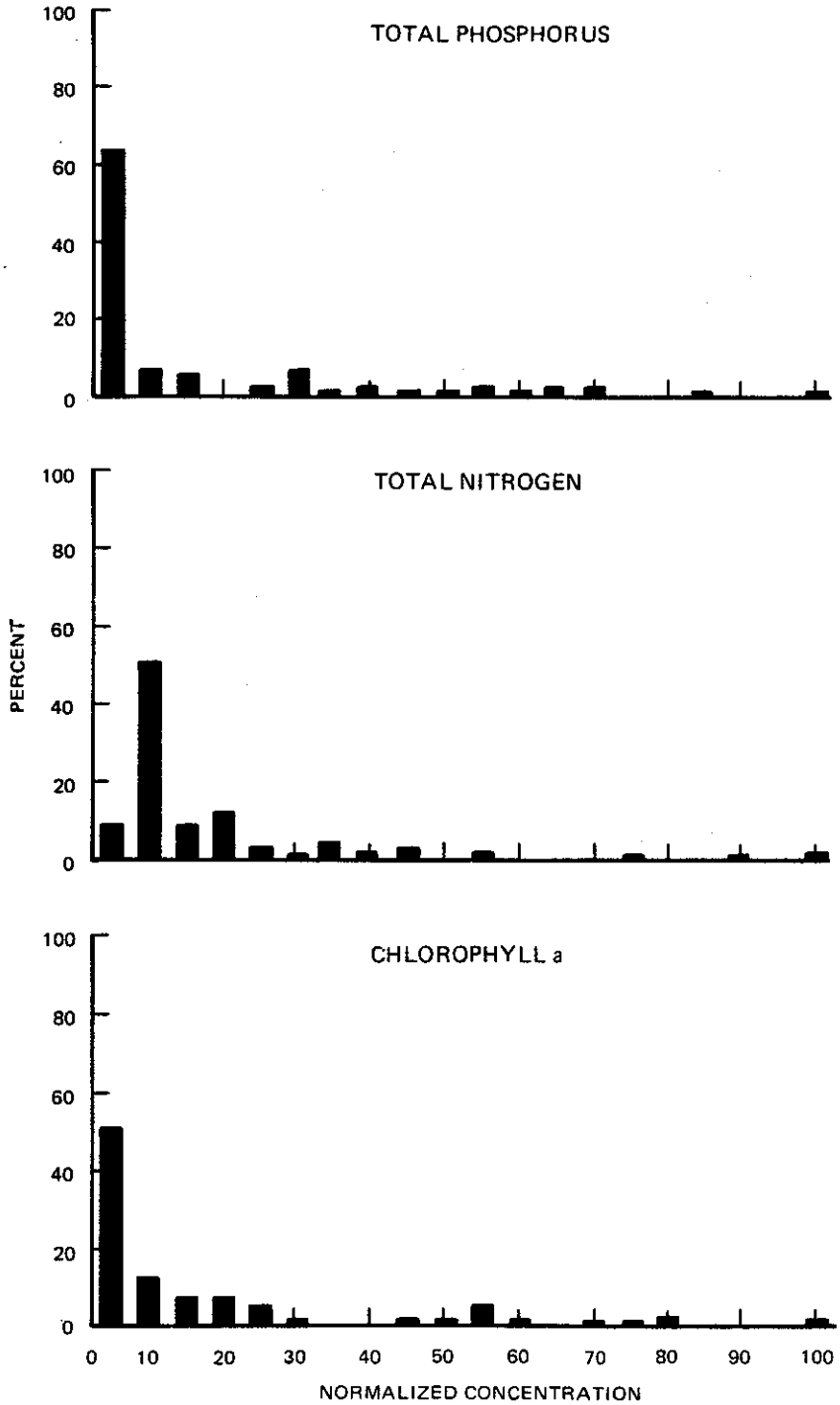


Figure 5. Normalized frequency distributions for total phosphorus, total nitrogen, and chlorophyll α from Eau Galle Lake. Normalized concentration is defined as (concentration/maximum concentration) × 100

approaches. One, it may be possible that the desired analyses are sufficiently robust that the deviations from normality will not impair the results. Two, a data transformation procedure can be applied to the data to produce more normal distributions. Three, if the deviations from normality are considered substantial and cannot be easily corrected by transformations, then nonparametric analyses can be applied to the data. The use of these three approaches is discussed in detail in Gaugush (1986).

68. If the sampling program is in the developmental stage and deviations from normality can be expected, then the problem is one of sampling design. Deviations from normality for water quality variables may result from the existence of distinct subpopulations of the target population. Each subpopulation may be normally distributed, but their combination may deviate from normality. The implication is that a stratified design with each stratum having a normal distribution is a beneficial approach. If stratification does not produce a normal distribution, then the investigator will have to rely on the data analysis methods discussed above.

Spatial Patterns

69. In reservoir water quality sampling, three spatial dimensions are of concern: (a) vertical, (b) lateral (along an axis perpendicular to the major hydrologic flow), and (c) longitudinal (along an axis parallel to the major hydrologic flow). Before proceeding to a discussion of these patterns, it is worthwhile to consider a study conducted under the EWQOS Program that was designed to assess the contribution of each of these dimensions to the overall variance.

70. Documenting the spatial variability in DeGray Lake was the objective of three intensive sampling efforts in 1978 and 1979. Discussion will center on the July 1978 sample; the reader can consult Thornton et al. (1982) for a complete discussion. Sampling transects were established from the dam to the headwaters (to account for longitudinal variability). Stations were located on each transect to sample

over the old river channel, along the littoral area on each side, and at intermediate distances between these locations (to account for lateral variability). Fifteen transects, with five stations per transect, were established along the length of the reservoir. Water samples were collected at depths of 0, 2, 4, 6, and 10 m and at 5-m intervals thereafter to within 0.5 m of the bottom (to account for vertical variability). Water samples were analyzed for chlorophyll *a* and TP concentrations and for turbidity. During July the lake was thermally stratified, and the data were divided into three depth strata: the epilimnion, the metalimnion, and the hypolimnion.

71. A variance component analysis of the data was used to determine the contribution of each dimension to the overall variability. Table 9 presents the results of the analysis. Lateral variability was low and rarely contributed a major portion of the variance. Longitudinal variability, except in the hypolimnion, contributed most of the variability. Vertical variability within each depth stratum contributed a significant fraction of the total variability. Exceptions to this pattern (i.e., little lateral variability and significant longitudinal and vertical variability) may occur in wide and shallow reservoirs with relatively long residence times (i.e., low flushing rates). Exceptions may also occur in reservoirs with extremely short residence times (longitudinal patterns do not develop because of rapid flushing) and in unstratified reservoirs (vertical patterns do not develop because of mixing). Most reservoirs can be expected to exhibit significant longitudinal and vertical variability with lateral variability contributing a small and often insignificant portion of the overall variance.

72. The implications of this study should be obvious. The possible existence of significant vertical and longitudinal gradients in water quality must be accounted for in the development of a sampling design. Consideration of lateral variability is less important, and the remainder of the discussion of spatial patterns in water quality will focus on vertical patterns and longitudinal gradients.

73. Vertical patterns are established in response to vertical gradients in light and temperature. The establishment of thermal

stratification effectively divides the reservoir into three depth strata (Figure 6). The effective isolation of the hypolimnion can lead to

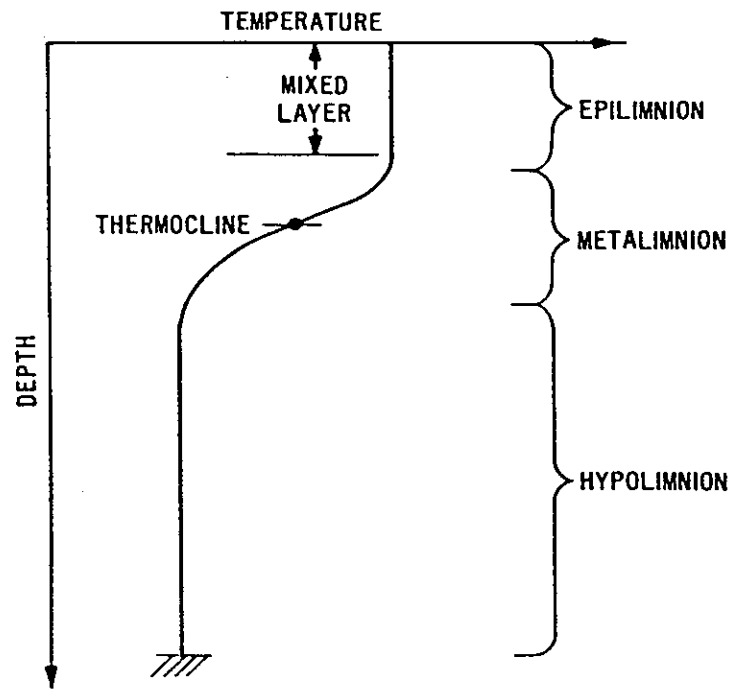


Figure 6. Typical pattern of thermal stratification and its ability to divide the reservoir into three vertical strata

oxygen depletion and the subsequent release of reduced constituents. The distribution of light with respect to depth is a major determinant of primary productivity, and this effect, in turn, influences the vertical distribution of plant nutrients. Figure 7 presents some typical vertical patterns and illustrates the influence of changes in light, temperature, and dissolved oxygen concentrations on the vertical distribution of chlorophyll *a*, soluble reactive phosphorus, and dissolved metals. For a detailed discussion of vertical stratification patterns, the reader can consult "Reservoir Water Quality" (OCE 1985), Wetzel (1975), and Rutner (1963).

74. The vertical patterns in water quality imply that a sampling design should be stratified with respect to depth. The vertical

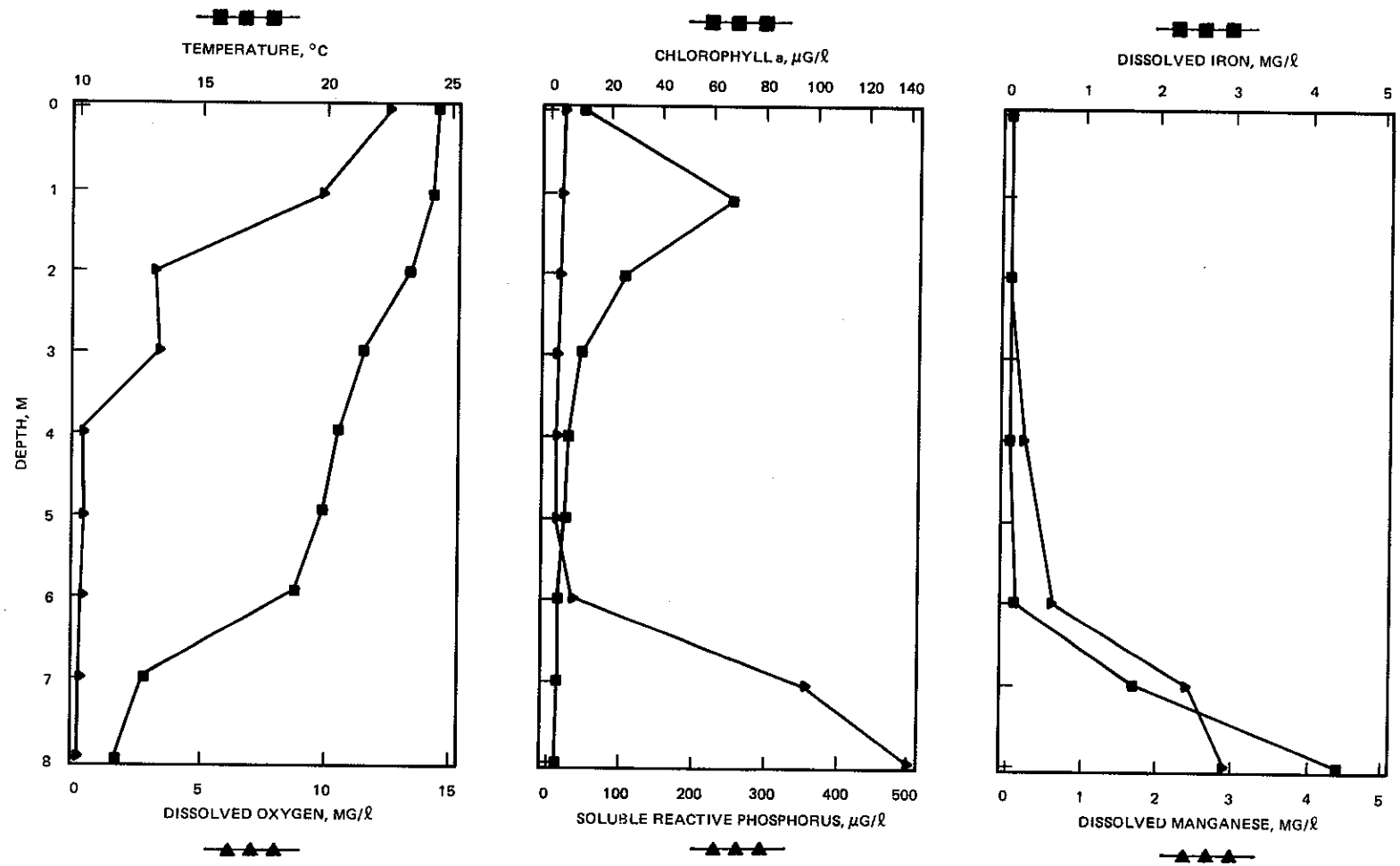


Figure 7. Vertical distributions of temperature, dissolved oxygen, chlorophyll *a*, soluble reactive phosphorus, dissolved iron, and dissolved manganese from Eau Galle Lake

distributions of light, temperature, and dissolved oxygen can be used to define depth strata. The strata that might be defined with these variables are listed in Table 10. The selection of strata or combinations of strata will depend on the specific objectives of the study.

Table 10

Depth Strata Defined by Vertical Patterns of Temperature, Light, and Dissolved Oxygen

Stratification Variable	Strata Defined
Temperature	epilimnion, metalimnion, hypolimnion
Light	photic, aphotic
Dissolved Oxygen	oxic, anoxic

75. Reservoirs can exhibit significant longitudinal gradients in physical, chemical, and biological parameters. The relatively long, dendritic morphometry, the location of the major inflow a considerable distance from the outflow, and the generally unidirectional flow from the headwaters to the dam foster the development of longitudinal gradients. The potential for longitudinal gradients is a complex function of hydrology and morphometry (Figure 8). Reservoirs with long residence times, high sedimentation, and a flow regime dominated by advective transport will have a greater potential for longitudinal gradients than reservoirs with short residence times, low sedimentation, and dispersive transport.

76. These gradients are continuous, but three characteristic zones can be identified (Thornton et al. 1980): a riverine zone, a zone of transition, and a lacustrine zone (Figure 9). The riverine zone can be characterized as a zone with high flow velocities, high concentrations of suspended matter and nutrients, and limited light availability. High concentrations of allochthonous organic matter may lead to a considerable oxygen demand, but anoxia is prevented from developing because

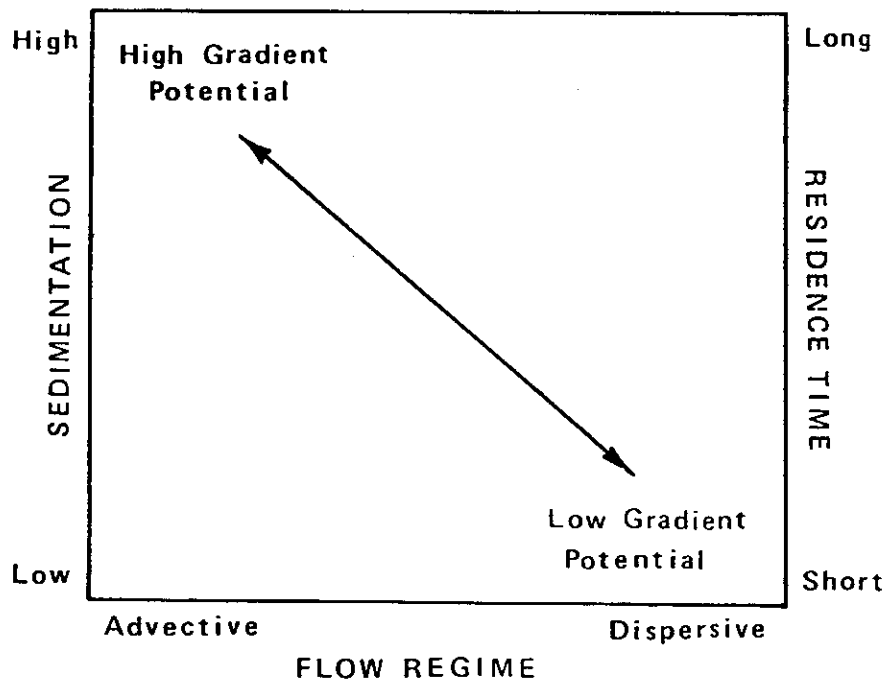


Figure 8. Potential for longitudinal gradients in reservoirs as function of sedimentation, flow regime, and residence time

the riverine zone is relatively shallow and well mixed. The zone of transition is characterized by reduced flow velocities resulting from the widening and deepening of the reservoir in this zone. Reduced flow velocities allow for greater sedimentation of suspended matter, and, as a result, light availability improves. Given the relatively high nutrient concentrations and the light availability, both phytoplankton productivity and biomass may peak in this zone. The characteristics of the lacustrine zone of the reservoir are similar to the conditions observed in natural lakes. Flow velocities are minimal, light availability is relatively high, and nutrient availability will most likely limit primary productivity. The deposition and subsequent decomposition of autochthonous organic matter in the hypolimnion of this zone may result in the depletion of oxygen in portions of the hypolimnion.

77. The presence and extent of these zones can readily be incorporated into a stratified sampling design. A combination of flow velocity and light availability profiles could be used to define the

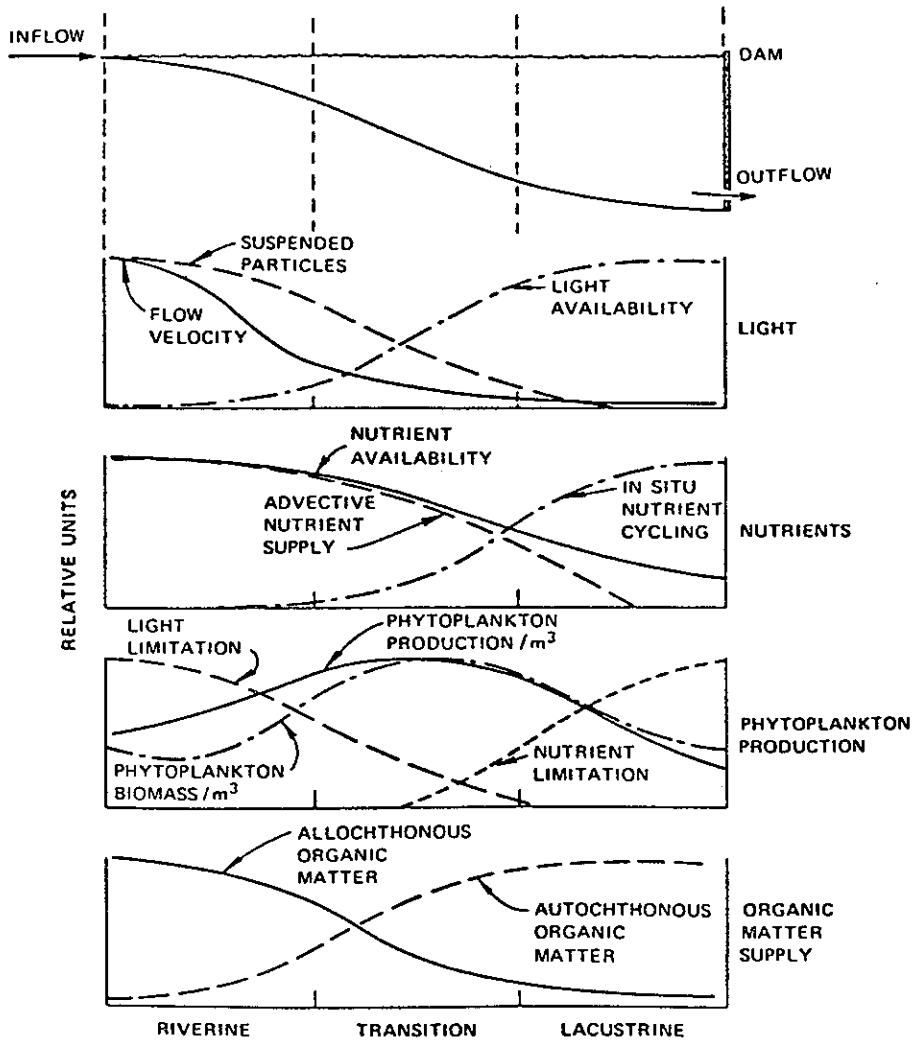


Figure 9. Longitudinal patterns in reservoir water quality (after OCE (1985))

longitudinal strata. The vertical strata discussed earlier could be combined with the longitudinal strata to form a two-dimensional stratification of the reservoir. A sampling design using both vertical and longitudinal stratification should account for the majority of the spatial variability in reservoir water quality.

Temporal Patterns

78. Just as spatial patterns in reservoir water quality are ultimately determined by the flow regime and the vertical and longitudinal changes in light and temperature, significant temporal patterns are determined by the temporal dynamics of the the flow regime and the seasonal changes in temperature and incident solar radiation. Temporal patterns of interest will be related primarily to changes in thermal stratification. The annual cycle of changes in thermal stratification is well documented in limnology texts, such as Wetzel (1975) or Rutner (1963). In dimictic lakes, the periods of mixing, spring and fall turnover, are preceded and followed by periods of stratification (i.e., winter and summer stratification). The duration and extent of these seasonal events can be used to stratify a sampling program temporally. In the warm monomictic reservoirs of the Southeast, the stratified period is preceded and followed by a period of winter circulation. In such a case, stratification of the temporal variability into only two strata may not be sufficient to account adequately for the temporal component of variability in reservoir water quality.

79. Temporal changes in the flow regime can aid in the temporal stratification of reservoir sampling designs. Figure 10 presents the seasonal dynamics of the mean daily flow into DeGray Lake contributed by its major tributary, the Caddo River. The flow regime can be divided into a high flow stratum (March - May) and a low or base flow stratum (July - February). The combination of thermal and flow stratification can be used to develop a design to account for a greater proportion of the temporal variability (Table 11).

80. The allocation of samples with respect to time or the sampling frequency can pose serious problems if a systematic design is used. A major assumption of statistical analyses is that the values of the sample are independent, meaning that each value in the target population has an equal probability of being chosen for the sample. In a systematic design, only the first sample value is randomly chosen; the remaining values are selected at constant intervals of time (see the

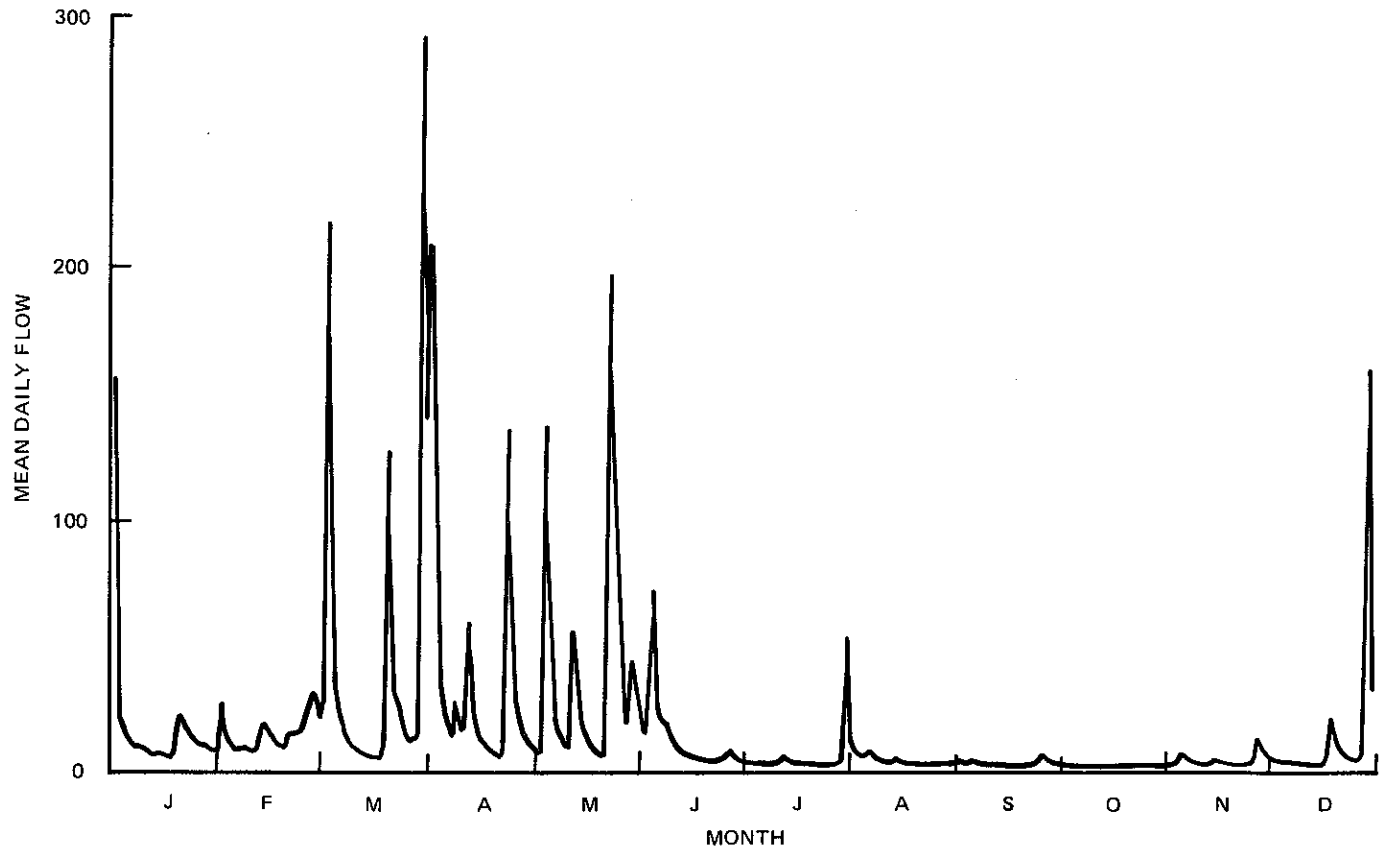


Figure 10. Mean daily flow (m^3/sec) into DeGray Lake from the Caddo River

Table 11

Temporal Strata Defined by Seasonal Changes in Flow Regime and Thermal Stratification

Stratum	Months
Winter circulation, high flow	March - May
Stratified	June - September
Winter circulation, low flow	October - February

section on systematic sampling in Part III). Such a design can result in the lack of independence if the successive sample values are correlated.

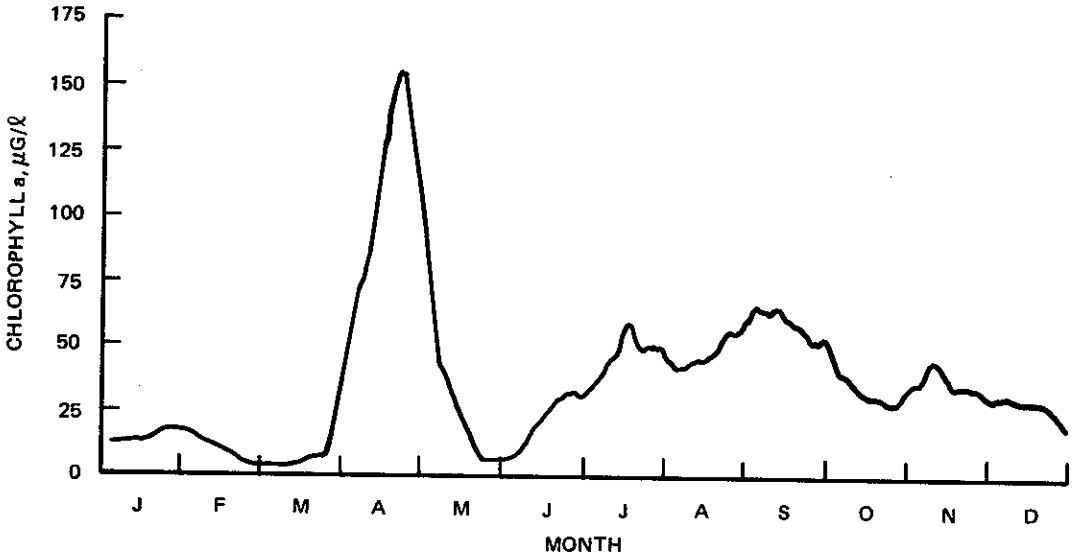
81. Serial correlation implies that the successive values are not independent and, in fact, successive values can be expressed as a function of the preceding values. Assume that daily samples for chlorophyll *a* were obtained from a reservoir (Figure 11a). From the data the correlation between the values can be calculated for

y_i and y_{i-1} where $i = 2, 3, 4, \ldots, k$

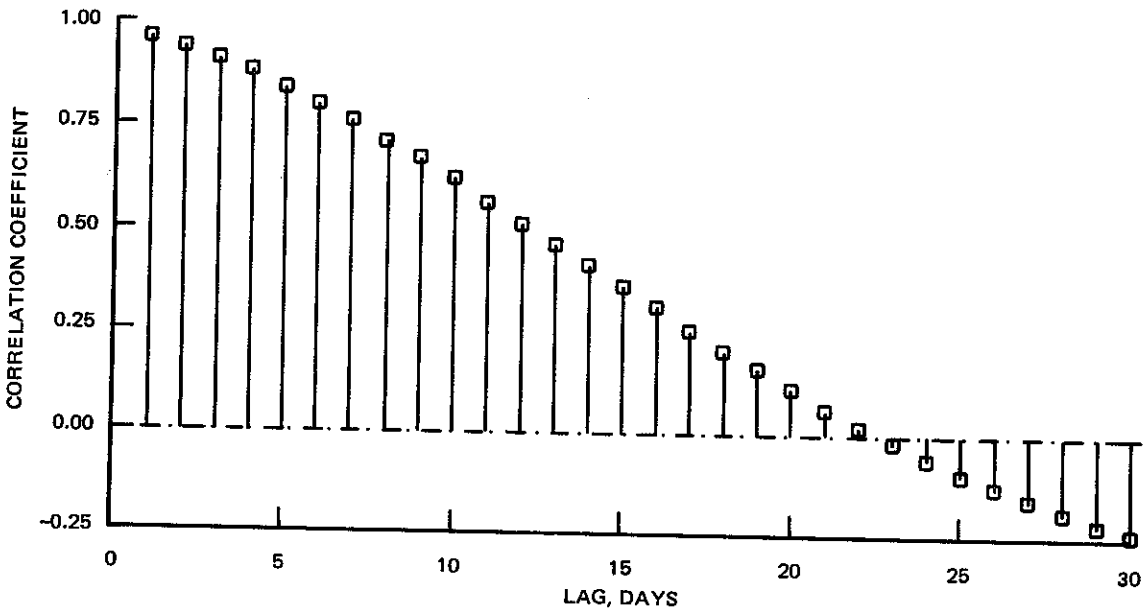
y_i and y_{i-2} where $i = 3, 4, 5, \ldots, k$

y_i and y_{i-n} where $i = n, n + 1, n + 2, \ldots, k$

Computing the correlation between a variable and the preceding values of the same variable produces autocorrelation coefficients that describe the dependence of the y_i values on the preceding values. Figure 11b presents the autocorrelation coefficients for the chlorophyll *a* data. Successive values of chlorophyll *a* are highly dependent on previous values. At a lag of 7 days, the autocorrelation coefficient is greater than 0.75. The term "lag" refers to the separation between successive observations.



a. Chlorophyll *a*



b. Autocorrelation coefficients

Figure 11. Seasonal dynamics of chlorophyll *a* and its autocorrelation coefficients

Therefore, a lag of 7 days implies successive observations of chlorophyll *a* at 7-day intervals.

82. The existence of autocorrelation or serial dependence in water quality data implies a very serious limitation to the use of systematic designs. At low sampling frequencies, less than one sample per month, serial dependence should not pose a problem. If higher sampling frequencies are to be used, a systematic design can be a liability. Systematic designs are commonly used because of their convenience, and this advantage can be retained by combining aspects of both random and systematic sampling. For example, assume that weekly samples are to be collected. The systematic nature of the design can be retained by sampling every week, but the day that the sample is taken can be chosen at random. This sampling design effectively treats each week as a separate stratum from which one unit is collected at random. This modification of the systematic design should result in an independent sample.

Interaction Between Spatial and Temporal Patterns

83. It should be intuitively obvious that spatial and temporal variability in reservoir water quality is not independent. Spatial patterns can be expected to change with time. The size or extent of both vertical and longitudinal strata varies with time. The size of the epilimnion will generally increase during the course of the stratified period while the hypolimnion will undergo a concomitant decrease in size. Longitudinal strata will also expand and contract in response to seasonal changes in the flow regime. During the period of high flow, the riverine zone will extend further downstream, displacing the transition zone downstream and shrinking the lacustrine zone. Under low flow conditions, the riverine zone may be very limited in extent, shifting the transition zone upstream and expanding the lacustrine zone.

84. A few examples from DeGray Lake will illustrate the interaction of spatial and temporal dynamics. The spatial distribution of chlorophyll *a* concentrations changed drastically with time (Figure 12). In May, a period of high flow, high concentrations of chlorophyll *a*

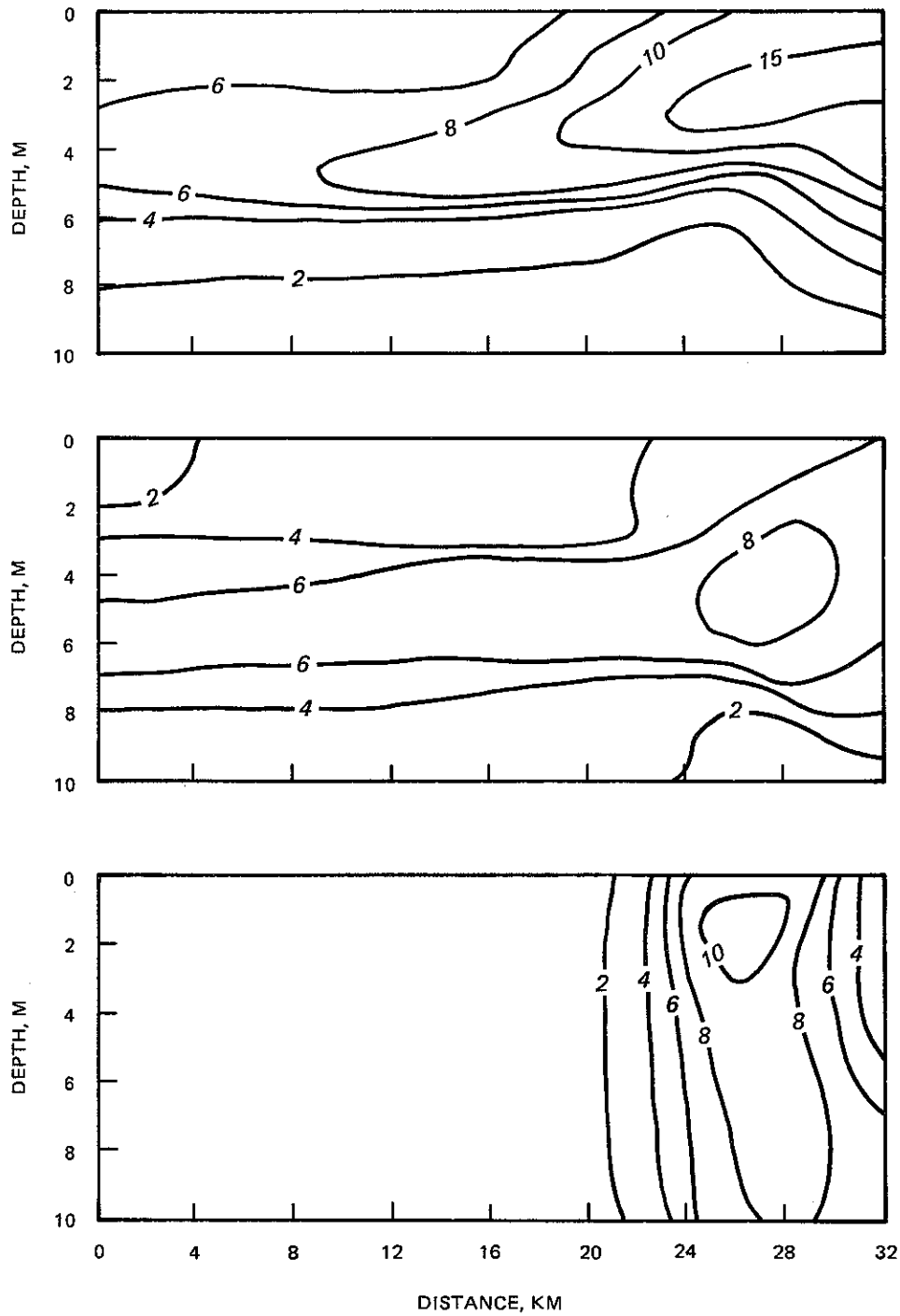


Figure 12. Longitudinal and vertical distributions of chlorophyll α in DeGray Lake during May, July, and October

extended much farther downstream than in July or October when flows were lower. Also, chlorophyll *a* was vertically stratified in May and July, but not in October. The spatial extent of anoxia changed with time, which influenced the spatial distribution of TP concentrations (Figure 13). The zone of anoxia increased in size and moved downstream from May to November. Phosphorus release from the sediments under anoxic conditions resulted in a zone of elevated phosphorus concentrations increasing in size and moving downstream during the same period. Longitudinal gradients in turbidity, TP, and chlorophyll *a* also changed with time (Figure 14). Changes in light, temperature, and flow regime altered both the steepness of the gradient and the position of the peak.

A Generalized Reservoir Sampling Design

85. Reservoir sampling designs must consider both spatial and temporal sources of variability when attempting to generate precise and cost-effective estimates of water quality. The spatial strata portrayed in Figure 15 can form the basis of a generalized reservoir sampling design. This design accounts for both vertical and longitudinal variability in water quality. The design can also account for temporal variability if the time-dependent changes in the size and position of the vertical and longitudinal strata are considered. This design will not be suitable for all water quality sampling programs on all reservoirs. Sampling designs must be tailored to meet the stated objectives within the site-specific conditions of a given reservoir. The generalized reservoir sampling design is offered as a conceptual starting point for the development of specific reservoir sampling designs.

86. The proposed generalized design can be thought of as a minimal design for sampling reservoir water quality. The term "minimal" refers to the fact that this design is the simplest that can be used to characterize reservoir-wide water quality conditions. Many water quality studies will require more complex designs with a greater number of longitudinal and/or vertical strata. Even in those cases where the minimal design is acceptable, an increase in the complexity of the design

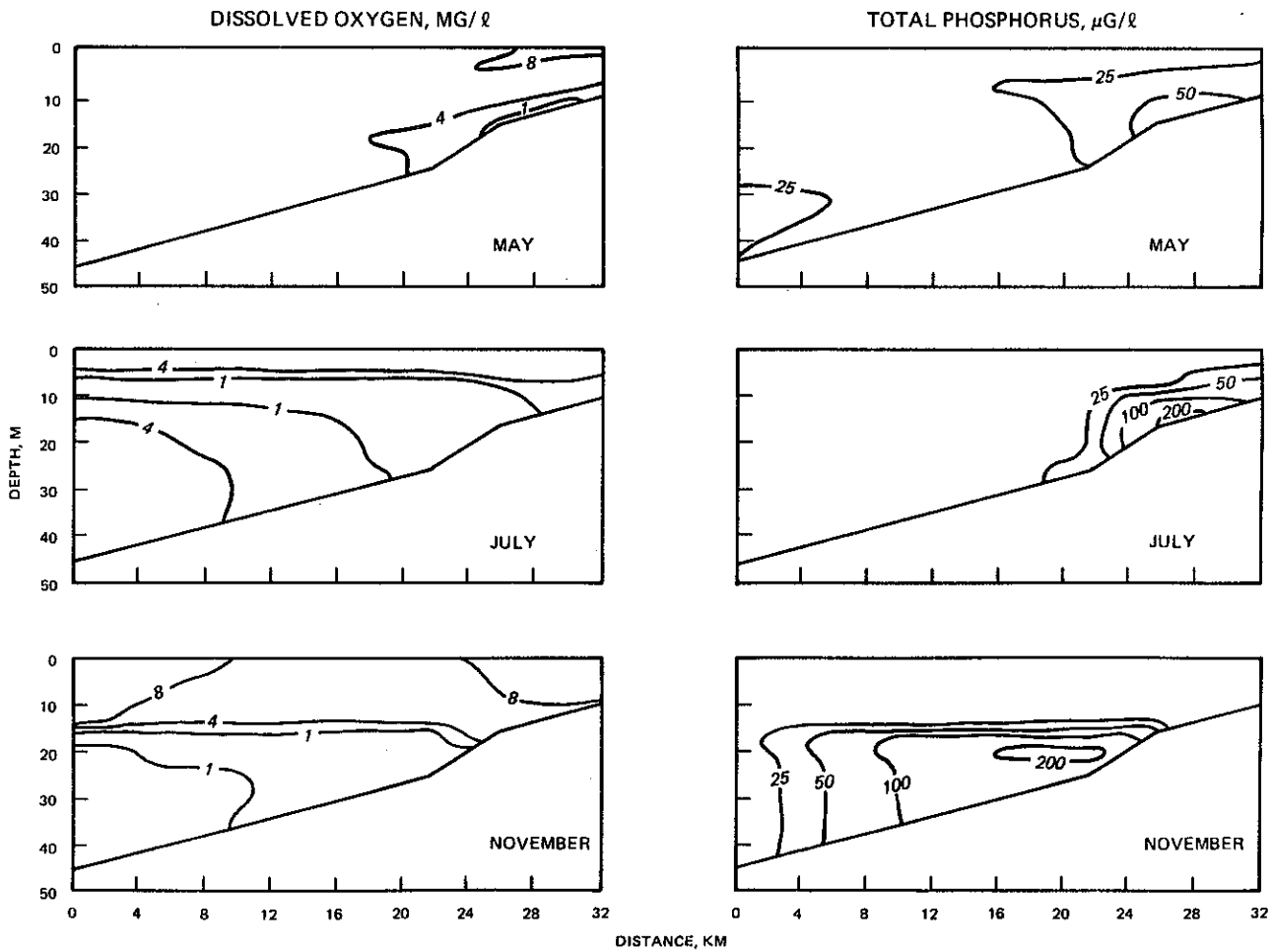


Figure 13. Longitudinal and vertical distributions of dissolved oxygen and total phosphorus in DeGray Lake during May, July, and November

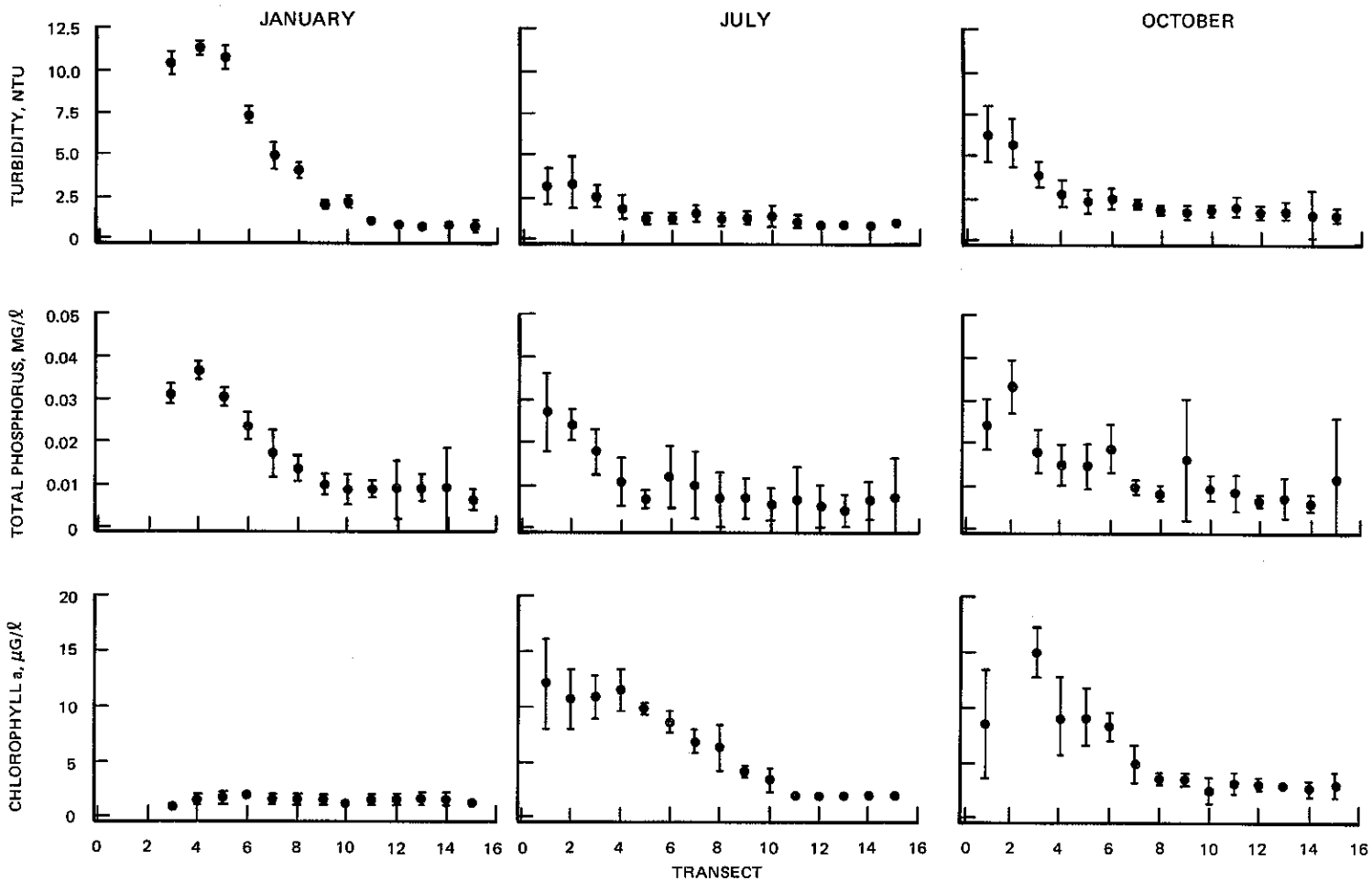


Figure 14. Longitudinal gradients in turbidity, total phosphorus, and chlorophyll *a* in DeGray Lake during January, July, and October (NTU = Nephelometric Turbidity Units)

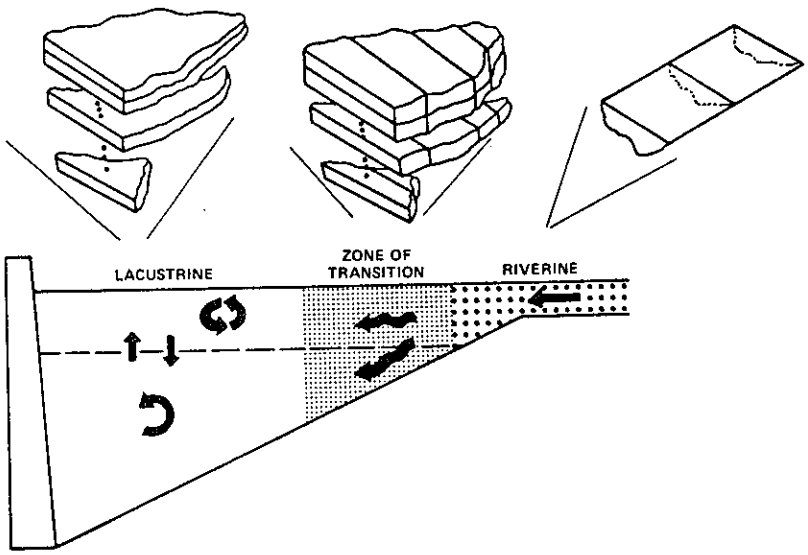


Figure 15. A generalized reservoir sampling design with both vertical and longitudinal stratification (after Thornton et al. (1980))

(if possible within manpower and funding constraints) will result in increased precision or reduced error in the water quality estimates. The minimal design considers three vertical strata (the epilimnion, metalimnion, and hypolimnion) and three longitudinal strata (the riverine, transition, and the lacustrine zones) for a total of nine cells (i.e., combinations of strata, 3 vertical v 3 longitudinal = 9 cells).

87. Strata boundaries can be defined by either a pilot study or an examination of existing data. Most often, water temperature and dissolved oxygen concentrations can be used to define the vertical strata. A series of vertical profiles taken along the major axis of the reservoir will identify the epilimnion, metalimnion, and hypolimnion. Longitudinal strata will be best defined by the longitudinal gradients in phosphorus, chlorophyll a and turbidity or suspended solids. Once the strata boundaries are defined, the total sample size can be determined using the formulae presented earlier. Given total sample size, the samples can be allocated to the strata equally (equal allocation), as a

function of their relative size (proportional allocation), or as a function of their variance (optimal allocation).

PART V: EVALUATION OF SAMPLING DESIGN

Introduction

88. Once the data have been collected from a sampling program, the data analysis phase of the program can be initiated. There are fundamentally two types of data analysis: (a) data analysis directed at characterizing the target population and (b) data analysis with the objective of evaluating the effectiveness of the sampling design. The first type of analysis will not be considered in this work, but an introduction to the subject with an emphasis on reservoir water quality data analysis can be found in Gaugush (1986). The second type of data analysis is rarely considered, and methods for the evaluation of sampling design will follow.

89. There are two important reasons for evaluating a sampling design. First, if the sampling design is to be used again, as in a long-term monitoring program, it is essential to determine if the design can be modified or improved to increase the precision of the estimates or reduce the cost of sampling. It may be possible to reduce the sample size of a particular variable with relatively low variance, or a variable may be dropped from the sampling program if it exhibits little variance with respect to the important water quality concerns. Second, even if the sampling will not be repeated, it is worthwhile to evaluate the effectiveness of the design to aid the development of future sampling programs.

90. The remainder of this report will consider the following three questions concerning sampling design and the three corresponding methods for addressing the questions:

a. How well does the sampling design "explain" or account for the observed variance?

Method: Variance Component Analysis

b. Are there redundancies in the data that can be removed to improve the sampling design?

Method: Cluster Analysis

c. How can a sample design be modified to improve the precision of the estimates?

Method: Error Analysis

Variance Component Analysis

91. Variance component analysis is a technique for quantifying the sources of variability in the data resulting from a given sampling design. The data analysis of the spatial variability in DeGray Lake presented in Part IV is an example of a variance component analysis. The analysis results in the determination of each design component's contribution to the overall variance. Based on the results of a variance component analysis, sampling effort in a given stratum can be reduced or even eliminated. In the DeGray example, it was shown that lateral variability contributed very little to the overall variability. In future sampling, lateral sampling could be eliminated to reduce sampling effort without seriously affecting the precision of the estimates.

92. The computational details of variance component analysis are beyond the scope of this work, but the reader can consult Winer (1971) for an excellent treatment of the method.

Cluster Analysis

93. Cluster analysis is a classification technique that may be used to group or identify similar objects. In the more typical data analysis situation, cluster analysis may be used to group a set of reservoirs according to their trophic state or by the composition of their phytoplankton. The use of cluster analysis in a typical data analysis mode can be found in Gaugush (1986). In the evaluation of a sampling design, cluster analysis can be used to identify redundancies in the design. For example, if a stratified random sampling design

requires the collection samples from 12 strata, cluster analysis may be able to identify groups of strata that can be combined to reduce sampling effort. The remainder of this section will provide a conceptual introduction to the method and the necessary background to interpret the results of a cluster analysis in a sampling design framework. Computational details of cluster analysis can be found in Gaugush (1986).

94. Cluster analysis is a multivariate technique that can be used to identify the possibility of reducing sampling effort. The term "multivariate" refers to the fact that cluster analysis can operate on a set of variables. It is the multivariate nature of cluster analysis that gives it a considerable advantage over variance component analysis as a method for the evaluation of sampling design. Whereas variance component analysis can suggest a reduction of effort based on one variable at a time, cluster analysis can use all of the variables sampled to suggest an improvement of the design. Cluster analysis has a second advantage over variance component analysis in that it identifies specific samples that can be combined in future sampling to reduce sampling effort.

95. Cluster analysis may identify groups of similar observations (i.e., the clusters) based on a number of criteria, but in most applications the criterion is either the correlation between groups of observations or the similarity of their means. At the start of the cluster analysis, each group of variables comprises its own cluster. At each step of the analysis, clusters are joined based on their similarity until all of the observations form a single cluster. The objective of the analysis is to find how many clusters between one cluster for each observation and one cluster for all observations best describe the data. An indicator referred to as the cubic clustering criterion (CCC) is probably the best available method for identifying the number of clusters within the data set (Gaugush 1986). During the cluster analysis, a CCC is calculated for each set of clusters between n clusters (where n is the number of observations) and one cluster. A local peak value of the CCC identifies the number of clusters that may define an acceptable grouping of the data.

96. As an example of the usefulness of cluster analysis, consider the following sampling design. A stratified random sampling program was used to estimate the annual mean concentrations of chlorophyll a, TP, TN, total organic carbon, and dissolved silica. Strata were defined both vertically and temporally. The epilimnion, metalimnion, and hypolimnion were defined as the vertical strata, and the months of the year were chosen as the temporal strata. One randomly selected sample was taken from each combination of depth and time strata for a total sample size of 36.

97. Cluster analysis can identify the number of clusters between 1 (i.e., all observations are similar) and 36 (i.e., all observations are dissimilar) that best describes the data. It should be obvious that if the analysis finds a number of clusters less than 36 that can adequately describe the data, a reduction in future sampling effort can result. For the sake of the example, assume that the CCC has identified eight clusters as an acceptable grouping of the data. The results of the clustering can be depicted graphically using a dendrogram or tree diagram (Figure 16). Along the bottom of the figure, the 36 strata sampled are listed and the y-axis of the diagram is the number of clusters. The clusters identified by the CCC are also numbered from 1 to 8 from left to right.

98. By an inspection of which strata fall into particular clusters, the clusters can serve to define new strata (Table 12). The clusters that result from the analysis appear to fall into logical groupings when one considers typical spatial and temporal patterns in water quality. The strata comprising the winter months (November-February) all fall into a single cluster. This cluster includes all depth strata during these months, which is intuitively reasonable because of the expected lack of significant vertical stratification during this period. The clusters for the summer (July-September) indicate that depth strata are significant during the thermally stratified period.

99. The results of the cluster analysis indicate that it may be possible to reduce the number of strata from 36 to 8. With this considerable reduction, it may also be possible to increase the total sample

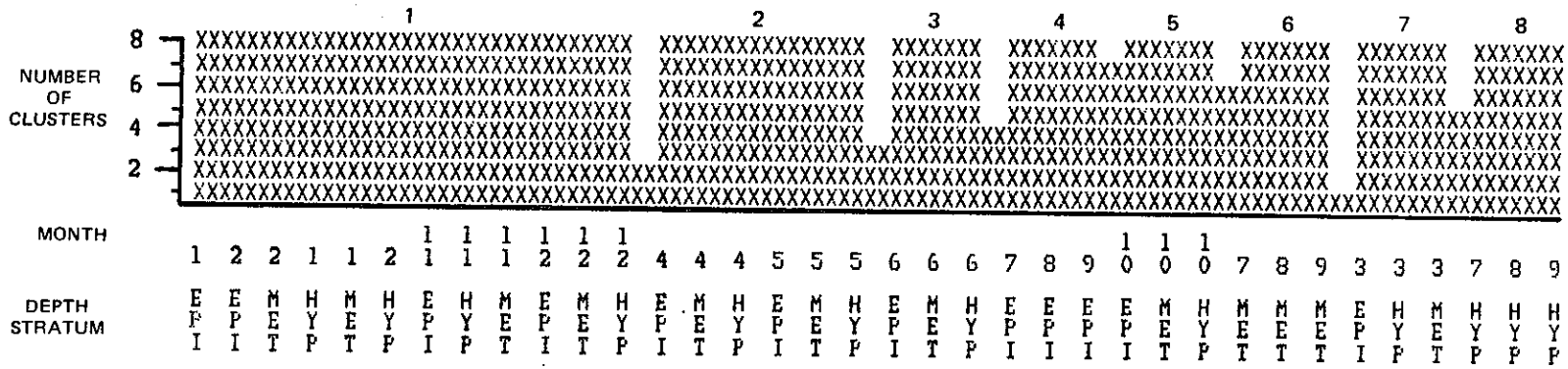


Figure 16. Dendrogram for the cluster analysis example

Table 12

New Strata Defined by the Results of the Cluster Analysis

Cluster	Existing Strata: Months	Existing Strata: Depths	New Strata
1	1, 2, 11, 12	epi, met, hyp*	Nov-Feb, no vertical strata
2	4, 5	epi, met, hyp	Apr-May, no vertical strata
3	6	epi, met, hyp	Jun, no vertical strata
4	7, 8, 9	epi	Jul-Sep, epilimnion
5	10	epi, met, hyp	Oct, no vertical strata
6	7, 8, 9	met	Jul-Sep, metalimnion
7	3	epi, met, hyp	Mar, no vertical strata
8	7, 8, 9	hyp	Jul-Sep, hypolimnion

* Epi = epilimnion; met = metalimnion; hyp = hypolimnion

size, allocate more than one sample per stratum, and increase the precision of the estimates without increasing costs. After a cluster analysis is performed, it would be important to develop a decision matrix for the sampling design suggested by the cluster analysis. Only through the use of the matrix can one be assured of an improvement in the sampling design.

Error Analysis

100. Error analysis is a method for the improvement of a sampling design that uses the observed distribution of variance to redefine a sampling program. The result of error analysis is a redistribution of samples to the existing strata to produce the minimum variance about the mean. Error analysis can be applied to the data of a stratified sampling design or to the data from a simple random or a systematic sample that has been subjected to poststratification (i.e., defining strata a posteriori).

101. To illustrate the use of error analysis, consider the following stratified random sampling design. The objective of the study was to estimate the annual mean epilimnetic chlorophyll *a* concentration. Based on limited previous data, a seasonal stratification was used with the strata defined by the durations of spring mixing, summer stratification, fall mixing, and winter stratification. A proportional allocation of 25 samples was used because of the limited data concerning variance within the specified strata. The results of this study are presented in Table 13.

Table 13

Results Derived from Sampling Epilimnetic Chlorophyll Concentrations*

Stratum	n	wt	$\bar{y}$	s^2	$s^2_{\bar{y}}$
Spring	4	0.16	73.30	4852.65	1213.16
Summer	9	0.36	41.81	341.80	37.98
Fall	4	0.16	31.93	45.08	11.27
Winter	8	0.32	13.93	93.41	11.68

* $\bar{y}_{st} = 36.35$; $s^2_{\bar{y}_{st}} = 37.47$

102. Error analysis considers the contribution of each stratum's error variance (variance about the mean $s^2_{\bar{y}_h}$) to the error variance about the stratified sample mean ($s^2_{\bar{y}_{st}}$). The error variance contributed by a given stratum is

$$w_h^2 s^2_{\bar{y}_h} \tag{53}$$

and the percentage contribution to the stratified sample error variance is

$$V_h\% = 100 \frac{w_h^2 s^2_{\bar{y}_h}}{s^2_{\bar{y}_{st}}} \tag{54}$$

where $V_h\%$ = distribution of variance. Error analysis also considers the distribution of sampling effort among the strata. The percentage contribution of each n_h to the n_{st} is given by

$$N\ \% = 100 \frac{n_h}{n_{st}} \tag{55}$$

where $N_h\%$ = distribution of samples. Based on the observed distribution of variance, error analysis calculates the optimal distribution of the samples to the strata from the following:

$$OPT_h\% = \frac{100\ w_h \left(n_h s^2_{\bar{y}_h}\right)^{1/2}}{\sum_{h=1}^{L} w_h \left(n_h s^2_{\bar{y}_h}\right)^{1/2}} \tag{56}$$

where $OPT_h\%$ = optimal distribution of samples. This can be simplified because

$$s_h^2 = n_h s^2_{\bar{y}_h} \tag{57}$$

which gives

$$OPT_h\% = \frac{100\ w_h s_h}{\sum_{h=1}^{L} w_h s_h} \tag{58}$$

The $N_h\%$, $V_h\%$, and $OPT_h\%$ for the example data set are presented in Table 14.

Table 14

Distribution of Sampling Effort, Error Variance, and Optimal Sampling Effort for the Chlorophyll Data of Table 13

Stratum	$N_h\%$	$V_h\%$	$OPT_h\%$
Spring	16.0	82.9	50.8
Summer	36.0	13.1	30.3
Fall	16.0	0.8	4.9
Winter	32.0	3.2	14.1

103. An examination of the disparity between the $N_h\%$ values and the $V_h\%$ values indicates there is considerable room for improvement of the sampling design. For example, the spring stratum was allocated only 16 percent of the samples, and yet it contributed over 80 percent of the error variance. The $OPT_h\%$ values account for both the stratum size and its contribution to the error variance.

104. Error analysis can also estimate the error variance of the stratified sample mean if future samples are allocated according to the optimal distribution. The error variance of the optimal design can be estimated by

$$VOPT = \sum_{h=1}^{L} \left[w_h^2 s_{\bar{y}_h}^2 \left(\frac{N_h\%}{OPT_h\%} \right) \right] \tag{59}$$

For the example data set

$$VOPT = (0.16)^2(1213.16)(16.0/50.8) + (0.36)^2\ (37.98)(36.0/30.3)$$

$$+ (0.16)^2\ (11.27)(16.0/4.9) + (0.32)^2\ (11.68)(32.0/14.1)$$

$$VOPT = 9.78 + 5.85 + 0.94 + 2.71$$

$$VOPT = 19.28 \tag{60}$$

The optimal design results in an expected error variance of 19.28, whereas the existing design resulted in an error variance of 37.47. If the optimal design were to be adopted for future sampling, the error variance would be expected to be nearly one half the error variance resulting from the original design.

105. From the above, it is apparent that error analysis can lead to a significant reduction in error variance (i.e., an increase in precision) without an increase in cost. The only assumption that is necessary is that the distribution of variance remains fairly constant over time. That is, if the optimal design is adopted, the distribution of strata variances must be similar to those observed in the past for the estimated reduction in error variance to occur. Since this assumption is made in the development of all sampling designs, it does not limit the usefulness of error analysis as an evaluation method.

REFERENCES

American Public Health Association. 1980. Standard Methods for the Examination of Water and Wastewater, 15th ed., American Public Health Association, New York.

Brown, E., Skougstad, M. W., and Fishman, M. J. 1970. Methods for Collection and Analysis of Water Samples for Dissolved Minerals and Gases, US Geological Survey, Reston, Va.

Cochran, W. G. 1977. Sampling Techniques, John Wiley and Sons, New York.

Gaugush, R. F., tech. ed. 1986. "Statistical Methods for Reservoir Water Quality Investigations," Instruction Report E-86-2, US Army Engineer Waterways Experiment Station, Vicksburg, Miss.

Greeson, P. E., et al. 1977. Methods for Collection and Analysis of Aquatic Biological and Microbiological Samples, US Geological Survey, Reston, Va.

Office, Chief of Engineers (OCE). 1985. "Reservoir Water Quality," Engineer Manual 1110-2- , Washington, DC.

Plumb, R. H. 1981. "Procedures for Handling and Chemical Analysis of Sediment and Water Samples," Technical Report EPA/CE-81-1, prepared by Great Lakes Laboratory, State University College at Buffalo, Buffalo, N. Y., for the US Environmental Protection Agency/Corps of Engineers Technical Committee on Criteria for Dredged and Fill Material, published by the US Army Engineer Waterways Experiment Station, Vicksburg, Miss.

Rutner, F. 1963. Fundamentals of Limnology, University of Toronto Press, Toronto, Canada.

Skougstad, M. W., et al., eds. 1979. Methods for Determination of Inorganic Substances in Water and Fluvial Sediments, US Geological Survey, Reston, Va.

Thornton, K. W., et al. 1980. "Reservoir Sedimentation and Water Quality - A Heuristic Model," Proceedings of a Symposium on Surface-Water Impoundments, American Society of Civil Engineers, Minneapolis, Minn., Vol I, pp 654-661.

_____. 1982. "Reservoir Water Quality Sampling Design," Water Reservoir Bulletin, Vol 18: pp 471-480.

US Environmental Protection Agency. 1982. "Handbook for Sampling and Sample Preservation of Water and Wastewater," Technical Report EPA-600/4-82-029, Environmental Monitoring and Support Laboratory, Office of Research and Development, Cincinnati, Ohio.

US Geological Survey (USGS). 1972. Recommended Methods for Water Data Acquisition, US Department of the Interior, Washington, DC.

_____. 1977. National Handbook of Recommended Methods for Water-Data Acquisition, US Department of the Interior, Reston, Va.

Waide, J. B. 1986. "General Guidelines for Monitoring Contaminants in Reservoirs," Instruction Report E-86-1, US Army Engineer Waterways Experiment Station, Vicksburg, Miss.

Wetzel, R. G. 1975. Limnology, W. B. Saunders, Philadelphia.

Winer, B. J. 1971. Statistical Principles in Experimental Design, McGraw-Hill, New York.

BIBLIOGRAPHY

Benjamin, J. R., and Cornell, C. A. 1970. Probability, Statistics, and Decision for Civil Engineers, McGraw-Hill, New York.

Blum, J. R., and Rosenblatt, J. I. 1972. Probability and Statistics, W. B. Saunders, Philadelphia.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. 1978. Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley and Sons, New York.

Cochran, W. G., and Cox, G. M. 1957. Experimental Designs, John Wiley and Sons, New York.

Everett, L. G., and Schmidt, K. D., eds. 1979. Establishment of Water Quality Monitoring Programs, American Water Resources Association, Minneapolis.

Green, R. H. 1979. Sampling Design and Statistical Methods for Environmental Biologists, John Wiley and Sons, New York.

Kennedy, R. H., Thornton, K. W., and Gunkel, R. C. 1982. "The Establishment of Water Quality Gradients in Reservoirs," Can. Wat. Res. Journal, Vol 7, pp 71-87.

Snedecor, G. W., and Cochran, W. G. 1967. Statistical Methods, Iowa State University Press, Ames, Iowa.

Sokal, R. R., and Rohlf, F. J. 1969. Biometry, W. H. Freeman, San Francisco.

Steel, R. G. D., and Torrie, J. H. 1980. Principles and Procedures in Statistics, 2d ed., McGraw-Hill, New York.

Stuart, A. 1962. Basic Ideas of Scientific Sampling, Hafner Publishing, New York.

Tukey, J. W. 1977. Exploratory Data Analysis, Addison-Wesley, Reading, Mass.

Wine, R. L. 1964. Statistics for Scientists and Engineers, Prentice-Hall, Englewood Cliffs, N. J.

Wonnacott, T. H., and Wonnacott, R. J. 1972. Introductory Statistics, John Wiley and Sons, New York.

Zar, J. H. 1974. Biostatistical Analysis, Prentice-Hall, Englewood Cliffs, N. J.

APPENDIX A: SAMPLING DESIGN FORMULAE

1. SAMPLE MEAN (p 8)

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i \tag{A1}$$

where

$\bar{y}$ = an estimate of the population mean

n = sample size

y_i = individual observations

2. SAMPLE VARIANCE (p 8)

$$s^2 = (n - 1)^{-1} \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 \tag{A2}$$

s^2 = an estimate of the population mean

3. SAMPLE VARIANCE OF SAMPLE MEANS (p 10)

$$s^2_{\bar{y}} = \frac{s^2}{n} \tag{A3}$$

4. STANDARD ERROR (p 10)

$$s_{\bar{y}} = \left(s^2_{\bar{y}}\right)^{1/2} = \left(\frac{s^2}{n}\right)^{1/2} \tag{A4}$$

5. CONFIDENCE LIMITS FOR THE SAMPLE MEAN (p 11)

$$L_L = \bar{y} - ts_{\bar{y}}$$

$$L_U = \bar{y} + ts_{\bar{y}} \qquad \text{(A5)}$$

where

L_L = lower confidence limit

L_U = upper confidence limit

t = Student's t value for $n - 1$ degrees of freedom and a specified probability

6. SAMPLE SIZE FOR A SIMPLE RANDOM SAMPLE (p 15)

$$n = \frac{t^2 s^2}{(r\bar{y})^2} \qquad \text{(A6)}$$

where r = the desired precision .

7. SAMPLE MEAN FOR STRATUM h (p 25)

$$\bar{y}_h = n_h^{-1} \sum_{i=1}^{n_h} y_{hi} \qquad \text{(A7)}$$

where

n_h = sample size for stratum h

y_{hi} = value of the ith unit in stratum h

8. SAMPLE VARIANCE FOR STRATUM h (p 25)

$$s_h^2 = \left(n_h - 1\right)^{-1} \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h\right)^2 \tag{A8}$$

9. STRATIFIED SAMPLE MEAN (p 26)

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h \tag{A9}$$

where

$w_h = N_h/N$, stratum weight

N = total number of units in the target population

N_h = number of units in stratum h

10. STRATIFIED SAMPLE VARIANCE (p 26)

$$s_{st}^2 = \sum_{h=1}^{L} \left(w_h^2\right)\left(s_h^2\right) \tag{A10}$$

11. STRATIFIED SAMPLE STANDARD ERROR OF THE MEAN (p 27)

$$s_{\left(\bar{y}_{st}\right)} = \left[\sum_{h=1}^{L} \frac{\left(w_h^2\right)\left(s_h^2\right)}{n_h}\right]^{1/2} \tag{A11}$$

12. STRATIFIED SAMPLE SIZE (p 28)

$$n_{st} = \frac{t^2 s_{st}^2}{\left(r\bar{y}_{st}\right)^2} \tag{A12}$$

or

$$n_{st} = \frac{t^2 \sum_{h=1}^{L} w_h^2 s_h^2}{\left(r \sum_{h=1}^{L} w_h \bar{y}_h\right)^2} \tag{A13}$$

13. STRATUM SAMPLE SIZE - EQUAL ALLOCATION (p 29)

$$n_h = \frac{n_{st}}{L} \tag{A14}$$

14. STRATUM SAMPLE SIZE - PROPORTIONAL ALLOCATION (p 29)

$$n_h = w_h n_{st} \tag{A15}$$

15. LINEAR COST FUNCTION (p 29)

$$C = C_0 + \sum_{h=1}^{L} C_h n_h \tag{A16}$$

where

C = total cost

C_0 = fixed cost or overhead

C_h = cost per sample in stratum h

16. STRATUM SAMPLE SIZE - OPTIMAL ALLOCATION (p 30)

a. Costs differ among strata

$$n_h = n_{st} \frac{\dfrac{w_h s_h}{\sqrt{C_h}}}{\sum_{h=1}^{L} \left(\dfrac{w_h s_h}{C_h}\right)} \tag{A17}$$

b. Costs do not differ among strata (p 30)

$$n_h = n_{st} \frac{w_h s_h}{\sum_{h=1}^{L} (w_h s_h)} \tag{A18}$$

17. SYSTEMATIC SAMPLE MEAN (p 37)

$$\bar{y} = n^{-1} \sum_{k=1}^{n} y_k \tag{A19}$$

18. SYSTEMATIC SAMPLE VARIANCE (p 37)

$$s^2 = (n - 1)^{-1} \sum_{k=1}^{n} \left(y_k - \bar{y}\right)^2 \tag{A20}$$

19. SYSTEMATIC SAMPLE STANDARD ERROR (p 37)

$$s_{\bar{y}} = \left(\frac{s^2}{n}\right)^{1/2} \tag{A21}$$

20. ERROR ANALYSIS (pp 74-76)

a. Percentage contribution of stratum h to overall error variance

$$V_h\% = 100 \frac{w_h^2 s^2_{\bar{y}_h}}{s^2_{\bar{y}_{st}}} \tag{A22}$$

b. Percentage contribution of stratum h to overall sample size

$$N_h\% = 100 \frac{n_h}{n_{st}} \tag{A23}$$

c. Optimal percentage contribution of stratum h to overall sample size

$$OPT_h\% = \frac{100 \quad w_h\left(n_h s^2_{\bar{y}_h}\right)^{1/2}}{\sum_{h=1}^{L} w_h\left(n_h s^2_{\bar{y}_h}\right)^{1/2}}$$

or (A24)

$$OPT_h\% = \frac{100 \; w_h s_h}{\sum_{h=1}^{L} w_h s_h}$$

d. Error variance of the optimal design

$$VOPT = \sum_{h=1}^{L} \left[w_h^2 s_{\bar{y}_h}^2 \left(\frac{N_h\%}{OPT_h\%} \right) \right] \tag{A25}$$