STATISTICAL METHODS FOR STUDYING BIOTIC RESPONSES TO PERSISTENT TOXIC SUBSTANCES

by

S.R. Esterby

NWRI CONTRIBUTION #85-106

National Water Research Institute
Canada Centre for Inland Waters
867 Lakeshore Road, P.O. Box 5050
Burlington, Ontario, Canada L7R 4A6

PERSPECTIVE-GESTION

Les expériences en laboratoire et les études sur le terrain sont nécessaires pour évaluer les effets que peuvent avoir les substances toxiques persistantes sur la santé des colonies aquatiques parce que, même si le rapport de cause à effet peut être établi en laboratoire, ces expériences ne peuvent reproduire tous les aspects complexes du milieu aquatique. C'est sous cet angle qu'on a étudié le rôle des statistiques, en tenant compte des différences qui existent entre les méthodes statistiques et la validité des résultats des expériences en laboratoire et des études sur le terrain. On met l'accent sur la continuité des étapes d'une étude, de sa conception jusqu'à l'analyse de ses résultats, et sur le fait qu'il faut utiliser les méthodes statistiques pour concevoir l'étude en question. On donne des exemples tirés d'études sur le terrain portant sur les invertébrés benthiques pour expliquer comment utiliser l'échantillonnage aléatoire dans le cas de ces dernières. On donne également des exemples de l'analyse de données recueillies dans le cadre d'études expérimentales, entre autres, des analyses partielles des données de l'étude 211 de la DEA publiées en 1983 et d'une étude menée en 1977 par le Laboratoire de biolimnologie des Grands Lacs.

RÉSUMÉ

On discute des applications des méthodes statistiques sous l'angle du rôle complémentaire des expériences en laboratoire et des études sur le terrain en établissant un rapport de cause à effet et en expliquant leur incidence sur l'environnement. On met l'accent sur la continuité des étapes d'une étude, de sa conception jusqu'à l'analyse de ses résultats, et sur le fait qu'il est important d'utiliser les méthodes statistiques pour concevoir l'étude en question. On étudie les techniques de conception et les méthodes d'analyse dans trois genres d'études, en expliquant comment le degré de contrôle influe sur les méthodes statistiques et la validité des conclusions. On utilise des exemples tirés d'études sur le terrain portant sur les invertébrés benthiques pour démontrer l'utilisation de l'échantillonnage aléatoire dans le cadre d'études sur le terrain et pour offrir certaines méthodes d'analyse statistique qui conviennent aux études expérimentales. On n'a pas essayé de passer en revue ce sujet très vaste. On s'est plutôt contenté de déterminer les principes statistiques qui sont utilisés dans de nombreuses études et d'en discuter.

STATISTICAL METHODS FOR STUDYING BIOTIC RESPONSES TO PERSISTENT TOXIC SUBSTANCES

S.R. Esterby

Aquatic Ecology Division

National Water Research Institute

Burlington, Ontario, Canada

TABLE OF CONTENTS

		Page
1	INTRODUCTION	2
2	STUDY DESIGN	3
	2.1 Experiment, Field Sampling Program and Observational Study Defined	4
	2.2 The Steps in Planning a Study	5
	2.3 Comparison of the Experiment and the Field Study	6
	2.4 Further Considerations: Replication, Hypotheses, Efficiency	9
3	METHODS OF ANALYSIS	11
	3.1 Experiments	12
	3.2 Field Studies	13
	3.3 Checking the Model	15
4	EXAMPLES FROM BENTHIC STUDIES	17
	4.1 Quantiative Study Design	17
	4.2 Design for Estimation of Population Mean	18
	4.3 Design for Comparison of Strata	20
	4.4 Separation of Natural and Contaminant-Induced Variability	21
	4.5 Examples of Non-Parametric and Cluster Analysis	22

TABLE OF CONTENTS cont'd.

		Page
4.6	Contaminants in Benthic Fauna and Sediments	25
4.7	Major Benthic Taxa in Lake Ontario: Density versus Depth	26
5	DISCUSSION	28
REFERENCES		29

ABSTRACT

The applications for statistical methods are discussed in the context of the complementary roles of laboratory experiments and field studies in establishing cause and effect and explaining environmental significance. The continuity of the steps from the design of the study to the analysis of the results and the importance of the use of statistical methods in study design are emphasized. The techniques of design and methods of analysis are discussed for three types of studies, showing how the degree of control affects the statistical methods and the strength of the conclusions. Examples drawn from field studies of benthic invertebrates are used to illustrate the use of random sampling in field studies and to give some methods of statistical analysis suitable for observational studies. A review of this very broad topic was not attempted. Rather, statistical principles, of use in many types of studies, were identified and discussed.

1 INTRODUCTION

The scope for the application of statistical methods in the study of the effects of toxic substances on aquatic communities is very broad because statistical methodology is available for the design of studies and the analysis of results, and because both laboratory experiments and field observations are necessary for the elucidation of the effects of toxic substances. Further, structural and functional responses and many different types of organisms are being studied. There are, however, some important statistical concepts which are applicable to the many aspects of the present topic. These concepts will be discussed as they apply to laboratory experiments and to field studies.

Laboratory experiments and field studies are complementary, and each, on its own, is limited by different inadequacies. Examples of recommendations from the literature which acknowledge their complementary roles are the U.S. National Research Council (1981) and Sloof and de Zwart (1983). In the National Research Council report, field observations and experimental work are deemed necessary to the understanding of the structure and the function of baseline ecosystems. Sloof and de Zwart (1983) recommend that they be applied simultaneously since field studies have low specificity to toxic stress and low signal to noise ratio, while laboratory studies suffer from poor ecological significance. These distinctions are of consequence to the statistical methods which can be applied and also to the strength of the conclusions which can be drawn.

The statistical concept which is fundamental to the present topic involves the separation of variation due to different sources, for example, in a field study, it is the separation of the variation associated with natural fluctuations of the ecosystems or with other modifying factors, from that associated with toxic substances which is important. The strength of the conclusions about cause and effect weaken as we go from laboratory experiments to field observations. There is a corresponding change in statistical methods and the nature of the statistical inferences that can be drawn, and thus the discussion will be developed using this correspondence.

The need to consider design and analysis as integral parts of a study will be documented. Study design will be emphasized because it is the critical step that determines how much information, related to the objectives of the study, can be extracted from the data produced. Several papers, in which recommendations were made for use of statistical methods, will be considered, and these will be related to the statistical terminology which will be defined. Methods of analysis will be discussed and examples from studies of benthic invertebrates will be given.

2 STUDY DESIGN

The term study is used here to include both experiments and programs under which different types of field observations are collected. There are many texts on the design of experiments, but few for the design of environmental field studies. Elliott (1977) and Green (1979) consider aspects of both design and analysis, and a number of authors, e.g. Miller (1984) and Hurlbert (1984), criticize other environmental scientists either for not using statistical methods or for using them improperly.

The incorporation of statistical principles in the design of field studies is perhaps the most important role for statistics. Miller (1984) observes that much biological research is conducted without addressing the question of how much data is required to estimate a quantity with a specified precision or to detect a change of a specified size. However, the points raised by Miller emphasize only one consideration in the design of a study, and, the one which often results in a superficial use of statistics. Before further discussing design techniques, some definitions are required.

2.1 Experiment, Field Sampling Program and Observational Study Defined

The distinction between experiments and planned surveys is that, in the former, the experimenter has control over the assignment of treatments to experimental units, whereas, in the latter, the observer has no control over the factors which determine the grouping of observational units (Cox, 1958). Thus, our major concern, the field study, would be classified as a planned survey. Neither Green (1979) nor Elliott (1977) have attempted to name such field studies further than to call them sampling programs, and, as such, have given the standard statistical references for the theory of survey sampling (Yates 1960, Stuart 1962, Cochran 1977).

Hurlbert (1984) has called sampling programs mensurative experiments and, what statisticians call simply experiments, he has named manipulative experiments. In the present paper (Figure 1), Cox's definition of an experiment will be used, and a field study, in which randomization is used, will be called a sampling program. When a study is subject to the limitation that randomization cannot be used, it will be called an observational study (Snedecor and Cochran 1980). Studies conducted in laboratory microcosms, mesocosms and *in situ* enclosures have the potential for the control of the assignment of treatments to experimental units and thus can be classified as experiments. Clearly, the ability to control sources of variation other than that due to treatments will be much less than in simpler laboratory experiments and larger error terms would be expected.

2.2 The Steps in Planning a Study

The link between the design of an investigation and the analysis of the resulting data needs to be emphasized. Cochran and Cox (1957) state that there are three parts to the planning of an experiment: 1) a clear and specific statement of the objectives, including questions, to be answered, hypotheses to be tested and effects to be estimated, 2) description of the experiment covering all details related to treatments, experimental materials, measurements to be made, number of replicates and mechanism for assigning treatments to experimental units; and 3) an outline of the methods of analysis. Box *et al*. (1978) take this further to position the experiment in the iterative process of scientific learning which includes: 1) the deductive phase consisting of abstraction of specific and testable hypotheses from existing knowledge and 2) the inductive phase which occurs when conclusions obtained from the results of an experiment provide inferences about general knowledge. Statistical inference allows the uncertainty to be quantified. Similarly, Green (1979) states that there must be a logical flow in an environmental study between the steps: purpose, question, hypothesis, sampling design, statistical analysis, tests of hypotheses, interpretation and presentation of results.

2.3 Comparison of the Experiment and the Field Study

Similar design techniques are available for the three types of studies (Table 1), and, in general, the purpose of the techniques is to permit valid statistical inferences to be drawn, which are free from bias and possess the best achievable precision. The statistical concepts of experimental design, free of mathematical technicalities, are given by Cox (1958). A clear distinction between the potential of experiments and surveys can be rendered by first reviewing the features of experiments.

Given that the objectives and hypotheses have been established and the treatments, experimental units and the nature of the observations decided upon, Cox (1958) gives the requirements for a good experiment as freedom from systematic errors, sufficient precision, a wide range of applicability, simplicity and provision for assessing the uncertainty in the conclusions. Techniques such as blocking (grouping of units into sets of like units) and the use of concomitant variables (adjustment of the main observation to account for different values of the concomitant variable) are used to increase precision by removing explainable variation from the error term (or unexplained variation). Random assignment of treatments to units protects against systematic errors by randomizing unknown patterns in the uncontrolled variation, and provides a means of assessing uncertainty. Although exact significance tests and probability limits can be calculated from the permutation distribution provided by randomization, in practice, they are calculated by assuming that the errors follow a certain distribution (usually normal). The assumption that they constitute a random sample from the distribution of assumed form is effected by the physical procedure of randomization. In summary, the aim is to eliminate as much of the uncontrolled variation, e.g., by grouping of the units, and then to randomize the remainder.

The essential difference between the experiment and the field study is the control over the choice of the levels of the factors and of the assignment of factors to units that is exercised by the individual(s) planning an experiment but is unavailable in field studies. It is this control, together with the randomization of the uncontrolled variation, that permits the establishment of cause and effect. The lack of it permits conclusions about associations and possible causes only.

Another limitation of the field study is the difficulty of performing physical randomization. Most of the well-known statistical methods such as analysis of variance, t-tests and their non-parametric counterparts are based on the assumption that the data are a random sample. The robustness to non-normality of t-tests and of the analysis of variance tests for comparing several means, and the availability of transformations, make the assumption of normality much less of a problem than failure of the assumption of randomness (Box et al., 1978). When the method of sampling is something other than random, bias may be introduced, e.g., due to periodic variation or a trend, and further there is no reliable way of estimating the standard error of the primary quantities being estimated (Cox, 1958). For the dependence known to be of the form of serial correlation, Scheffé (1959) shows that, under the assumption of normality, the probability of a type I error, for a two-tail test of the mean, ranges from 0.00001 to 0.14 for serial correlations of -0.4 and 0.4, respectively, when the nominal significance level is 0.05. Thus, with negatively correlated errors the test would be conservative, that is, too few significant results would be obtained, and with positively correlated errors too many significant results would be obtained. In general, with correlated errors, the true value of the significance level is different from the nominal, but unknown.

Methods for increasing precision, analogous to blocking and concomitant variables in experimental design are also available for field study design (Table 1). Stratified random sampling consists of grouping units into homogeneous strata and of drawing a random sample from within each strata. Adjustment for sources of bias in field studies can be accomplished by the analysis of covariance if observations are made on a variable which will influence the response variable of interest. Again the principle is the removal of terms from the error component to the component of explainable variation. In some observational studies, it will be possible to use some of the design techniques (Table 1). Since randomization is not possible, considerably more effort must be expended in the statistical analysis to ascertain that the assumptions of the analysis are not violated. Despite this, considerably weaker conclusions are unavoidable. Not only are we limited to conclusions about asociations, but also, in observational studies, bias and the validity of statistical inferences are a problem.

2.4 Further Considerations: Replication, Hypotheses, Efficiency

Estimation of the error variance from the experiment or field study requires replication. In an experiment it is replication of the treatments that is required; in field sampling it is replication at the level at which differences in response are being investigated. For example, in a quantitative study to estimate the number of species in a defined area, the replication must consist of sampling a number of sites within the area, not just replicate samples at one site.

Hurlbert (1984) discusses at length the misuse of statistical analyses in field experiments when improper or no replication was performed. He has drawn on many examples from the ecological literature.

The major points which he makes, related to replication, are: 1) methods of statistical inference should not be used if there is no replication in an experiment; 2) when factors such as spatial or temporal variation are important, treatments must be assigned within groups of units defined by spatial and temporal considerations; 3) sub-sampling within a treatment should not be confused with replication of the treatment; and 4) the term replicability, used, e.g. in studies involving microcosms, to mean that experimental units are similar prior to application of treatments, should not be confused with replication which means applying the same treatment to more than one of these similar experimental units. Point 2 is covered by the design method of blocking and thus Hurlbert's discussion of interspersion and randomization deals with the idea of separating experimental units into blocks to remove known sources of variation from the error term, followed by random assignment of treatments within the blocks.

Since many of the principles of design in experiments and field studies are similar, it follows that the similar considerations regarding replicates are appropriate. Green (1979) summarizes them succinctly in stating that replicate samples should be taken within each combination of time, location and any other observable variable thought to influence response. Elliott (1977) gives some concrete advice on how to accomplish this objective in sampling programs for benthic invertebrates, but these methods are of much broader applicability.

The logic of inductive inference allows only for a hypothesis to be shown to be false. In any particular investigation, the null hypothesis is formed and a test of significance provides the probability of obtaining a result at least as unfavorable to the null hypothesis as that observed in the present experiment. An alternative hypothesis must also be formulated since this will help in the choice of a suitable statistic, one particularly sensitive to departures, from the null hypothesis, of the form specified by the alternative hypothesis. The alternative hypothesis will also determine whether a one- or two-tailed test of significance will be formed.

One further aspect is the efficiency of the statistical procedure which is to be used. Efficient methods require fewer observations for estimation or hypothesis testing than non-efficient ones, and, in hypothesis testing, this will be balanced against the power of the test to detect departures from the null hypothesis.

3 METHODS OF ANALYSIS

In the analysis of data from all types of studies, graphical examination of the data prior to more formal analysis is important. In fact, plots should be used throughout the analysis, the type depending upon the method of analysis. Plotting methods are particularly useful in the detection of departures from the assumptions made in the analysis. For observational studies, plots and descriptive statistics may be all that is possible, whereas, for well-designed studies (Green 1979 and Hurlbert 1984), they may be all that is necessary.

3.1 Experiments

Methods for the analysis of experiments include analysis of variance, with the associated techniques of multiple comparisons, analysis of covariance and response surface methods, and the specific methods for bioassay experiments. There are many good texts on the analysis of variance and related topics, and these vary in mathematical difficulty (e.g. Cochran and Cox 1957, Scheffé 1959 and Box *et al*. 1978). The classical references for bioassay are Finney's books (1964, 1971). More recent references, specifically for toxicity tests, can be found in Sládeček *et al*. (1984), including a review by Sprague (1969). Capizzi *et al*. (1985) propose improved methods of analysis of chronic aquatic toxicity tests which account for the dependence of response on concentration, the existence of censored data, and the need for transformations.

Laboratory microcosms and *in situ* enclosures are attempts to bridge the gap between over-simplified laboratory experiments and the many complications of field studies. Factorial experiments can be used to simultaneously assess the effects of more than one toxic substance. Used together with blocking techniques, and possibly partial replication, factorial experiments make it possible to study the response of an organism to mixtures of substances.

The functional responses of organisms to toxic substances have been studied mostly through laboratory experiments. Additional statistical methods that could be useful to survival studies are techniques for the study of failure-time data (Kalbfleisch and Prentice 1980).

3.2 Field Studies

Since natural systems are under observation in field studies, before investigating the relationship between the response of the feature of interest to the level of another factor, it is necessary to determine how to characterize this feature. For example, if an organism exhibits a random spatial distribution, the number of organisms per unit area can be characterized by fitting a Poisson distribution. This leads to the consideration of three components: 1) the feature(s) of a population or organism which have been used to describe status and the statistical methods which have aided the description; 2) the ways in which these features might be expected to change due to toxic and other stresses; and 3) the modifications to the statistical methods or the adoption of new methods so that the relationship between the features of the population or organism and the factors representing stress can be examined.

Features which have been used to characterize populations include: 1) properties of individual species, 2) composition of the population, 3) abundance of the components of the population and 4) derived variables such as diversity indices, biotic indices and linear combinations of abundances. Points 1) and 4) can be analyzed by univariate methods, whereas 2) and 3) require multivariate methods with the distinction that 2) consists of qualitative data (actually binary) and 3) of quantitative data. For a univariate response, the mean level or other measure of centre, measures of dispersion, and the form of frequency distribution (particularly with respect to random versus contagious spatial distribution) are of interest. This will usually be accomplished by fitting an appropriate probability distribution (Elliott 1977, Snedecor and Cochran 1980) to the data or by the use of procedures based on ranks. For the multivariate data sets, the aim is to condense the data so that relationships between variables can be seen and groups of similar objects can be identified and the methods generally used are the largely descriptive techniques of clustering and ordination (Pielou 1969, Gordon 1981).

The types of changes which might be expected as a result of toxic stress include: 1) spatial differences in the properties of individuals or populations, 2) temporal changes in these properties, and 3) relationships between the properties of the individual or population and the level of the contaminant or other modifying factor, where level refers to both continuous and categorical variables. Because of the need to adjust for factors other than toxic stress, these other factors should be included in the model. Familiar methods for univariate responses are regression, analysis of covariance, estimation of the components of variance from nested designs, and fitting of frequency distributions (Draper and Smith 1981, Snedecor and Cochran 1980, Box _et al_. 1978). Alternatively, non-parametric procedures are available for comparing samples and for tests of randomness versus alternatives such as trend (Lehmann 1975). Univariate procedures for changes in regression relationships (Esterby and El-Shaarawi 1981), changes in distribution (Pettitt 1979) and classification (El-Shaarawi _et al_. 1981) are available. As well as the multivariate procedures of clustering and ordination, procedures such as canonical variables (Pielou 1969) can be used to study the relationship between sets of biological and chemical and physical variables.

At the functional level, field studies can be used to characterize the occurrence of abnormalities such as tumors in fish. Here methods of categorical data analysis such as contingency table analysis and methods for proportions, are appropriate (Snedecor and Cochran 1980, Plackett 1974).

3.3 Checking the Model

The specification of the design can be written in the form of a model which expresses the functional relationship between the response variables, the variables whose effect we want to study, and an error term. In general, the form for one response variable is

$$y = \text{structural model} + \text{error}$$

or

$$y = f(\underline{x}_1) + \varepsilon.$$

When we speak of removing terms from the error component of the model to the component of explained variation this can be represented as moving from a model

$$y = f(\underline{x}_1) + \varepsilon(\underline{x}_2)$$

to

$$y = f(\underline{x}'_1) + \varepsilon(\underline{x}'_2)$$

where the dimension of $\underline{x}'_1$ is greater than that of $\underline{x}_1$ and the dimension of $\underline{x}'_2$ is smaller than that of $\underline{x}_2$, since, in the second case, fewer variables contribute to random error. Regression models are familiar, but all statistical procedures are based on models. When parametric procedures are used, tests of the validity of the assumptions of the statistical model might consist of 1) tests of fit of a particular distribution to a set of data (e.g. χ^2 goodness-of-fit tests) or 2) residual analysis (Draper and Smith 1981) after an analysis of variance or regression analysis. The residual

$$\varepsilon = y - f(\underline{x}_1)$$

contains the information about failure of the assumptions of the model.

The assumption upon which most non-parametric tests are based is randomness and, in fact, many of the tests consist of testing a hypothesis of randomness using a statistic that is sensitive to departures from randomness of the form specified by the alternative hypothesis.

Several approaches may be taken for the use of methods of statistical inference in observational studies. A model can be postulated, the corresponding statistical analysis performed and then the assumptions examined using diagnostic checking. This proceeds iteratively, redefining the model as indicated by the checking procedures in the previous step, until the diagnostic checks provide no evidence of departures from the assumptions. This is what must be done when regression methods are used on happenstance data. The problems in the analysis of happenstance data and the remedies, in terms of experimental design methods, are given in Box *et al*. (1978).

4 EXAMPLES FROM BENTHIC STUDIES

The examples have been chosen from studies of benthic invertebrates because a number of well designed studies have been reported and the book by Elliott (1977) provides a good foundation in statistical methods for the design of sampling programs and the analysis of the resulting data.

4.1 Quantitative Study Design

The objective of a quantitative study is to estimate the number per unit area of each species in some well-defined population, and may include further specifications related to the reasons for collecting these numbers. Elliott (1977) gives the major considerations in designing such a study as the determination of 1) the dimensions of the sampling unit, 2) the number of units to sample and 3) the method of choosing the units to be sampled. It has been shown that a small sampling unit is most efficient when the spatial distribution is contagious (references given by Elliott are Beall 1939, Finney 1946 and Taylor 1971). The number of sampling units can be determined if the form of the frequency distribution of a species is known. The idea behind methods for choosing sampling units is to draw random samples from homogeneous populations. This leads to stratified random sampling if the substrate is not homogeneous. Optimal allocation of the total number of samples can be accomplished by choosing the number of samples for each stratum in proportion to the standard deviation of the stratum.

When the objective is the estimation of the population mean, the estimation techniques of survey sampling theory are appropriate. Often however, comparisons between strata are the primary concern, and then techniques such as the analysis of variance or covariance are required.

4.2 Design for Estimation of Population Mean

Three studies, in which random sampling was used in the estimation of the population mean, are Russell (1972), Cuff and Coleman (1979) and Resh (1979). Both Resh and Cuff and Coleman provide a clear explanation of how the random sampling was accomplished; the former being an example of simple random sampling, the latter of stratified random sampling.

Resh estimated the density of the larvae of the caddisfly *Chematopsyche pettitti* in a riffle of uniform depth (10-12 cm) with substrate particle size of 4-12 cm in diameter located in Rock Creek, Carroll Co., Ind. The area was divided into a series of 30.5 x 61.0 cm rectangular quadrats and 26 quadrats were chosen at random. Two adjacent samples were taken at each location. The number of larvae of *C. pettitti* was counted in each sample, and the means, obtained from untransformed and logarithmically transformed counts, were reported.

A negative binomial distribution was fitted to the data, but it should be noted that Resh's analysis did not properly reflect the sampling design. All 52 samples were used to fit the distribution and the conclusion that the spatial distribution of C. pettitti is contagious was drawn from this. Use of the 52 samples in this fashion does not separate the variation due to location and that due to sample within location. It would be more appropriate to fit a distribution to the sum of the counts of the two samples at each station. Inferences about the mean of a negative binomial distribution can be obtained from likelihood intervals using either the maximum likelihood function for the mean, μ, or the joint likelihood function for the two parameters μ and k (Esterby and El-Shaarawi 1984). To do this the likelihood function should be written with parameters μ and k (see Elliott 1977, p. 24) instead of p and k, noting that μ=pk.

Cuff and Coleman (1979) used stratified simple random sampling, with proportional allocation of the number of stations to the strata, to estimate the number of crustaceans, of polychaetes and of mollusc individuals/0.3 m^2 in Western Port (Victoria, Australia). The total number of stations which could be sampled was limited to 41 due to time and labor considerations. The random selection of stations in a stratum was achieved by sub-dividing a map into areas corresponding to 60 x 60 m areas and selecting the required number of stations for each stratum using a random number table. Proportional allocation means that the number of stations selected from a stratum is proportional to the total number of stations in each stratum, i.e. to the total number of 60 x 60 m areas in this case. The location of the stations, marked on the map, was then determined using shipborne radar. The data were transformed using logarithms and 95% confidence limits for the population means were reported. No explanation for this choice of stratal boundaries was given. Comparison of within-stratum variances, suggests that they did not obtain more homogeneous regions by stratification. This would explain why, in their evaluation of alternate sampling designs, stratified random sampling did not produce substantially more precise estimates than simple random sampling.

Russell (1972) also used stratified random sampling, with proportional allocation, and although he didn't explain how the random sampling was accomplished, the objective of defining homogeneous regions by stratification seems to have been given more consideration. A preliminary survey was conducted to permit contours of equal abundance to be constructed and from this, strata were determined.

4.3 Design for Comparison of Strata

Green (1968) used stratified random sampling of Crib Island, Queensland, for the filter-feeding bivalve *Notospisula parva* Petit. A preliminary random sampling showed high aggregation and a dependence of density, x, on distance from mean low water. Thus a transect from mean low water to mean high water was divided into ten strata and six 43 cm^2 core samples taken from each stratum on each sampling date. After transformation using log (x + 1), an analysis of covariance was performed to assess changes in density with time. Again using the results of a preliminary study, Green and Hobson (1970) determined three faunal zones using association analysis on the presence or absence of 54 species of larger invertebrates in 400 samples in the region of Barnstable Harbor, Massachusetts. The centre of two of these zones, which were at different tidal heights, were then randomly sampled monthly during one year for one of the more abundant species, *Gemma gemma*. An analysis of variance on the logarithmically transformed densities was conducted to determine if there were spatial and temporal differences in the density of *G. gemma*.

4.4 Separation of Natural and Contaminant-Induced Variability

Resh (1979) discussed why life history information should be used in designing benthic studies. DeMarche (1976) demonstrated a relationship between the characteristics of a stream community and the temporal changes in the sediments of a small unpolluted river. Considerations such as these need to be made in order to design sampling programs that will allow the separation of the variability due to modifying factors other than toxic substances. Relative to the study designs described above, this means adding another level of sub-division, but the same principles are involved. Further, the types of changes expected due to exposure of an organism or community would be useful information. An example is the reduction in the coefficient of variation of the percentage of chironomids in samples in a section of a stream with higher copper concentrations relative to a section with lower concentrations (Winner *et al*. 1980).

4.5 Examples of Non-Parametric and Cluster Analysis

Examples of the analysis of studies of benthic communities or organisms in the presence of toxic substances were not located. Methods appropriate for such studies will be briefly described for three observational studies and then similar methods will be applied to two observational studies of benthic organisms.

Univariate non-parametric methods have been used for assessing the change in a benthic community, as exhibited in an index derived from the species count vector, in relation to organic pollution (Bell *et al*. 1981), and for testing for differences between concentrations of a toxic substance in water at two locations (El-Shaarawi *et al*. 1985). The minimal assumptions of the non-parametric procedures were important, since, in the first case, choice of an appropriate distribution is difficult, and, in the second case, the presence of numerous samples below detection limits was handled by ranking observations. Ranks were also used for clustering stations on the basis of concentrations of many toxic substances measured in water and sediment when some samples were below detection limits (El-Shaarawi *et al*. 1985).

Bell *et al.* (1981) reduced the species-count vectors to a single number such as a diversity index or a biotic index and considered two situations: 1) a single set of samples is collected and a measure of pollution is the correlation between the univariate biological descriptor and the time since a pollution event or the initiation of continued pollution, or between the biological descriptor and the distance from the source of pollution; and 2) a set of samples is collected from each of a number of regions, of varying levels of pollution (previously defined), in a body of water and the identical distribution of the biological descriptor in each region indicates no relationship between the level of pollution and the status of the biological population. Permutation statistics were used, which, for the first case, is based on the regression coefficient and, for the second case, is an analogue of the between-region sum of squares in the analysis of variance. For example, the statistic in the first case would be

$$h = \sum_{j=1}^{L} t_j f_j$$

where for a total of L samples, t_j = time of the jth sample and f_j = biological descriptor. To test the null hypothesis of randomness (no relationship with time), the statistic, h, is calculated for all possible permutations of the L samples and the significance level is given by the proportion of samples at least as unfavorable to the hypothesis as the observed h. Note that randomness means that all permutation values are equally probable. For example, if a diversity index was used for f_j in the form of the statistic h which involves distance, high values of h indicate a pollution effect since diversity would be expected to increase with the distance from the source of pollution. The rank tests which correspond to these two cases are the rank correlation coefficient and the Kruskall-Wallis statistic. Another possibility for the first case, which was not considered by Bell et al., is the detection of a change in a univariate biological variable at an unknown time due to a pollution event. This comes under the topic of change-point methods and a non-parametric method is given by Pettitt (1979). Parametric methods of detection (Brown et al. 1975) and estimation (Esterby and El-Shaarawi 1981) are also available.

To test the hypothesis of no difference between the concentration of a toxic substance in water at a station in the lower and one in the upper Niagara River, against the alternative of higher concentrations in the lower river, each pair of concentrations was ranked and the sign test was applied (El-Shaarawi et al. 1985). The two stations were sampled at about the same time (usually the same day) over seven years and, by pairing concentrations by date, changes over time which affected both stations could be eliminated. Further the analysis was made possible by ranking the data since the level of the toxic substance was reported either quantitatively, as below the detection limits, or as measurable.

The multivariate set of data consisted of the concentrations of 28 chlorinated organic substances in water and suspended sediments at six stations along the length of the Niagara River (El-Shaarawi et al. 1985).

The stations were ranked separately for each substance and the standardized Euclidean distance calculated between the vectors of ranks for each pair of stations. A complete linkage clustering performed on these distances showed that the two lower river stations were similar and the three upper river stations were similar.

4.6 Contaminants in Benthic Fauna and Sediments

Fox et al. (1983) reported the concentrations of 12 organochlorine contaminants in sediments, oligocheates and amphipods at five Lake Ontario stations sampled in April and July, 1981. Because of the limited nature of the data set, any analysis must be based on minimal assumptions and this is in agreement with the few analyses done by the authors. Only the relationship between contaminant concentration in the organisms and the sediment will be considered here. The important features can be determined from plots such as those shown in Figure 2. From these plots it appears that, except for PCBs, there are usually only two groups rather than separate points corresponding to each station due largely to the separation of station 210 from the others. The stations were ranked separately for each contaminant and the Euclidean distance was calculated between the rank vectors of all pairs of stations for each combination of compartment and date. The complete linkage cluster analysis (Figure 3) shows that, when the similarity of the stations is assessed using all 12 contaminants together, station 210 is different from the others only for sediments and oligocheates in April and that otherwise station 209 is similar to 210, as is 208 for amphipods in July. The Spearman rank correlation coefficient (Snedecor and Cochran 1980) is seen to be more conservative than the usual product moment correlation coefficient (Table 2) but provides the same general conclusions for PCBs and HCB. It can also be calculated when some samples are below detection limits, as was the case for one sample when oligocheates collected in July were analyzed for 1,3,5-TCB (Table 2).

4.7 Major Benthic Taxa in Lake Ontario: Density versus Depth

Golini (1979) reported the abundance of the major benthic taxa obtained from a survey of surficial sediments of Lake Ontario in September 1977 in which sample locations were determined by a polygonal grid. He observed that the abundance varied inversely with depth. This was based upon the calculation of mean relative abundance of stations in depth zones as determined by contours on a lake map (Figure 5). As a preliminary step in a further examination of this relationship, scatter plots of the density of the major taxa versus depth were prepared (examples are given in Figure 4). For all major taxa, the relationship with depth, which is most pronounced, is the increase in the variability of the density at shallower depths. The increase in scatter began at about 100 to 130 meters for Nematoda, Pontoporeia, Sphaeriidae and Chironomidae, whereas the scatter appeared to increase with decreasing depth over all depths for Oligochaeta. This suggests differences between stations at the same depth, for the shallower depths, that are not due to depth. A grouping of the stations on a basis other than depth was sought by applying to the densities of the seven major taxa a non-hierarchical, nearest-centroid clustering method which uses the Euclidean distance and has the number of clusters fixed (Anderberg 1973). The densities for each major taxa were standardized by subtracting the mean and dividing by the standard deviation of the taxa prior to clustering.

Dominating the distribution of stations amongst clusters for the total number of clusters, k, equal to 2 to 8, is the first large cluster consisting primarily of deeper stations (Table 3). The difference between the clusters can be measured by R, the between cluster sum of squares expressed as a proportion of the total sum of squares, both summed over all seven standardized densities. By determining eight clusters, 66 percent of the total variation is accounted for by the between cluster sum of squares and the one large cluster is split into two. To better understand what is happening as the stations are divided into groups, various plots and summary statistics for the individual taxa were prepared. For eight clusters, a number of the stations in the 10-35 m and 36-99 m depth contours, which were previously in the first cluster, were assigned to the eighth cluster (Figure 5). The effect on the densities of individual taxa in the clusters was examined and the example of Pontoporeia is given here. No relationship between either density or the degree of scatter and depth is evident for cluster 8 (Figure 6b). Further, the increase in scatter with decreasing depth is considerably less for cluster 1 when eight clusters are determined than for cluster 1 when there are five clusters (Figures 6a and c). This is partially due to the movement of stations with higher densities to cluster 8 (Figure 7). From this it appears that depth alone does not separate the stations into groups with similar densities and further analysis would be required to better characterize the groups.

5 DISCUSSION

In the previous sections, the use of statistical methods in the process of collecting scientific information has been considered. There are several other points which are important because they are related to the motivation for many studies of toxic substances. Whether we are concerned with obtaining scientific information to make defensible predictions about the potential hazards of toxic substances or to devleop water quality criteria for toxic substances, the variability of the system being assessed or monitored must be scientifically characterized and the characterization incorporated in the prediction or the statement of the criteria. It is clear that this cannot be given superficial treatment if the prediction or criterion is to be useful. In its simplest form this could consist of determining the form of the probability distribution that adequately fits the type of data collected under the conditions stated in a water quality regulation. Having established the critical level required to protect the system, the nature of the variability in the data can be used to set the exact specification of the regulation. The effect of the form of the probability distribution on the probability of non-compliance has been illustrated in the case of bacterial criteria (Esterby 1982) and similar considerations apply to other types of water quality criteria.

A further consideration is the use of empirical models. Many environmental scientists equate models to systems-analysis type models when predictions are required. It is important to recognize that in order to construct mechanistic models we must have a reasonable understanding of the system being modelled. The process of understanding a biological system will be aided by the construction of empirical models since this helps us to abstract the essential ideas from our experimental results and their consequences. Statistical methods are a means of constructing empirical models.

REFERENCES

Anderberg, M.R. (1973). Cluster Analysis for Applications. Academic Press, New York, 359 pp.

Beall, G. (1939). "Methods of estimating the population of insects in a field". Biometrika, 30, 422-439.

Bell, C.B., Conquest, L.L., Pyke, R., and Smith, E.P. (1981). "Some nonparametric statistics for monitoring water quality using benthic species counts". Envirometrics 81: Selected Papers, pp. 100-120. SIAM, Philadelphia.

Box, G.E.P., Hunter, W.G, and Hunter, J.S. (1978). Statistics for Experimenters. Wiley, New York, 653 pp.

Brown, R.L., Durbin, J., and Evans, J.M. (1975). "Techniques for testing the constancy of regression relationships over time" (with discussion). J.R. Statist. Soc. B, 37, 149-192.

Capizzi, T., Oppenheimer, L., Mehta, H., and Naimie, H. (1985). "Statistical considerations in the evaluation of chronic aquatic toxicity studies". Environ. Sci. Technol. 19, 35-43.

Cochran, W.G. and Cox, G.M. (1957). Experimental Designs. Second edition, Wiley, New York, 611 pp.

Cochran, W.G. (1977). Sampling Techniques. Third edition. Wiley, New York, 428 pp.

Cox, D.R. (1958). Planning of Experiments. Wiley, New York, 308 pp.

Cuff, W., and Coleman, N. (1979). "Optimal survey design: lessons from a stratified random sample of macrobenthos". J. Fish Res. Board Can., 36, 351-361.

DeMarche, B.G.E. (1976). "Spatial and temporal patterns in macrobenthic stream diversity". J. Fish Res. Board. Can., 33, 1261-1270.

Draper, N.R. and Smith, H. (1981). Applied Regression Analysis. Second edition. Wiley, New York, 709 pp.

Elliott, J.M. (1977). Some methods for the statistical analysis of samples of benthic invertebrates. Second edition. Freshwater Biological Association, Scientific Publication No. 25. Ferry House, U.K., 160 pp.

El-Shaarawi, A.H., Esterby, S.R., and Dutka, B.J. (1981). "Bacterial density in water determined by Poisson or negative binomial distributions". Appl. Environ. Microbiol. 41, 107-116.

El-Shaarawi, A.H., Esterby, S.R., Warry, N.D., and Kuntz, K.W. (1985). "Evidence of contaminant loading to Lake Ontario from the Niagara River". Can. J. Aquat. Sci., 42, 1278-1289.

Esterby, S.R., and El-Shaarawi, A.H. (1981). "Inference about the point of change in a regression model". Appl. Statist. 30, 277-285.

Esterby, S.R. (1982). "Fitting probability distributions to bacteriological data: considerations for surveys and for water quality guidelines". Journal Français d'Hydrologie, 13, 189-203.

Esterby, S.R., and El-Shaarawi, A.H. (1984). "Spatial and temporal ditribution of total and fecal coliform concentrations, 1966-1970". In A.H. El-Shaarawi, Ed., Statistical Assessment of the Great Lakes Surveillance Program, 1966-1981, Lake Erie. Scientific Series No. 136, Environment Canada, Burlington, Ont., pp. 151-232.

Finney, D.J. (1946). "Field sampling for the estimation of wireworm populations". Biometrics, 2, 1-17.

Finney, D.J. (1964). Statistical Method in Biological Assay. Second edition. Griffin, London, 668 pp.

Finney, D.J. (1971). Probit Analysis. Third edition, Cambridge University Press, Cambridge, 333 pp.

Fox, M.E., Carey, J.H., and Oliver, B.G. (1983). "Compartmental distribution of organchlorine contaminants in the Niagara River and the western basin of Lake Ontario". J. Great Lakes Res., 9, 287-294.

Golini, V.I. (1979). *Benthic Macro- and Meioinvertebrates of Lake Ontario 1977: Distribution and Relative Abundance*. Fisheries and Marine Service Manuscript Report No. 1519, Department of Fisheries and Oceans, Burlington, Ontario, 41 pp.

Gordon, A.D. (1981). *Classification*. Chapman and Hall, London, 193 pp.

Green, R.H. (1968). "Mortality and stability of a low diversity subtropical

intertidal community". *Ecology*, 49, 848-854.

Green, R.H. and Hobson, K.D. (1970). Spatial and temporal structure in a temperate intertidal community, with special emphasis on *Gemma gemma* (Pelecypoda:Mollusca)". *Ecology*, 51, 1000-1011.

Green, R.H. (1979). *Sampling Design and Statistical Methods for Environmental Biologists*. Wiley, New York, 257 pp.

Hurlbert, S.H. (1984). "Pseudoreplication and the design of ecological field experiments". *Ecological Monographs* 54, 187-211.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

Lehmann, E.L. (1975). *Non-Parametrics, Statistical Methods Based on Ranks*. Holden-Day, Inc., Oakland, California, 457 pp.

Miller, D.R. (1984). "Distinguishing ecotoxic effects". In P.J. Sheehan, D.R. Miller and G.C. Butler, Eds. *Scope 22, Effects of Pollutants at the Ecosystem Level*. Wiley, New York.

Pielou, E.C. (1969). *An Introduction to Mathematical Ecology*. Wiley, New York, 286 pp.

Pettitt, A.N. (1979). "A non-parametric approach to the change-point problem". Appl. Statist. 28, 126-135.

Plackett, R.L. (1974). The Analysis of Categorical Data Griffin, London, 159 pp.

Resh, V.H. (1979). "Sampling variability and life history features: Basic considerations in the design of aquatic insect studies". J. Fish Res. Board Can 36, 290-311.

Russell, H.J. (1972). "Use of a commercial dredge to estimate a hardshell clam population by stratified random sampling". J. Fish Res. Board Can., 29, 1731-1735.

Scheffé, H. (1959). The Analysis of Variance. Wiley, New York, 477 pp.

Sládeček, V., Hawkes, H.A., Alabaster, J.S., Daubner, I., Nöthlich, I., de L.G. Solbé, J.F., and Uhlmann, D. (1984). "Biological examination". In M.J. Seuss, Ed. Examination of Water for Pollution Control, Volume 3, Biological Bacteriological and Virological Examination. Pergamon Press, Oxford.

Sloof, W., and DeZwart, D. (1983). "Bio-indicators and chemical pollution of surface waters". Environmental Monitoring and Assessment 3, 237-245.

Snedecor, G.W. and Cochran, W.G. (1980). Statistical Methods. Seventh edition. The Iowa State University Press, Ames, Iowa, 507 pp.

Sprague, J.B. (1969). "Measurement of pollutant toxicity to fish - I. Bioassay methods for acute toxicity". Water Res. 3, 793-821.

Stuart, A. (1962). Basic Ideas of Scientific Sampling. Griffin, London.

Taylor, L.R. (1971). "Aggregation as a species characteristic". In Statistical Ecology, Vol. 1, Ed. G.P. Patil *et al.*, pp. 357-372. Pennsylvania State Univ. Press, University Park.

U.S. National Research Council. (1981). *Testing for Effects of Chemicals on Ecosystems*. National Academy Press, Washington, 103 pp.

Winner, R.W., Boesel, M.W., and Farrell, M.P. (1980). "Insect community structure as an index of heavy-metal pollution in lotic ecosystems". *Can. J. Aquat. Sci.*, 37, 647-655.

Yates, F. (1960). *Sampling Methods for Censuses and Surveys*. Griffin, London.

TABLE 1 Techniques for the design of studies and their consequences

Experiments		Field Sampling Programs		Observational Studies	
Design Technique	Consequence	Design Technique	Consequence	Design Technique	Consequence
Random assignment of treatments to units	Eliminate bias Valid tests of significance	Random selection of observational units	Eliminate bias Valid tests of significance		
Replication	Error estimation	Replication	Error estimation	Replication	Error estimation
Blocking	Increase precision	Stratification	Increase precision	Stratification	Increase precision
Concomitant variable	Increase precision	Covariate	Increase precision	Covariate	Attempt to eliminate bias Increase precision

TABLE 2 **Correlation coefficients between organisms and sediment concentrations in July for the data of Fox *et al.* (1983)**

Contaminant	Correlation Coefficient or Significance Level					
	Oligocheates			Amphipods		
	r	r_s	α_s	r	r_s	α_s
PCBs	0.93	0.70	0.12	0.35	0.50	0.23
HCB	0.96	0.90	0.04	0.47	0.30	0.34
HCBD		0.98	0.04		0.90	0.04
1,3,5-TCB		1.00	0.008		0.70	0.12

r is the product moment correlation coefficient between concentrations (Fox *et al.* 1983).
r_s and α_s are the Spearman correlation coefficient and corresponding significance level.
PCBs polychlorinated biphenyls, HCB hexachlorobenzene, HCBD hexachlorobutadiene, 1,3,5-TCB 1,3,5-trichlorobenzene

TABLE 3 **Number of stations per cluster and R, the sum of squares ratio, obtained for the cluster analysis of the seven major taxa reported by Golini (1979)**

Number of Clusters	Number of Stations per Cluster 1	2	3	4	5	6	7	8	R
2	132	19							0.28
3	128	2	21						0.42
4	127	2	15	7					0.53
5	125	3	1	13	9				0.56
6	125	4	2	13	2	5			0.63
7	119	4	2	1	11	5	9		0.65
8	100	4	3	1	10	5	5	23	0.66

FIGURE CAPTIONS

Figure 1 Relationships between types of studies.

Figure 2 Examples of plots of contaminant level in organisms and sediments for the two sampling period using the data of Fox *et al*. 1983. The outer axes in the first two plots show which stations had a 1,3-DCB level below detection limits for the organism or sediment or both.

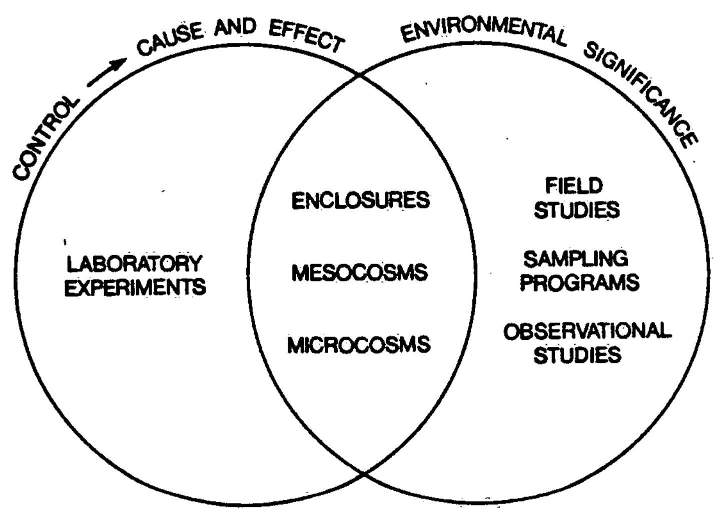
Figure 3 Dendrograms showing the Euclidean distance between stations or groups of stations by compartment and date obtained from the ranks for all 12 contaminants.

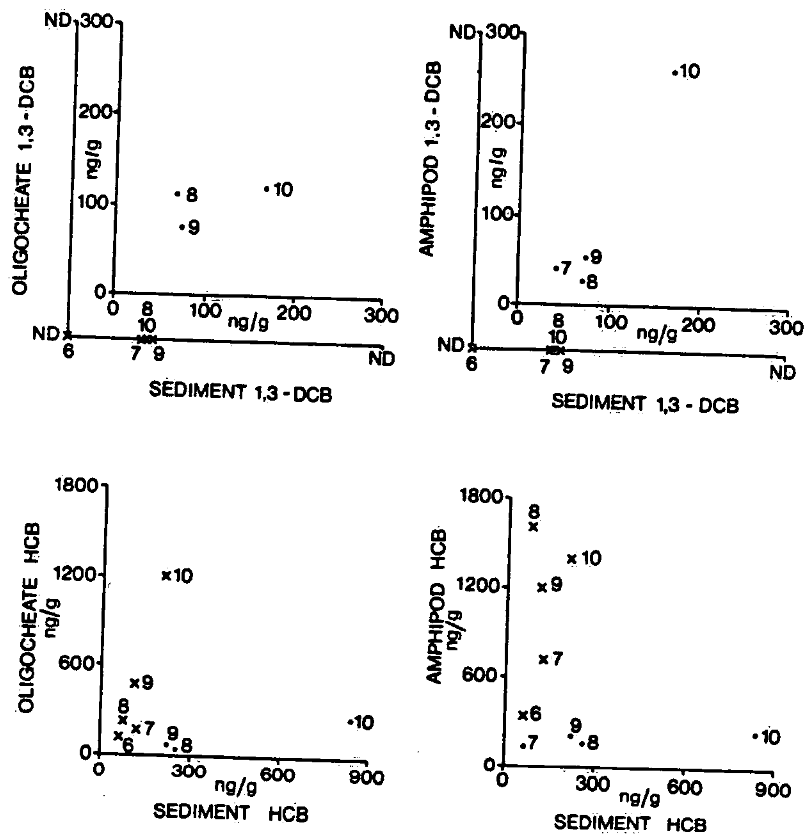
Figure 4 Scatter plots of the number/m^2 of *Pontoporeia* and Oligocheata versus depth for 151 Lake Ontario stations as reported by Golini (1979).

Figure 5 Location of the Lake Ontario stations in the eight clusters as determined using the standardized densities of the seven major benthic invertebrate taxa reported by Golini (1979).

Figure 6 Scatter plots of the number/m^2 of *Pontoporeia* versus depth for stations in clusters 1 and 2 when a total of eight clusters, k=8, were determined, and in cluster 1 when k=5.

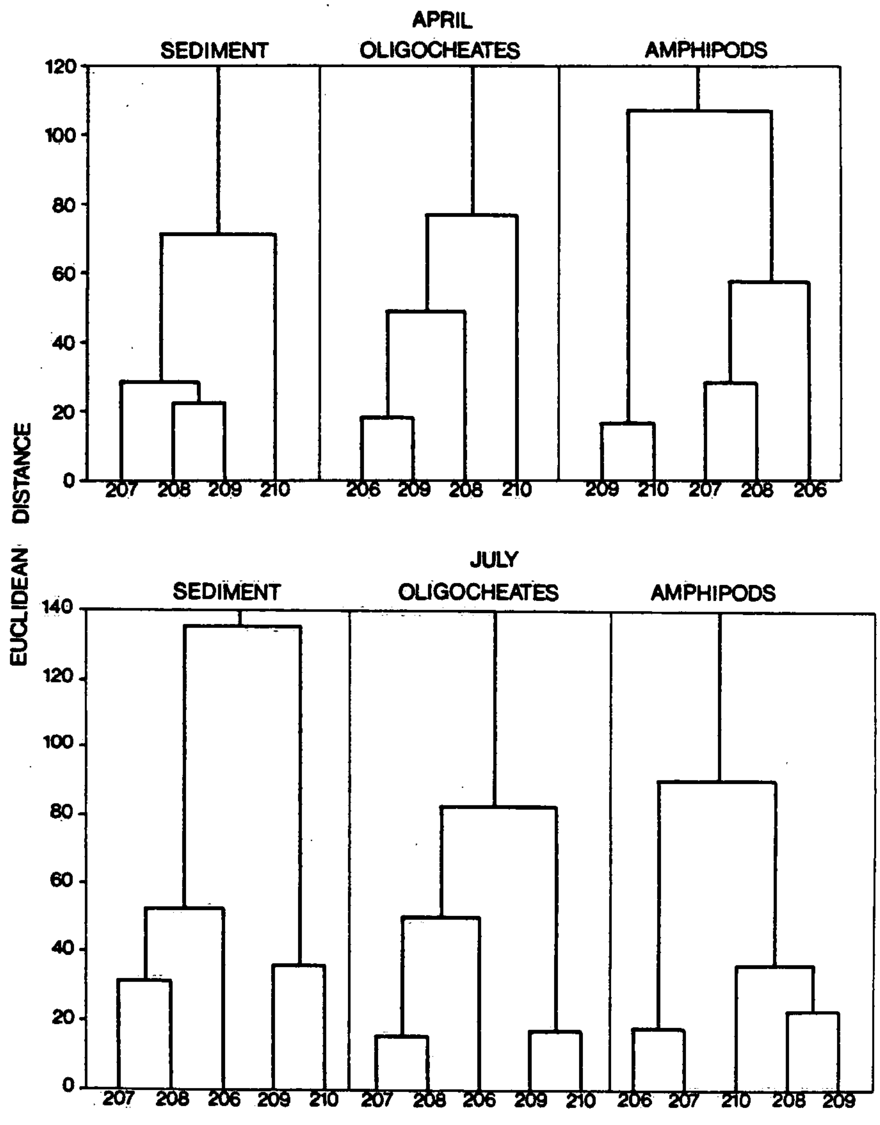
Figure 7 Frequency histograms for *Pontoporeia* for clusters 1 and 8 when the total number of clusters, k, equals 8 and for cluster 1 when k=5.

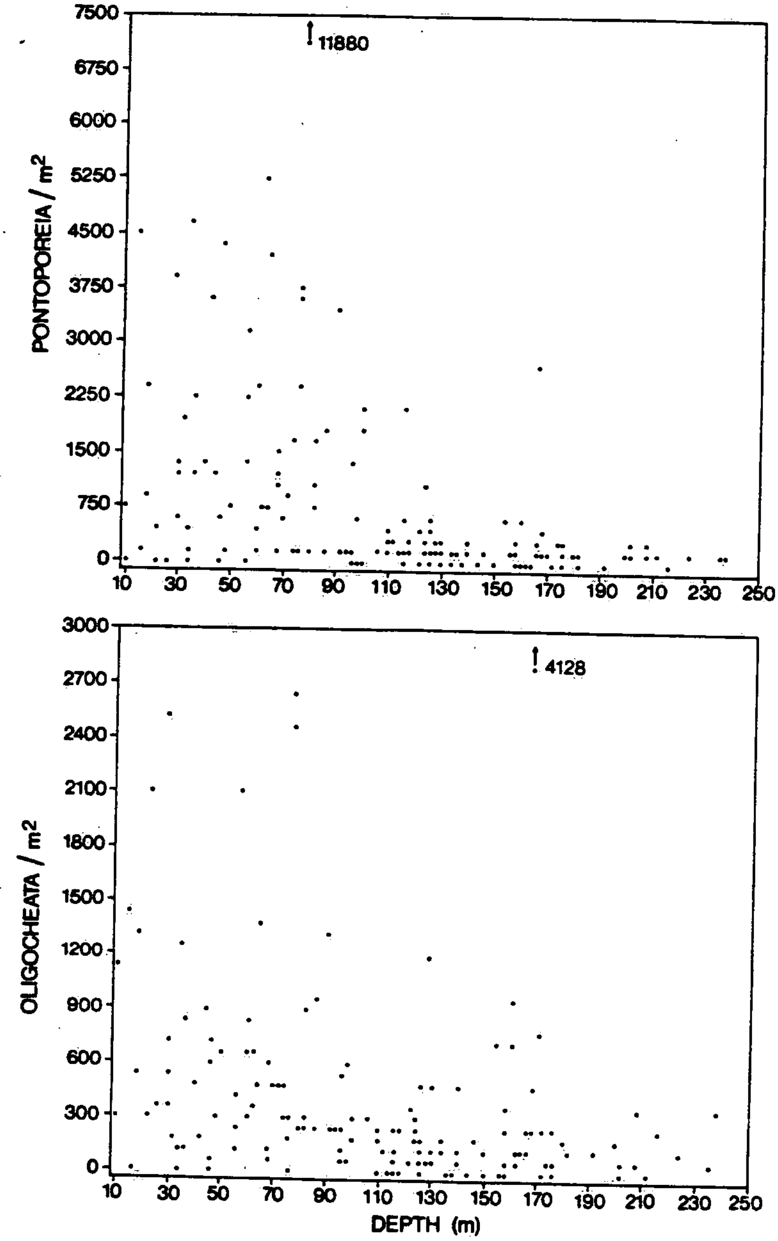

CONTROL → CAUSE AND EFFECT
ENVIRONMENTAL SIGNIFICANCE
LABORATORY EXPERIMENTS
ENCLOSURES
MESOCOSMS
MICROCOSMS
FIELD STUDIES
SAMPLING PROGRAMS
OBSERVATIONAL STUDIES



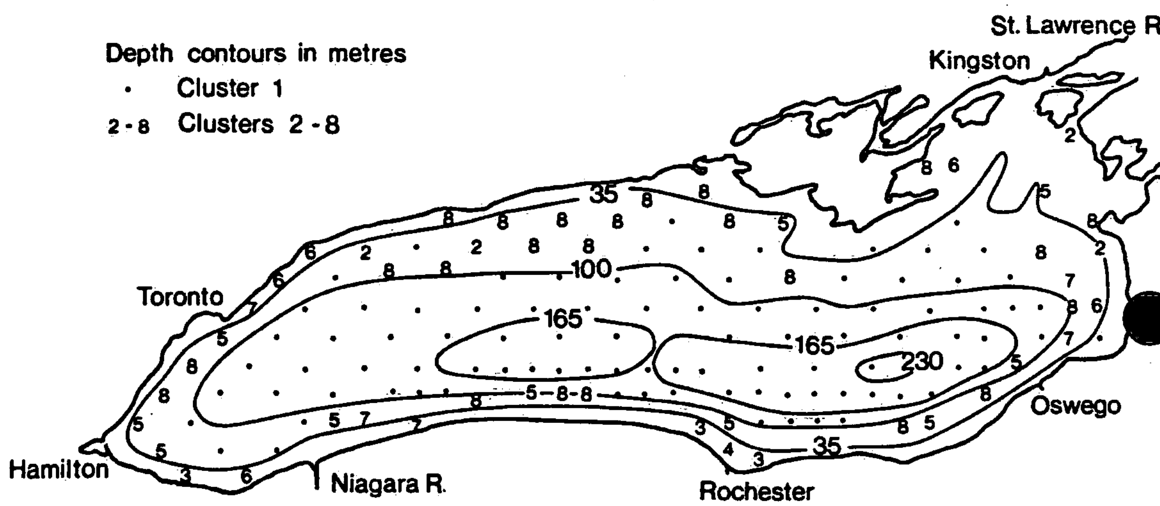
ND - Below detection limits
• - April
× - July

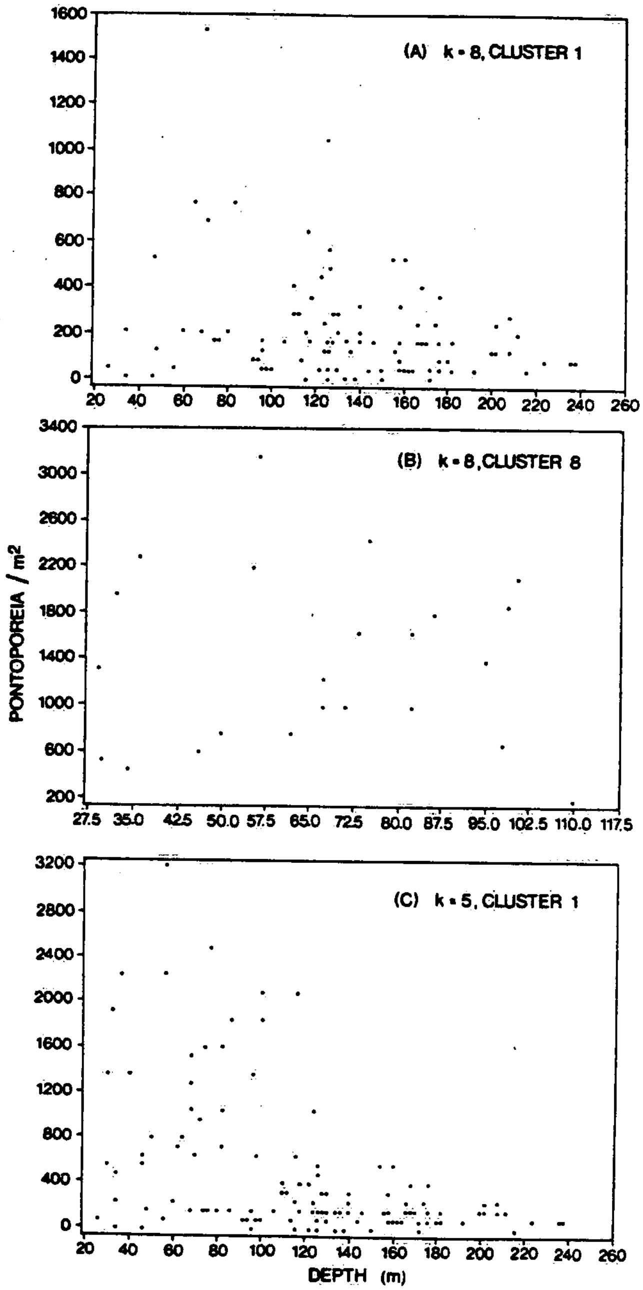
station		station	
6	206	9	209
7	207	10	210
8	208		

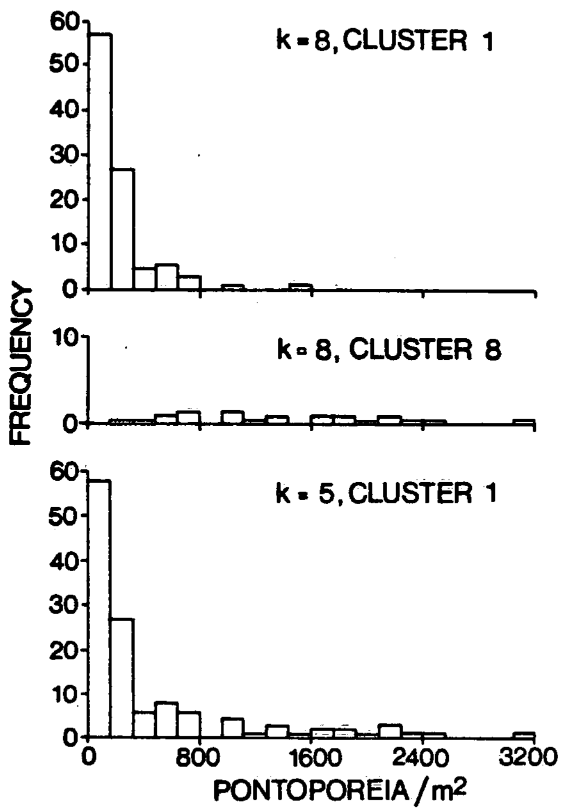

APRIL
SEDIMENT
OLIGOCHEATES
AMPHIPODS
120
100
80
60
40
20
0
207 208 209 210
206 209 208 210
209 210 207 208 206
EUCLIDEAN DISTANCE
JULY
SEDIMENT
OLIGOCHEATES
AMPHIPODS
140
120
100
80
60
40
20
0
207 208 206 209 210
207 208 206 209 210
206 207 210 208 209



11880
PONTOPOREIA / m²
7500
6750
6000
5250
4500
3750
3000
2250
1500
750
0
10 30 50 70 90 110 130 150 170 190 210 230 250
4128
OLIGOCHEATA / m²
3000
2700
2400
2100
1800
1500
1200
900
600
300
0
10 30 50 70 90 110 130 150 170 190 210 230 250
DEPTH (m)


Depth contours in metres
• Cluster 1
2 - 8 Clusters 2 - 8
St. Lawrence R.
Kingston
Toronto
Hamilton
Niagara R.
Rochester
Oswego
35
100
165
165
230
35


(A) k = 8, CLUSTER 1
(B) k = 8, CLUSTER 8
(C) k = 5, CLUSTER 1
PONTOPOREIA / m²
DEPTH (m)


FREQUENCY
k = 8, CLUSTER 1
60
50
40
30
20
10
0
k = 8, CLUSTER 8
10
0
k = 5, CLUSTER 1
60
50
40
30
20
10
0
0
800
1600
2400
3200
PONTOPOREIA / m2