AQUATIC EFFECTS TECHNOLOGY EVALUATION (AETE) PROGRAM

Review of Potentially Applicable Approaches to Benthic Invertebrate Data Analysis and Interpretation

AETE Project 2.1.3a

Review of Potentially Applicable Approaches to Benthic Invertebrate Data Analysis and Interpretation

Discussion Paper

Prepared for:

Aquatic Effects Technology Evaluation (AETE) Program
Natural Resources Canada
555 Booth Street
Ottawa, Ontario
K1A 0G1

Prepared by:

Roger H. Green
Department of Zoology
University of Western Ontario
London, Ontario
N6A 5B7

March 1999

AQUATIC EFFECTS TECHNOLOGY EVALUATION PROGRAM

Notice to Readers

Review of Potentially Applicable Approaches to Benthic Invertebrate Data Analysis and Interpretation

The Aquatic Effects Technology Evaluation (AETE) program was established to review appropriate technologies for assessing the impacts of mine effluents on the aquatic environment. AETE is a cooperative program between the Canadian mining industry, several federal government departments and a number of provincial governments; it is coordinated by the Canada Centre for Mineral and Energy Technology (CANMET). The program was designed to be of direct benefit to the industry, and to government. Through technical and field evaluations, it identified cost-effective technologies to meet environmental monitoring requirements. The program included three main areas: acute and sublethal toxicity testing, biological monitoring in receiving waters, and water and sediment monitoring.

The technical evaluations are conducted to document certain tools selected by AETE members, and to provide the rationale for doing a field evaluation of the tools or provide specific guidance on field application of a method. In some cases, the technical evaluations included a go/no go recommendation that AETE takes into consideration before a field evaluation of a given method is conducted.

The technical evaluations are published although they do not necessarily reflect the views of the participants in the AETE Program. The technical evaluations should be considered as working documents rather than comprehensive literature reviews. The purpose of the technical evaluations was to document specific monitoring tools. AETE committee members would like to stress that no one single tool can provide all the information required for a full understanding of environmental effects in the aquatic environment.

For more information on the monitoring techniques, the results from their field application and the final recommendations from the program, please consult the AETE Synthesis Report to be published in the spring of 1999.

Any comments concerning the content of this report should be directed to:

Geneviève Béchard
Manager, Metals and the Environment Program
Mining and Mineral Sciences Laboratories - CANMET
Room 330, 555 Booth Street, Ottawa, Ontario, K1A 0G1
Tel.: (613) 992-2489 Fax: (613) 992-5172
E-mail: gbechard@nrcan.gc.ca

PROGRAMME D'ÉVALUATION DES TECHNIQUES DE MESURE D'IMPACTS EN MILIEU AQUATIQUE

Avis aux lecteurs

Examen des méthodes potentielles d'analyse et d'interprétation des données sur les invertébrés benthiques

Le Programme d'évaluation des techniques de mesure d'impacts en milieu aquatique (ÉTIMA) visait à évaluer les différentes méthodes de surveillance des effets des effluents miniers sur les écosystèmes aquatiques. Il est le fruit d'une collaboration entre l'industrie minière du Canada, plusieurs ministères fédéraux et un certain nombre de ministères provinciaux. Sa coordination relève du Centre canadien de la technologie des minéraux et de l'énergie (CANMET). Le programme était conçu pour bénéficier directement aux entreprises minières ainsi qu'aux gouvernements. Par des évaluations techniques et des études de terrain, il a permis d'évaluer et de déterminer, dans une perspective coût-efficacité, les techniques qui permettent de respecter les exigences en matière de surveillance de l'environnement. Le programme comportait les trois grands volets suivants : évaluation de la toxicité aiguë et sublétale, surveillance des effets biologiques des effluents miniers en eaux réceptrices, et surveillance de la qualité de l'eau et des sédiments.

Les évaluations techniques ont été menées dans le but de documenter certains outils de surveillance sélectionnés par les membres de l'ÉTIMA et de fournir une justification pour l'évaluation sur le terrain de ces outils ou de fournir des lignes directrices quant à leur application sur le terrain. Dans certains cas, les évaluations techniques pourraient inclure des recommandations relatives à la pertinence d'effectuer une évaluation de terrain que les membres de l'ÉTIMA prennent en considération.

Les évaluations techniques sont publiées bien qu'elles ne reflètent pas nécessairement toujours l'opinion des membres de l'ÉTIMA. Les évaluations techniques devraient être considérées comme des documents de travail plutôt que des revues de littérature complètes. Les évaluations techniques visent à documenter des outils particuliers de surveillance. Toutefois, les membres de l'ÉTIMA tiennent à souligner que tout outil devrait être utilisé conjointement avec d'autres pour permettre d'obtenir l'information requise pour la compréhension intégrale des impacts environnementaux en milieu aquatique.

Pour des renseignements sur l'ensemble des outils de surveillance, les résultats de leur application sur le terrain et les recommandations finales du programme, veuillez consulter le Rapport de synthèse ÉTIMA qui sera publié au printemps 1999.

Les personnes intéressées à faire des commentaires concernant le contenu de ce rapport sont invitées à communiquer avec M^{me} Geneviève Béchard à l'adresse suivante :

Geneviève Béchard
Gestionnaire, Programme des métaux et de l'environnement
Laboratoires des mines et des sciences minérales - CANMET
Pièce 330, 555, rue Booth, Ottawa (Ontario), K1A 0G1
Tél.: (613) 992-2489 / Fax : (613) 992-5172
Courriel : gbechard@nrcan.gc.ca

Table of Contents

1. Introduction

This review is intended to be supplementary and complementary to the Taylor & Bailey (1997) report *Technical Evaluation on Methods for Benthic Invertebrate Data Analysis and Interpretation* which covered basic design and statistical analysis principles but was limited in the literature which was reviewed, and did not consider multivariate (hereafter abbreviated MV) approaches to benthic invertebrate data analysis and interpretation in any depth. Little attention was paid to freshwater acidic deposition ("acid rain") studies or to estuarine or marine pollution studies. A number of powerful new statistical approaches to monitoring of point source pollution impacts have been developed and applied successfully by marine workers, including to mine waste pollution problems.

Taylor & Bailey de-emphasize the applicability of MV statistical analysis on two grounds. The first is that we want hypothesis-testing statistical methods and that the sampling designs likely to be used for mining monitoring studies would not provide the necessary error degrees of freedom for MV hypothesis-testing methods to be used. The second is that other, more descriptive, MV statistical methods (e.g. ordination and clustering) are not appropriate and not recommended even when used as the first step in a two-step approach - where a descriptive method is used to summarize the structure in the MV data and then a hypothesis-testing method is applied to the first step results. With respect, I disagree on both grounds.

I should also emphasize that my philosophy is one of emphasizing good principles of design and statistical analysis, and suggesting options within that framework to the end user with pros and cons clearly stated, rather than declaring "this is the way". There is more than one legitimate way to skin the cat, especially in applied statistical analysis. Environmental biology has seen too much coming and going of design/statistics fads (Green 1993b), of the *approach du jour*, whether it be diversity indices, the BACI-P design, the reference sites approach, reliance on metrics, or whatever. I will not contribute to it by describing one way of going about things and saying that is how mine discharge monitoring should always be done.

1.1 Correlated variables in multivariate (MV) data

In this review I will first discuss some principles involved in the analysis of MV data. There is the question of whether the variables are in fact correlated. If they are then there are some theoretical considerations to keep in mind when analysing them or testing hypotheses about them. One option is to use only one or a few of the variables in univariate analyses - but which ones? Another option is to use a descriptive MV method to summarize the correlation structure in the MV data and then apply hypothesis-testing statistics as a second step. Finally there is the periodically fashionable option of reducing the MV information to one or more synthetic indices, "metrics", or other descriptors which are then used in univariate analyses.

1.2 Models with MV response data

Hypothesis testing with MV response variable data is discussed in terms of five models: an Analysis of Variance (ANOVA) design as the predictor (including the "reference sites approach" and repeated measures designs), continuous predictor variables, relating three or more sets of variables e.g. the Sediment Quality Triad, and an ANOVA design *plus* continuous variables as predictors i.e. an Analysis of Covariance (ANCOVA) model. Finally some recommendations for hypothesis-testing approaches are made and discussed.

1.3 Interpretation and display of results

A common complaint about MV hypothesis-testing statistics is that the results are difficult to interpret. I discuss interpretation of MV analysis output, effective graphical display of MV analysis results, and interpreting the results in a manner appropriate for the statistical sophistication of the audience.

1.4 Connection between study design and statistical analysis approach

There is a necessary connection between study design and statistical analysis approach, including MV analysis. The ANOVA design should follow logically from the allocation of samples

in space and time. All tests of hypotheses involve an effect (e.g. a difference between putatively impacted areas and control areas) whose variance is compared to an error variance (e.g. differences among control areas). There must be a meaningful error variance, representing "null hypothesis" variation on a spatial (or temporal) scale commensurate with the "impact" contrast. Sometimes there isn't and the best one can do is to use "pseudoreplicate" error (sampling error) for testing impact (Hurlbert 1984). Also there must be a reasonably good estimate of the error variance, which is equivalent in practice to having adequate error degrees of freedom. A rule of thumb is that a minimum of 10 error degrees of freedom (df) is adequate for robustness of an ANOVA F-test in the face of moderate violations of assumptions e.g. homogeneity of variance and normality, in a balanced design with random sampling (Harris 1985).

Power to detect a real difference, e.g. between impact and control areas, is a function of the magnitude of the difference, the error variance, and the sample size. Study designs which are inadequate (no error df to do desired tests of hypotheses) or marginally adequate (low error df resulting in tests that aren't robust and have low power), or which yield data violating parametric statistics assumptions, can often be analysed (and hypotheses tested) using nonparametric or randomisation procedures.

1.5 Literature re: history of development of approaches

Literature reporting study design and statistical analysis approaches to monitoring point sources, from situations other than freshwater mining and pulp and paper discharges, will be cited throughout. There are no sections explicitly devoted to such "exotic" examples. Emphasis will be on point source discharges that are high in metals, for example freshwater/estuarine acidification (Yan et al. 1996; Locke et al. 1994; CJFAS 1994; Somerfield *et al.* 1994a, 1994b; Metcalfe-Smith and Green 1992; Hinch and Stephenson 1987; Yan and Strus 1980) and industrial discharges (Metcalfe-Smith *et al.* 1996), and marine oil & gas (Olsgard *et al.* 1997; Peterson *et al.* 1996; Kennicutt *et al.* 1996b; Hyland *et al.* 1994; Clarke 1993; Warwick 1993; Chapman *et al.* 1991) as well as mining discharges (Austen and Somerfield 1997; Somerfield *et al.* 1994a, 1994b). Workers in marine environments have developed powerful statistical tools which have been well

tested in the field, especially in recent years for impact and monitoring studies re: oil and gas operations (where metal contaminants are a concern, as they are in mining discharges).

There have been three main threads. One, associated with the Plymouth Marine Laboratory (UK), initially emphasized descriptive MV statistical methods but recently has added tests of hypotheses using randomisation methods. Most of their methods are incorporated in the PRIMER statistical package (Carr 1996) and discussed in an accompanying manual (Clarke and Warwick 1994a). Three key references, spanning the period 1982-1993, are (Clarke 1993; Bayne *et al.* 1988; Field *et al.* 1982). A second thread has been the development and application of sophisticated ANOVA designs (with both univariate and multivariate responses) for marine environmental studies. There is an early book by Green (1979), and papers plus a recent book by Underwood have been influential (1997; 1993; 1992; 1981). Other references to this approach are (Green and Montagna 1996; Kennicutt *et al.* 1996a, 1996b; Hyland *et al.* 1994; Green 1993b; Hinch and Green 1989; Green 1989; Clarke and Green 1988; Green 1987, 1984). The third and most recent thread has been the development and application of the Sediment Quality Triad (SQT) concept which relates the three components: benthic biological community response, contaminant concentrations in sediment, and experimentally determined toxicity of the sediments to organisms. Originally proposed by Chapman and Long, there has been much interest in the SQT approach by workers in both marine and freshwater environments (Chapman *et al.* 1997a, 1997b; Green and Montagna 1996; Chapman 1996; Green *et al.* 1993; Chapman *et al.* 1991, 1987).

There is a large literature on biomonitoring using organism uptake and body burden of metal contaminants. Molluscs are often used. Some references are (Bryan and Langston 1992; Metcalfe-Smith and Green 1992; Hinch and Green 1989; Hinch and Stephenson 1987; Imlay 1982; Eganhouse and Young 1978; Davies and Pirie 1978; Luoma and Jenne 1977; Phillips 1976; Smith *et al.* 1975). Multivariate statistical analysis of such data in an ANOVA design is presented in Metcalfe-Smith *et al.* (1996).

2. Correlated variables

When there is more than one variable in a data set, either in a response (Y) set or in a predictor (X) set, the procedure for analysing the data should be carefully chosen and explicitly stated. If only one or a subset of the variables is used then there should be *a priori* reasons for the choice or else the rationale, criterion and procedure for selection should be stated. Partly this is to avoid suspicion that variables were chosen after determining which variables showed something. After all, 1 variable in 20 will "show something" (e.g. a Control versus Impact area difference) in the sense of being significant at the 5% level, even with null hypothesis data.

2.1 Whether the response variables are correlated and what you can do if they aren't

First there is the question of whether the variables *are* correlated. This is usually obvious but it can be tested. The parametric test of the null hypothesis "no non-zero correlations among p variables" is called Bartlett's sphericity test, and the formula, calculations, and a worked example are given in Green (1979). An equivalent randomisation test which avoids the parametric assumptions (e.g. normality) is described in Clarke and Green (1988) and implemented in the PRIMER package (Carr 1996; Clarke and Warwick 1994a).

If the variables are uncorrelated (not very likely with real field data) then a univariate analysis can be done on each variable, but a correction (e.g. Bonferroni's correction) should be applied to determine the significance level to be used in each univariate analysis so that if the null hypothesis is true the probability that *any* of the univariate tests is significant will be 5%. For example if there are five independent (uncorrelated) variables and a univariate ANOVA (Control vs. Impact) is done on each, then each test should be done at a $1-(1-.05)^{1/5} = 0.0102 = 1.02\%$ significance level. To put it the other way around, if each test were done at a significance level of 5%, then if the null hypothesis is true there is a $1-(1-.05)^{5} = 0.226 = 22.6\%$ chance that at least one of the five tests will be "significant".

2.2 If the response variables are correlated

If the variables are correlated (which is likely) then it is necessary to take the among-variables correlation structure into account, in some way. If one or a subset of the variables is used for analysis, then information in the other variables will be included or left out depending on the correlation structure. One should not just arbitrarily choose variables to analyse because that would arbitrarily (and blindly) include or omit information that was in the original set of variables. There are "variable subset selection" methods which use the correlation structure to choose one or more of the variables so as to maximize the information contained in all of the variables in the original set (Green 1979; Orloci 1978, 1976, 1973). A similar result is usually obtained by doing a Principal Components Analysis (PCA) on the original set of variables and then using the variable having the strongest relationship with the first Principal Component (PC1), adding in the variable having the strongest relationship with PC2, and so on until the desired subset size is obtained. The variables in the subset thus obtained will not be very strongly correlated and will contain as much as possible of the information in the original set of variables.

Alternately the PC scores for the largest few PCs from a PCA could be used, and they would be completely uncorrelated (Principal Components are always orthogonal to each other), but unless the PCs (which are linear additive functions of the original variables) are easily interpreted it is better (easier to explain to an audience) to use a relatively uncorrelated subset of the original variables. If they are organisms you can show pictures of them. It's hard to show pictures of Principal Components. Nonetheless, a successful example of using PC scores as Y variables is in Gray *et al.* (1988).

If the original set of correlated variables is a response to impact (e.g. organism abundances as Y variables) then the correlations themselves are not a problem. Multivariate statistical analysis handles correlation structure just fine - after all that is what it is designed to do. The problem is often that there are too many Y variables in relation to the amount of data, which in practice expresses itself as too few error degrees of freedom. This was Taylor and Bailey's concern. Thus

a reduction in number of Y variables to a smaller number (a subset) of them is often a necessary preliminary step to doing hypothesis-testing MV statistical analysis (e.g. MV ANOVA). Selection of a subset of variables which maximizes the information contained in all of the variables in the original set is exactly what is wanted.

2.3 Correlated predictor variables

If the original correlated variables are predictors, say in a regression or ANOVA or ANCOVA model, then the correlations *are* a problem because predictor variables are supposed to be independent. At the worst the analysis will fail due to collinearity (no unique stable solution). At the best all tests of significance of particular predictor variables or evaluations of their relative importance in prediction will be compromised. A relatively uncorrelated subset of the original variables, which retains as much of the information in the original set as possible, can be used as predictors instead. As noted above, use of PC scores as the "new" variable set for a MV ANOVA may not be the best approach. However, two examples of it working well with a predictor set of variables are in Kennicutt *et al.* (1996a) and Metcalfe-Smith *et al.* (1996).

2.4 Description of structure in MV correlated data

Apart from hypothesis testing, descriptive multivariate analyses (e.g. ordination and clustering) are often useful for displaying the structure that is in the correlations among the variables. The samples may be along one or more environmental gradients, for example, and an ordination analysis (PCA, Correspondence Analysis, Non-metric Multidimensional Scaling, etc.) can show that the inter-variable correlations result from such an environmental gradient. Alternatively, the samples may cover discontinuous areas which are different for natural or pollution-related environmental reasons (e.g. a Control and an Impact area), and a cluster analysis can use the inter-variable correlations to group the samples by those areas. Ordination and clustering can be done as a first step in a two-step analysis sequence, where the second step is a hypothesis-testing analysis such as MV ANOVA or contingency table analysis, but they are often sufficiently explanatory by themselves that formal tests are unnecessary. For example see Figure 4.2 in Green

(1979), which relates species composition to a Before-After and Control-Impact (BACI) design.

2.5 Reduction of structure in MV data to a single synthetic variable - indices and metrics

The information in MV response data (e.g. species abundances) is often boiled down to indices, metrics and their ilk. This seems to be a fashion that comes and goes. A diatribe against it plus a literature review is in section 3.5 of my book (Green 1979).

First of all, more information than necessary is lost. Data requiring more than one dimension cannot be reduced to a single number without loss of the information that was in the dimensions >1. Second, such derived synthetic variables often exhibit bad statistical behavior when used as response variables. This is especially true when they are in whole or in part concocted as ratios of original variables (Atchley *et al.* 1976). There are better analysis approaches when a ratio is the response of interest (Green 1986). Third, indices and metrics are generally poor at distinguishing pollution impact from "pure" or harsh natural environments. For example, diversity indices typically have higher values in slightly eutrophied conditions than in very oligotrophic or very eutrophic conditions, and they have lower values in clean estuaries than in the nearby clean rivers or the nearby clean marine environment (estuaries are osmotically stressful). Fourth, there are many indices and metrics that have been derived, supposedly to respond to different things, and yet when they are all calculated for the same environmental impact study they are usually highly correlated (i.e. redundant). In other words, whatever they are measuring in theory, in practice they are responding similarly. Fifth, when indices and metrics indicate pollution impact they rarely explain *why* they do, i.e. what is going on.

Unfortunately (in my opinion), the use of indices and metrics to describe biological community composition in pollution monitoring has recently gone into a new wave of popularity, e.g. Fore *et al.* (1994), and the US EPA has emphasized metrics in its monitoring protocols. My personal theory is that the reason indices and metrics remain popular is that many of the senior people in industry, environmental consulting firms and government who are responsible for environmental

monitoring and impact studies have engineering backgrounds. Engineers like formulae, simple answers and one dimension at a time. They tend to be strong in the component of intelligence which is good for mathematics (and engineering), as opposed to the (quite separate) component of intelligence which is good for geometry, geography, graphs showing more than one dimension - and MV statistics. But the reality is that biological response to the environment, including to pollution, usually requires more than one dimension to describe it. One-dimensional approaches can miss a lot of that reality. They summarize it well (to the extent that it can be), but leave out the complexity.

2.6 Literature re: statistical analysis of correlated variables

I will close this section by mentioning some good general references to the statistical analysis of correlated variables, and to descriptive MV statistical analysis. Re: the former, Seber (1984) is a basic reference. To enter at an easier level try Manly (1994). For more of an orientation to ecological applications see Legendre and Legendre (1997). Re: descriptive MV statistical analysis, Pielou (1984) is a good place to start. Green (1979) may be useful. For how to use PCA in ecological applications, i.e. how many PCs to use and try to interpret, see Jackson (1993a). A nice ecological application of PCA is in Sprules (1977). Kruskal (1964a, 1964b) are the original papers on Non-metric Multidimensional Scaling (NM-MDS). For comparisons of different kinds of ordination applied to the same data see Gray *et al.* (1988) and Jackson (1993b). For computer implementation of descriptive MV statistical methods, most statistical packages will do PCA. The coefficients which relate PCs to the original variables are called different things and standardized in different ways in different packages and don't look the same, so be prepared for that. (Stay away from Factor Analysis. The FA model is wrong for our kinds of data. It may be good for some applications in the social sciences but it's not for us.) NM-MDS is in PRIMER, NT-SYS, and elsewhere. Correspondence Analysis (similar to PCA but for contingency table type count data which have chi square rather than normally distributed errors) is in SIMCA and elsewhere. Legendre and Legendre (1997) give a worked example from which a program can easily be written. Clustering is in SAS, Minitab, NT-SYS, CLUSTAN and elsewhere. Nemec and Brinkhurst (1988a, 1988b) describe a method for determining how many clusters are "significant"

in a cluster analysis, and so do Clarke and Warwick (1994a). A variety of descriptive MV statistical analysis programs in BASIC are given in Orloci (1978).

3. Hypothesis-testing MV statistical analysis with correlated response variables

General aspects of this have been discussed above. Now I will consider the various kinds of environmental monitoring related hypotheses and the statistical models related to them.

3.1 MV response with ANOVA design as the predictor

3.1.1 Introduction

Probably the most common situation is prediction of an MV response from an ANOVA design. A simple example would be a contrast of benthic community species abundances between Control and Impact areas, or between Before and After times in the Impact area. Conceptually this is no different from a univariate (UV) one-way ANOVA with two groups. The only difference is that the distributions of the samples in the groups are contrasted in a space whose dimensionality is equal to the number of response variables. For a UV ANOVA the null hypothesis is that the groups have similar distributions on one axis (i.e. in one dimension). For an MV ANOVA with two response variables it is that the groups have a similar distribution in a 2-dimensional space, and so on. A major difference between MV and UV ANOVA is that in the former, but not in the latter, correlations among the response variables contain relevant information about group distributions and group differences. Because of this, in an MV ANOVA two groups can have significantly different - even non-overlapping - distributions while at the same time having similar and not significantly different distributions on each of the response variables taken separately. Thus, as noted in section 2 above, with correlated response variables it is necessary to take the among-variables correlation structure into account. MV ANOVA does this, whereas UV ANOVAs done on each variable do not.

Of course an MV ANOVA design can be more complicated than a simple one-way two group ANOVA. It can have more than two groups (treatment levels); be factorial e.g. the Before vs. After Control vs. Impact, or BACI, design (Green 1979); or nested; or repeated measures; or whatever. Any ANOVA design which can be analysed with one response variable can be analysed

with more than one, as an MV ANOVA, if there are sufficient error df.

3.1.2 Testing hypotheses with MV ANOVA

In any MV ANOVA the test statistic is different than in a UV ANOVA and has different degrees of freedom. Several test statistics are usually tabulated in statistical package output and usually do not differ in indicating significance or nonsignificance. Pillai's Trace is conservative and is my preference (see Green (1979) for discussion). It is usually easier to get significance in MV tests than to interpret the significant results, so I prefer to use a conservative test. Roy's Greatest Root is different from the others, in that it tests group differences in only one dimension - the dimension which best separates the groups. It is a more powerful test because it is not looking for group separation in the entire p-dimensional space (p = the number of response variables), and that saves degrees of freedom. It is appropriate in situations where only one dimension, or gradient, of differential group response is expected - e.g. where different concentrations of a contaminant (a contamination "gradient") are imposed on samples from a community (Gray *et al.* 1988).

3.1.3 Statistical computing software and references re: MV ANOVA and CDA

MV ANOVA is easy to implement in most of the comprehensive general-purpose statistical packages (e.g. SAS, Minitab, SPSS, Systat)). In procedures such as GLM (acronym for "general linear model") and ANOVA one enters a model statement of the form "response = predictors", which for a UV ANOVA (abundances of one species) with two groups (Control vs. Impact areas) might be "y = CI". To do the same analysis as an MV ANOVA, or MANOVA, one simply enters a model statement with more than one response variable e.g. "y_1 y_2 y_3 = CI". That alone would produce three separate UV ANOVAs. To obtain the MV analysis output including the MV test statistics one typically adds a subcommand "MANOVA - - - ". Canonical Discriminant Analysis (CDA) displays the group distributions in a few dimensions, the ones which best separate the groups, and is typically done following a significant MANOVA result as an aid to interpreting it.

There are a number of good general and fairly accessible references for MANOVA and CDA

(Legendre and Legendre 1997; Harris 1985; Green 1979; Pimentel 1978; Cooley and Lohnes 1971, 1962). Seber (1984) is at a higher level. Pielou (1984) explains CDA as an ordination method i.e. without the MANOVA hypothesis-testing step. Personally I don't think that CDA *is* an ordination method, or that it should be done unless a MANOVA test has shown significance, but Pielou's explanation of how CDA is done (with a worked example) is excellent. There are numerous examples of ecological/environmental application of MANOVA and CDA based on sophisticated ANOVA designs. Two are the GOOMEX project (Kennicutt *et al.* 1996b) and the following seven papers), and my lab's studies on the use of freshwater bivalves for monitoring heavy metal pollution (Metcalfe-Smith *et al.* 1996; Hinch and Green 1989; Hinch and Stephenson 1987).

Examples where randomisation rather than parametric tests are used ("ANOSIM" tests) are from the Plymouth lab (Somerfield *et al.* 1994b; Clarke and Warwick 1994a, 1994b; Clarke 1993). Randomisation tests are typically for fairly simple ANOVA designs, e.g. one and two-way layouts without interaction. There are unresolved problems with testing interactions by randomisation tests, which is rather limiting because often tests of interaction are tests of the hypotheses of greatest interest e.g. the Before-After by Control-Impact interaction in a BACI design. Valid testing of interactions by randomisation is currently being worked on by M.J. Anderson of Sydney University, collaborating with K.R. Clarke, P. Legendre and C.J. ter Braak, but for now it is necessary to use classical parametric MANOVA for testing such hypotheses. Actually I would usually do so anyway, even if such randomisation tests existed, just as I do now for ANOVA tests in simpler designs where there are valid randomisation tests. As noted in section 1 above, MANOVA tests are robust when done with balanced designs providing at least 10 error df, and some element of random sampling (Harris 1985).

3.1.4 Repeated measures designs

Repeated measures designs deserve particular comment. In environmental monitoring we often sample at more than one time, and we usually establish sites at the first time. We may, and in fact should, establish them randomly within areas that represent conditions (e.g. Control, Impact) but at subsequent times we usually revisit the same sites. The samples (e.g. grab samples) taken at each site *are* assumed to be re-randomised at each time but the sites are not. This is a repeated measures design (sites are "repeatedly measured") and should be analysed as such. In this situation it would *not* be valid to analyse it as a CI by Times two-way factorial ANOVA. There would be random replication *within* site-times (the grab samples), but sites would be the meaningful level of replication for testing hypotheses and sites at one time are not independent of sites at other times - because they are the *same* sites at every time!

Repeated measures ANOVA comes out of the medical and the social sciences. If you randomly assign patients to drug treatment groups (perhaps including a Control group) then the same people will come back in at subsequent times to have their blood pressure, heart rate, body temperature, or whatever, measured. Thus observations are not independent among times - they are repeated measures on the same replicates. You can't randomly sacrifice people - hence the need for a repeated measures statistical model. Repeated measures ANOVA has come into ecology and environmental studies more recently. See the review paper by Green (1993b) for explanation of repeated measures designs (both simple and complex), how to analyse and interpret them, the necessary assumptions, worked examples, and references.

To prevent confusion, it should be mentioned that references to multivariate analysis of repeated measures can have two quite different meanings. One is that a repeated measures analysis of data having one Y variable (e.g. abundance, density or cover of some species) has been done in the multivariate *mode*. This means that the times are treated as variables and a MANOVA is done in which the distribution of groups (e.g. Control vs. Impact areas) is evaluated in a t-dimensional space, t being the number of times. Given the number of times and the number of sites

we usually have in environmental monitoring, that is not the way we are likely to do the analysis. We generally do it in the univariate mode, which is analogous to a split plot design - a vector of times is nested within each replicate. The other sense of "multivariate analysis of repeated measures" is that there is more than one Y variable. As with other ANOVA designs, a repeated measures ANOVA can be done with one or with more than one Y variable, given sufficient error df.

3.1.5 The reference sites approach

The "reference sites approach" (Bailey *et al.* 1996; Wright 1995; Reynoldson *et al.* 1995), which has antecedents in the earlier work of Wright et al. (1984), should be mentioned here. The advocates of this approach agree on the underlying concept (I like it too), which is in essence that reference conditions should be described by variation (usually multivariate) among a large number of sites in apparently unimpacted conditions widely distributed within the habitat type of interest; putatively impacted sites are evaluated (in effect tested) based on where they fall on the reference sites' probability distribution (e.g. within or beyond 1-.). This can be done either for biological response variables such as species abundances or for environmental variables, or for both (e.g. for the biological response variables "conditioned on" the natural environmental variables used as covariates). However, the principal advocates of the reference sites approach part company on exactly how to go about this statistically.

The point I wish to make here is that the conceptual design is essentially a one way ANOVA (perhaps with stratification by habitat type e.g. depth zones in lakes or stream order in lotic habitats), with two groups (reference sites and putatively impacted sites), but with the *a priori* constraint that the estimate of error variation comes entirely from the reference sites' group. The putatively impacted sites are evaluated only in terms of the reference sites' distribution. Their own distribution does not contribute to the evaluation. In any case, when there is only one test site (as is typical in the reference sites approach) then an ANOVA will estimate the error variance only from the reference sites group, by default.

This in fact makes sense because there is ample evidence that pollution impact often changes not only the means of the Y (response) variables but also their variances (Green and Montagna 1996; Green 1993b; Underwood 1993; Warwick and Clarke 1993). Evidence of heterogeneity of variance (typically a higher variance among impacted sites than among reference sites) suggests a violation of the homogeneity of variance assumption for parametric ANOVA (whether UV or MV) , but in a sense that is irrelevant because it is itself evidence of impact. When there is heterogeneity of variance in an ANOVA design which includes a control group it is in fact philosophically proper to use the control group as the reference re: its error distribution as well as re: its mean values, and that is what the reference sites approach does. Of course if there were not heterogeneity of variance then *both* groups could provide an estimate of the common within-group variance, and a pooled estimate of error variance would be obtained as normally done in an ANOVA.

3.1.6 Problems and philosophy re: MV ANOVA

As discussed in section 2 above, and also by Taylor & Bailey, there is often a problem with too many Y variables in relation to the amount of data, which comes out as too few error degrees of freedom for MANOVA tests of hypotheses. In a one-way ANOVA design, for example, as a general rule the number of observations minus one, less the number of groups, has to exceed the number of Y variables in order for the MANOVA to be possible (i.e. in order for there to be *any* error df for tests of hypotheses). But, as also noted above, we want at least 10 error df in tests, for them to be robust as well as for them to be possible. As a rough rule of thumb the number of observations (less g where g is the number of groups) should be 2 to 3 times the number of Y variables. Some MV analyses, such as Canonical Correlation Analysis (see section 3.2.1 below) which tests relationships between sets of continuous variables (e.g. biological response and environmental predictor variables), are more sensitive than MANOVA to violations of assumptions. An even larger ratio of observations to variables is recommended for them.

It should be noted that the term "observations" as used here refers to the level of replication

being used to obtain the error for the hypothesis test one is talking about. If the study design includes sites within Control and Impact areas, and grab samples within sites, then it is likely that the "among-sites" level of replication would be used as the error for tests. The "among-samples within sites" replication, if used for this purpose, would be what Hurlbert refers to as pseudoreplication (Hurlbert 1984). Thus it is usually the number of sites less g, not the number of samples, that should be 2 or 3 times the number of Y variables. That is more difficult and costly to accomplish than simply increasing the number of samples by dropping the grab a few more times at each site, which is undoubtedly a reason why pseudoreplicated designs are still common.

Another reason is that sometimes there just *aren't* enough reference sites, or a large enough area in which to allocate enough reference sites. Mining monitoring, e.g. these AETE studies, is often like that. The BACI-P design proposed by Stewart-Oaten *et al.* (1986) uses "replication in time" to deal with such situations, but apart from concerns about validity of hypothesis-testing (e.g. Green 1993b), this statistical design is intended for the BACI study design situation i.e. impact studies with Before *vs.* After impact times. It is not appropriate for monitoring programs having no pre-operation times.

I have made the argument (Green *et al.* 1993) that better impact/monitoring studies would be done if less money were spent on the latest technology for measuring variables and instead it was spent on having more sites (*not* more replicate sampling at sites - see Cuff and Coleman (1979)) so that more robust and powerful statistical models and tests of hypotheses (including MV) could be applied. It is also a common practice to try to measure too many variables, but as discussed in section 2 and also later in this section, there are various ways to reduce the original number of Y variables while retaining the maximum possible amount of information, and they should be used. If we do both, i.e. have more sites in our monitoring study designs and *also* apply a first-step analysis to reduce the number of response variables before doing MV hypothesis-testing statistical analysis, then we will greatly improve how we go about these sorts of studies.

3.2 MV response with continuous variables as predictors

A second kind of environmental monitoring related hypothesis is whether biological response variables Y are related to (can be predicted from) a set of continuous variables X, e.g. environmental variables.

3.2.1 Canonical Correlation Analysis (CCA)

The classical parametric statistical model for relating two sets of variables is Canonical Correlation Analysis. However, as mentioned above, it is not a very robust model. It assumes linear additive relationships both among Y and among X variables, and also between Y and X variables, as well as having the usual assumptions of normality and homogeneity of variance. More to the point, all the empirical evidence suggests that Canonical Correlation Analysis is sensitive to violations of those assumptions. The assumption of linear additive relationships among variables is often violated by real data. For the Y variables, species abundances are not usually linearly related. Logarithmic transformation can help to approximate inter-species linear relationships, but the greatest problem is with relationships between species abundances (Y) and environmental variables (X) which are often not only nonlinear but also nonmonotonic (species are most abundant where the environment is optimum for them).

Canonical Correlation Analysis can work well for other kinds of variables. Green (1972) used it successfully to relate variation in bivalve mollusc shell morphology to water chemistry, and Marcus and McDonald (1992) recommend using it to relate effluent toxicity to observed instream effects for the U.S. EPA Complex Effluent Toxicity Testing Program. For species abundances versus environmental variables relationships, however, this statistical model should be used only with great caution. There are other ways to model and test this kind of environmental monitoring related hypothesis. Green (1993a) provides an extensive review of approaches to this problem, and in another paper (Green *et al.* 1993) apply the approaches to data from a pollution gradient in Vancouver Harbour. The information in these two papers will not be presented in detail here.

Briefly, the approaches fall into two categories: one-step and two-step.

3.2.2 One-step methods additional to CCA

Besides Canonical Correlation Analysis (described in most MV statistics textbooks e.g. Seber (1984)), one-step methods include Mantel's test (Mantel 1970) and Procrustes analysis (Gower 1975; Schonemann and Carroll 1970). An ecological application of Mantel's test is in Anderson and Underwood (1997). Methods not mentioned in Green (1993a; 1993b), but which should have been, are Canonical Correspondence Analysis (Jongman *et al.* 1987; ter Braak 1986) and an ANOSIM algorithm (Clarke and Warwick 1994a; Clarke and Ainsworth 1993). Both use randomisation for testing the null hypothesis of no relationship between the two sets of variables, as does Canonical Correspondence Analysis. So does Mantel's test as implemented in many programs and statistical packages. Jackson (1995, 1993a, 1993b) has built on Procrustes Analysis (which was originally an entirely descriptive method) by adding a randomisation test. He illustrates its use on ecological data.

In general, if a parametric test of the Y set vs. X set relationship is to be done, then there has to be an adequate number of observations in relation to the number of variables. The same comments and rules of thumb made above in relation to MANOVA and CDA apply here as well. If the test is done by randomisation, it can be done with a number of variables and a number of observations that would pose problems for a parametric test. However, I believe that this advantage of randomisation tests, compared to parametric tests, is somewhat illusory. The fact that a test can be done doesn't necessarily mean that it should be done, or that it will be robust, powerful or interpretable. The problem with parametric tests on data having too many variables in relation to number of observations is not just a technical one. There are *reasons* why one can't or shouldn't do them, the principal one being that the model is overparameterized. This remains true if a randomisation test is done on those data.

3.2.3 Two-step methods

Two-step approaches discussed by Green (1993a, 1993b) involve reduction of and summary of the information in one or both of the sets of variables prior to modelling and testing the relationship between the sets. Ordination (e.g. PCA) can be used, as previously described, and so can a variable subset selection algorithm. Each PC or subset variable from the Y set can be regressed on the X set by multiple regression, followed by Bonferroni's correction (the PCs and subset variables will be independent, or approximately so). Alternatively, PCs derived from the Y set can be regressed or otherwise related to PCs derived from the X set. See Sprules (1977) for PCA approaches of this kind applied to lake zooplankton communities. Cluster analysis can be applied to one or both sets of variables. If applied to the Y set only, the clusters of observations from the cluster analysis can be used as biologically defined groups in a Canonical Discriminant Analysis (CDA), in a multidimensional space defined by the environmental variables. See Green and Vascotto (1978) and Green (1979) for an example. If cluster analysis is applied to both sets of variables then the cluster frequency count data for the Y set can be related to that for the X set by contingency table analysis (chi-square or loglinear model). The Y-X relationship as expressed in the contingency table of counts can be displayed graphically using Correspondence Analysis.

3.2.4 Relating more than two sets of variables e.g. the Sediment Quality Triad

The Sediment Quality Triad approach, which was mentioned in section 1, involves relating three sets of variables: a biological community response set , a sediment contaminant concentrations set, and an experimentally determined toxicities set. References were given in section 1. In particular see Green *et al.* (1993). To have good sediment contaminant concentrations data, one must have good sediment samples, which may not be reliably available in mining monitoring e.g. the AETE studies. Also, the biological communities must be comparable i.e. not from different biogeographic regions. Neither of these requirements may be satisfied.

3.3 MV response with both ANOVA design and continuous variables as predictors

The two models described above can be combined, so that the biological response is predicted by both an ANOVA design and by continuous environmental variables . This is an analysis of covariance, or ANCOVA (Seber 1984; Cox and McCullagh 1982; Cooley and Lohnes 1971, 1962; Cochran 1957). As with ANOVA, ANCOVA can easily be done as an MV analysis. Whether UV or MV, ANCOVA used on observational data in environmental monitoring gets tricky because predictors are supposed to be independent, i.e. not correlated. The usual ANOVA design in environmental monitoring studies has to do with Control versus Impact areas, or Before impact versus After impact times. Obviously in most such cases the ANOVA design predictor *would* be correlated with (confounded with) predictor variables measuring environmental contaminants. For example the distribution of Control sites on a contaminant covariate would be different than the distribution of Impact sites. To put it another way, ANCOVA is meant to use within-group variation in X to adjust within-group variation in Y. ANCOVA is *not* meant to use a between-*group* difference in X to adjust a between-group difference in Y. Beyak and Green (unpublished) showed in simulation studies that biases to significant tests are greatest when this kind of confounding is accompanied by substantial error in estimation of the X variables (the covariates). Some MV ANCOVA was used in the GOOMEX studies (Kennicutt *et al.* 1996b), and by Metcalfe-Smith (1996).

4. Interpretation of results of hypothesis-testing MV analysis

Presenting examples of multivariate analysis results here and discussing how to interpret them is impractical, and would amount to writing a MV statistics reference book as part of this report. Fortunately general reference books exist which are good at explaining how to interpret hypothesis-testing multivariate analysis results, as provided by computer output (Legendre and Legendre 1997; Manly 1994; Harris 1985; Seber 1984; Green 1979; Pimentel 1978; Cooley and Lohnes 1971, 1962). It is important to keep in mind that parametric hypothesis-testing statistical analyses such as MANOVA/CDA, ANCOVA and CCA usually provide information about and test the significance of more than one "component", even for simple ANOVA designs. The jargon re: these components is confusing. In MANOVA and CDA they are typically referred to as discriminant functions or canonical discriminant functions, in Canonical Correlation Analysis as canonical variates. The latter is a more general term so here I will refer to canonical variates regardless of the MV analysis model. The canonical variates are the axes of the multidimensional space in which the results of the MV analysis are displayed.

For example, a MANOVA on data with two groups would produce one canonical variate regardless of how many response variables there were, because the number of canonical variates produced by a MANOVA is the number of variables, or one less than the number of groups, whichever is less. This makes intuitive sense, geometrically. All the information about the separation of two groups should be expressable in one axis running through the multidimensional "variable space". Similarly, if there are three groups (and at least two variables), complete representation of the distributions of the three groups should be possible in two dimensions, represented by two canonical variates. And so on. However some canonical variates may be trivial. Suppose there are three groups and two variables, but the three groups lie more-or-less in line within the two-dimensional variable space. It is quite likely that only one of the two possible canonical variates would be statistically significant. In Canonical Correlation Analysis the number of canonical variates is the number of variables in the set (Y or X) which has the fewest variables. The same sort of logic applies.

When the hypothesis-testing multivariate analysis is not the parametric kind but instead is the randomisation-testing kind, as with the ANOSIM methods as implemented in the PRIMER package, most of the above does not apply. Fortunately the documents provided with PRIMER (Carr 1996; Clarke and Warwick 1994a) are very clear on interpretation of analysis results, with numerous examples.

Obviously it helps to have a knack for geometry when trying to understand and interpret the results of multivariate statistics. And it follows that effective graphical display is critically important. Most everyone who has gotten into hypothesis-testing multivariate statistical analysis has discovered that it is much easier to obtain significant results than it is to convincingly interpret those results. Some good examples of application of such methods with effective graphical display as an aid to interpretation are in Green and Vascotto (1978), Clarke (1993), Somerfield *et al.* (1994b), Manly (1994), Metcalfe-Smith *et al.* (1996), and some of the GOOMEX papers (e.g. Montagna and Harper 1996). Green (1979) has a section on graphical display of results, with examples. There are books about "tricks" for displaying multivariate data (Everitt 1978; Andrews 1972) but I have never found them to be particularly useful.

Finally, one has to be aware of the statistical sophistication, or lack thereof, of the intended audience. For most general audiences it is best to avoid verbiage which will inevitably be jargon-filled, and instead emphasize effective graphical display. Also, some methods are inherently easier to understand than others. As mentioned in section 3 above, variable subset selection is easier to understand and explain than PCA, and both methods will often accomplish the same purpose (reduction of number of variables). More people have an intuitive understanding of cluster analysis and contingency table analysis than of Non-Metric Multidimensional Scaling and Canonical Discriminant Analysis. There is often more than one way to skin the cat in this business and there is nothing wrong with using methods that are more likely to be comprehensible to your intended audience.

5. Classical versus randomisation approaches

Are classic parametric statistical methods appropriate, and safe to use, on real environmental monitoring data? Or is it best to routinely play it safe and use nonparametric or randomisation methods? First of all, nonparametric statistical methods have largely been superseded by randomisation methods. The former have to be derived for every particular case whereas the latter are in principle quite general. One simply sets up the data in the design layout that pertains, does the parametric analysis and obtains the test statistic value, then randomly re-allocates the data and calculates the test statistic and does that thousands of times until a null hypothesis test statistic distribution is adequately defined. Then the test statistic value for the actual data is compared to the null hypothesis distribution obtained by randomisation. It is pretty straightforward for simple ANOVA designs, but not for more complicated designs like factorial and nested ANOVAs. The problem is that one has to decide *what* to randomise. In a factorial design with replication, for example, the parametric ANOVA tests whether there is an interaction in addition to the main effects. It is the marginal distributions of the main effects model that are tested. But if the data are totally re-randomised then any main effects will vanish along with any interaction, so the null hypothesis is the wrong one ("no interaction given no main effects", instead of "no interaction"). This is one disadvantage of the randomisation testing approach - that it is not clear how to do it properly for some ANOVA designs which we would like to use.

Another disadvantage is that randomisation provides a test but doesn't fit a predictive model as parametric methods do. So the null hypothesis that two sets of variables are unrelated may be rejected, but no model is fit which describes *how* they are related. Interpretation is therefore more difficult. Finally, it is not obvious to me that parametric methods are so unreliable with typical data that randomisation methods are needed. As discussed in section 3 above, parametric methods are quite robust to failures of assumptions as long as the design meets certain criteria. My feeling is that the advantages of the classical parametric statistical methods outweighs concerns about their reliability. I hasten to add that biostatistician colleagues who I greatly respect, K.R. Clarke for example, take the opposite stance on this issue and always use randomisation methods for testing

hypotheses - at least for biological data. Of course one can always use both methods and use the randomisation test as a check on the parametric test. I often do this. The problem is what to do when they disagree, and one method accepts but the other rejects the null hypothesis. The proper, conservative approach is to accept the null hypothesis when either method does so.

Good general references for randomisation tests *per se* are Manly (1991) and Edgington (1995). Manly's RT package does randomisation tests for a variety of hypotheses.

6. General recommendations and conclusions

Monitoring studies should be planned so they provide enough sites, within areas representing conditions, to be able to robustly and powerfully test hypotheses about differences among areas. Put the effort and money into sites, not into variables requiring high-tech measurement or into replicate sampling at sites. Your study design should provide meaningful error variances for tests of hypotheses. This will usually be "among sites" or "sites by times interaction" (the latter in a repeated measures design).

There should be at least 10 error degrees of freedom for robust hypothesis testing. If there are <10 error df check the parametric result with a randomisation test, if one is possible. Since parametric statistical analyses lead to explanatory or predictive models and randomisation tests do not, it is best to use parametric hypothesis-testing methods and to use randomisation tests as checks on them. Some would disagree.

Do not test response variables one univariate test at a time, or arbitrarily choose one or a few of them for hypothesis-testing statistical analysis. Choosing an organism or a contaminant to study because it is important *a priori* is not arbitrary, but choosing one or a few organisms or contaminants to be the Y variables in statistical analyses after the data are in hand - without clearly stated criteria for doing so - *is* arbitrary.

With correlated MV response data you should use MV hypothesis-testing statistics. Bear in mind that the more Y variables there are, the more sites you need. If you have too many Y variables, as often happens when the biological community is the response and you have identified all the organisms in all the samples, then you should reduce the number by variable subset selection, ordination or clustering. If there are correlated predictor (X) variables, reduce them to a smaller number of relatively uncorrelated variables by variable subset selection, ordination or clustering. Then use some hypothesis-testing statistical analysis to relate the responses Y to the ANOVA design and/or the predictor variables X. Which analysis is best to use will depend on the

original model and on the approach used to reduce the Y or the X sets, if that was done.

There is as much art as science in choosing a statistical analysis sequence. Preferences are as much aesthetic as scientific. This should *not* be taken as an excuse for doing any old analysis one wants. The most appropriate methods should always be used. It is just that there is often more than one legitimate way to do it and the judgement of which is the most appropriate way, given the objective, the data and the intended audience, may differ among scientists - and that's OK. If a method shows an impact and can be successfully defended to the audience as showing the impact, then that's all that is needed. Similarly, if a method indicates no impact and can be successfully defended to the audience as capable of showing an impact had there been one, that's all that is needed too.

Some people insist on direct (one-step) methods, or nothing. I happen to like two-step methods, in part because each step provides different information. The description of the structure of the multivariate biological response Y, or of the redundancy among predictor X variables, is useful and important in its own right and not just a mindless data-processing step before statistically relating Y and X. Furthermore, a PCA or a cluster analysis, or variable subset selection, are no harder to interpret than a one-step MANOVA and CDA. After all, we do two-step analyses all the time without really thinking about it. Selecting which variables to use or applying transformations to those we do use before doing ANOVAs, are just as much "first steps" in analysis as are the use of variable subset selection or ordination or clustering to reduce the dimensionality of the description of the response. PCA, for example, is just a rotation of the original variable axes and reduction of the original correlated variables to a few uncorrelated PCs is really just a transformation. Some people don't understand and can't interpret transformations of any kind, log or square root for example. That's tough on them, but one does transformations anyway if they need to be done in order for a valid hypothesis-testing analysis to follow. The procedures I have described aren't any different in that regard.

Two-step approaches remain popular with leading biostatisticians in the environmental impact

and monitoring field and also with environmental biologists who do the work. For example the two-step approach described by Field *et al.* (1982), NM-MDS ordination on the species abundances followed by relating that reduced description of the response Y to the environmental variables X, has been much cited and much used by marine workers over the following 11 years (Clarke 1993), and it and similar approaches are still widely used. There are no general differences between marine and freshwater environments that would suggest the inapplicability of such approaches for freshwater pollution monitoring.

For just testing whether there *is* a Y *vs* ANOVA design relationship or a Y *vs* X relationship (without description, explanation or interpretation of the relationship), Mantel's test is always OK.

Use conservative MV tests and test statistics. It is easy to get significance with MV tests. Concentrate on convincingly explaining why it was significant.

Do most of your statistical analysis in a good standard heavy duty statistical package (e.g. SAS, Minitab, SPSS, Systat) rather than ad hoc "benthic community data analysis programs". Go to other packages or programs as necessary for particular methods e.g. randomisation tests: Primer, RT; NM-MDS: Primer, NT-SYS; Correspondence Analysis: SIMCA; clustering: CLUSTAN.

Be cautious about using CCA and ANCOVA with environmental monitoring observational data. Both methods relate Y responses to X predictor variables. CCA is not robust unless there is a very large ratio of sites to variables, larger than is usually practical. The environmental predictors (e.g. contaminants) are rarely independent of the groups (e.g. Control *vs* Impact areas) in an ANCOVA.

When monitoring over time, as data from sampling the same sites at different times accumulate, a repeated measures design may be appropriate.

The Reference Sites approach is conceptually good but it requires a large number of sites, larger

than is usually practical.

Contaminant uptake, body burden and enzyme response (e.g. metallothionein) are good response variables for environmental monitoring, if the right organism is available. Fish typically move too much. Bivalve molluscs are good but occur in sedimentary environments which may not be present. Caged specimens of either could be used.

Avoid using indices or metrics to reduce MV response data, except where it is necessary to compare with other studies. Even then do it in addition to proper MV statistical analysis.

In writing the paper/report or presenting the talk, avoid the use of technical statistical verbiage. To that end, use "accessible" statistical methods when possible.

7. Bibliography

Anderson, M.J. and A.J. Underwood 1997. Effects of gastropod grazers on recruitment and succession of an estuarine assemblage: a multivariate and univariate approach. Oecologia, *109*: 442-453.

Andrews, D.F. 1972. Plots of high dimensional data. Biometrics, *28*: 125-136.

Atchley, W.R., C.T. Gaskins, and D. Anderson 1976. Statistical properties of ratios.1.Empirical results. Syst. Zool., *25*: 137-148.

Austen, M.C. and P.J. Somerfield 1997. A community level sediment bioassay applied to an estuarine heavy metal gradient. Marine Environ. Res., *43*: 315-328.

Bailey, R.C., R.H. Norris, and T.B. Reynoldson 1996. Study design and data analysis in benthic macroinvertebrate assessments of freshwater ecosystems using a reference site approach. Ninth Annual Technical Information Workshop, 44th Annual Meeting of the North American Benthological Society. (Abstract)

Bayne, B.L., K.R. Clarke, and J.S. Gray 1988. Biological Effects of Pollutants: Results of a Practical Workshop. Mar. Ecol. Prog. Ser., *46*: 1-278.

Bryan, G.W. and W.J. Langston 1992. Bioavailability, accumulation and effects of heavy metals in sediments with special reference to UK estuaries: a review. Envir. Pollut., *76*: 89-131.

Canadian Journal of Fisheries and Aquatic Sciences (CJFAS). 1992. Aquatic Acidification Studies in the Sudbury, Ontario, Canada, Area. Volume 49, Supplement No. 1.

Carr, M.R. 1996. PRIMER User Manual (Plymouth Routines in Multivariate Ecological Research). Natural Environment Research Council, UK, Plymouth, UK.

Chapman, P.M., R.N. Dexter, and E.R. Long 1987. Synoptic measures of sediment contamination, toxicity and infaunal community structure (the Sediment Quality Triad). Mar. Ecol. Prog. Ser., *37*: 75-96.

Chapman, P.M., E.A. Power, R.N. Dexter, and H.B. Andersen 1991. Evaluation of effects associated with an oil platform, using the sediment quality triad. Env. Tox. Chem., *10*: 407-424.

Chapman, P.M. 1996. Presentation and interpretation of Sediment Quality Triad data. Ecotoxicology, *5*: 327-339.

Chapman, P.M., B. Anderson, R.S. Carr, V. Engle, R.H. Green, J. Hameedi, M. Harmon, P.S. Haverland, J. Hyland, C.G. Ingersoll, E.R. Long, J. Rodgers, Jr., M.H. Salazar, P.K. Sibley, P.J.

Smith, R.C. Swartz, B. Thompson, and H. Windom 1997a. General guidelines for using the Sediment Quality Triad. Mar. Poll. Bull., *34*: 368-372.

Chapman, P.M., E.A. Power, and G.A. Burton, Jr. 1997b. Integrative assessments in aquatic ecosystems. In: Sediment Toxicity Assessment. G.A. Burton, Jr. ed. Lewis, Boca Raton, Florida, 313-340.

Clarke, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. Austr. J. Ecol., *18*: 117-143.

Clarke, K.R. and M. Ainsworth 1993. A method of linking multivariate community structure to environmental variables. Mar. Ecol. Prog. Ser., *92*: 205-219.

Clarke, K.R. and R.H. Green 1988. Statistical design and analysis for a 'biological effects' study. Mar. Ecol. Prog. Ser., *46*: 213-226.

Clarke, K.R. and R.M. Warwick 1994a. Change in marine communities: an approach to statistical analysis and interpretation. Natural Environment Research Council, UK, Plymouth, UK.

Clarke, K.R. and R.M. Warwick 1994b. Similarity-based testing for community pattern: the two-way layout with no replication. Mar. Biol., *118*: 167-176.

Cochran, W.G. 1957. Analysis of covariance: its nature and uses. Biometrics, *13*: 261-281.

Cooley, W.W. and P.R. Lohnes 1962. Multivariate Procedures for the Behavioral Sciences. Wiley, New York.

Cooley, W.W. and P.R. Lohnes 1971. Multivariate Data Analysis. Wiley, New York.

Cox, D.R. and P. McCullagh 1982. Some aspects of analysis of covariance. Biometrics, *38*: 541-561.

Cuff, W. and N. Coleman 1979. Optimal survey design: lessons from a stratified random sample of macrobenthos. Can. J. Fish. Aquat. Sci., *36*: 351-361.

Davies, I.M. and J.M. Pirie 1978. The mussel *Mytilus edulis* as a bio-assay organism for mercury in seawater. Mar. Poll. Bull., *9*: 128-132.

Edgington, E.S. 1995. Randomization tests. Marcel Dekker, New York.

Eganhouse, R.P. and D.R. Young 1978. *In-situ* uptake of mercury by the intertidal mussel, *Mytilus californianus*. Mar. Poll. Bull., *9*: 214-217.

Everitt, B. 1978. Graphical Techniques for Multivariate Data. Heineman, London, pp. 1-117.

Field, J.G., K.R. Clarke, and R.M. Warwick 1982. A practical strategy for analysing multispecies distribution patterns. Mar. Ecol. Prog. Ser., *8*: 37-52.

Fore, L.S., J.R. Karr, and L.L. Conquest 1994. Statistical properties of an index of biological integrity used to evaluate water resources. Can. J. Fish. Aquat. Sci., *51*: 1077-1087.

Gower, J.C. 1975. Generalized Procrustes analysis. Psychometrika, *40*: 33-52.

Gray, J.S., M. Aschan, M.R. Carr, K.R. Clarke, R.H. Green, T.H. Pearson, R. Rosenberg, and R.M. Warwick 1988. Analysis of community attributes of the benthic macrofauna of Frierfjord/Langesundfjord and in a mesocosm experiment. Mar. Ecol. Prog. Ser., *46*: 151-165.

Green, R.H. 1972. Distribution and morphological variation of *Lampsilis radiata* (Pelecypoda, Unionidae) in some central Canadian lakes; A multivariate statistical approach. J. Fish. Res. Board Can., *29*: 1565-1570.

Green, R.H. 1979. Sampling design and statistical methods for environmental biologists. Wiley, New York.

Green, R.H. 1984. Statistical and nonstatistical considerations for environmental monitoring studies. Environ. Monit. Assess., *4*: 293-301.

Green, R.H. 1986. Some applications of linear models for analysis of contaminants in aquatic biota. In: Statistical aspects of water quality monitoring. A.H. El-Shaarawi and R.E. Kwiatowski, eds. Elsevier, New York.

Green, R.H. 1987. Statistical and mathematical aspects: distinction between natural and induced variation. In "Methods for Assessing the Effects of Mixtures of Chemicals", by V.B. Vouk, G.C. Butler, D.V. Upton, D.V. Parke, S.C. Asher, eds., pp. 335-354. Wiley, Chichester UK.

Green, R.H. 1989. Inference from observational data in environmental impact studies: what is valid? what is possible? *47th Session* International Statistical Institute. Paris.

Green, R.H. 1993a. Relating two sets of variables in environmental studies. pp. 151-165. (Abstract)

Green, R.H. 1993b. Application of repeated measures designs in environmental impact and monitoring studies. Austr. J. Ecol., *18*: 81-98.

Green, R.H., J.M. Boyd, and J.S. Macdonald 1993. Relating sets of variables in environmental studies: the Sediment Quality Triad as a paradigm. Environmetrics, *4*: 439-457.

Green, R.H. and P.A. Montagna 1996. Implications for monitoring: study designs and interpretation of results. Can. J. Fish. Aquat. Sci., *53*: 2629-2636.

Green, R.H. and G.L. Vascotto 1978. A method for the analysis of environmental factors controlling patterns of species composition in aquatic communities. Water Research, *12*: 583-590.
Harris, R.J. 1985. A Primer of Multivariate Statistics. Academic Press, New York.

Hinch, S.G. and R.H. Green 1989. The effects of source and destination on growth and metal uptake in freshwater clams reciprocally transplanted among south-central Ontario lakes. Can. J. Zool., *67*: 855-863.

Hinch, S.G. and L.A. Stephenson 1987. Size- and age-specific patterns of trace metal concentrations in freshwater clams from an acid-sensitive and a circumneutral lake. Can. J. Zool., *65*: 2436-2442.

Hurlbert, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecol. Monogr., *54*: 187-211.

Hyland, J., D. Hardin, M. Steinhauer, D. Coats, R. Green, and J. Neff 1994. Environmental impact of offshore oil development on the outer continental shelf and slope off Point Arguello, California. Marine Environ. Res., *37*: 195-229.

Imlay, M.J. 1982. Use of shells of freshwater mussels in monitoring heavy metals and environmental stresses: a review. Malac. Rev., *15*: 1-14.

Jackson, D.A. 1993a. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology, *74*: 2204-2214.

Jackson, D.A. 1993b. Multivariate analysis of benthic invertebrate communities: the implication of choosing particular data standardizations, measures of associations, and ordination methods. Hydrobiologia, *268*: 9-26.

Jackson, D.A. 1995. PROTEST: A PROcrustean randomization TEST of community environment concordance. Ecoscience, *2*: 297-303.

Jackson, D.A. and H.H. Harvey 1993. Fish and benthic invertebrates: community concordance and community-environment relationships. Can. J. Fish. Aquat. Sci., *50*: 2641-2651.

Jongman, R.H.G., C.J.F. ter Braak, and O.F.R. van Tongeren 1987. Data analysis in community and landscape ecology. Pudoc Wageningen, The Hague.

Kennicutt, M.C.J., P.N. Boothe, T.L. Wade, S.T. Sweet, R. Rezak, F.J. Kelly, J.M. Brooks, B.J. Presley, and D.A. Wiesenburg 1996a. Geochemical patterns in sediments near offshore production platforms. Can. J. Fish. Aquat. Sci., *53*: 2554-2566.

Kennicutt, M.C.J., R.H. Green, P.A. Montagna, and P.F. Roscigno 1996b. Gulf of Mexico Offshore Operations Monitoring Experiment (GOOMEX), Phase 1:Sublethal responses to contaminant exposure -introduction and review. Can. J. Fish. Aquat. Sci., *53*: 2540-2553.

Kruskal, J.B. 1964a. Non-metric multidimensional scaling: a numerical method. Psychometrika, *29*: 115-129.

Kruskal, J.B. 1964b. Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. Psychometrika, *29*: 1-27.

Legendre, P. and L. Legendre 1997. Numerical ecology. Elsevier, Amsterdam.

Locke, A., W.G. Sprules, W. Keller, and J.R. Pitblado. 1994. Zooplankton communities and water chemistry of Sudbury area lakes: changes related to pH recovery. Can. J. Fish. Aquat. Sci., *51*: 151-160.

Long, E.R. and P.M. Chapman. 1985. A sediment quality triad: measures of sediment contamination, toxicity and infaunal community composition in Puget Sound. Mar. Poll. Bull., 16: 405-415.

Luoma, S.N. and E.A. Jenne 1977. The availability of sediment-bound cobalt, silver and zinc to a deposit-feeding clam. In: Biological Implications of Metals in the Environment. H. Drucker and R.E. Wildung, eds. U.S.N.T.I.S. Springfield (Conf.750929), pp. 213-231.

Manly, B.F.J. 1991. Randomization and Monte Carlo methods in biology. Chapman and Hall, London.

Manly, B.F.J. 1994. Multivariate statistical methods: a primer. Chapman and Hall, London.

Mantel, N. 1970. A technique of nonparametric multivariate analysis. Biometrics, *26*: 547-558.

Marcus, M.D. and L.L. McDonald 1992. Evaluating the statistical bases for relating receiving water impacts to effluent and ambient toxicities. Env. Tox. Chem., *11*: 1389-1402.

Metcalfe-Smith, J.L., R.H. Green, and L.C. Grapentine 1996. Influence of biological factors on concentrations of metals in the tissues of freshwater mussels (*Elliptio complanata* and *Lampsilis radiata radiata*) from the St. Lawrence River. Can. J. Fish. Aquat. Sci., *53*: 205-219.

Metcalfe-Smith, J.L. and R.H. Green 1992. Ageing studies on three species of freshwater mussels from a metal-polluted watershed in Nova Scotia, Canada. Can. J. Zool., *70*: 1284-1291.

Montagna, P.A. and D.E. Harper 1996. Benthic infaunal long-term response to offshore production platforms in the Gulf of Mexico. Can. J. Fish. Aquat. Sci., *53*: 2567-2588.

Nemec, A.F.L. and R.O. Brinkhurst 1988a. The Fowlkes-Mallows statistic and the comparison of two independently determined dendrograms. Can. J. Fish. Aquat. Sci., *45*: 971-975.

Nemec, A.F.L. and R.O. Brinkhurst 1988b. Using the bootstrap to assess significance in the cluster analysis of species abundance data. Can. J. Fish. Aquat. Sci., *45*: 965-970.

Olsgard, F., P.J. Somerfield, and M.R. Carr 1997. Relationships between taxonomic resolution and data transformations in analyses of a macrobenthic community along a established pollution gradient. Mar. Ecol. Prog. Ser. 1-9.

Orloci, L. 1973. Ranking characters by a dispersion criterion. Nat. Lond., *244*: 371-373.

Orloci, L. 1976. Ranking species by an information criterion. J. Ecol., *64*: 417-419.

Orloci, L. 1978. Multivariate Analysis in Vegetation Research. W. Junk, The Hague, 1-451.

Peterson, C.H., M.C.J. Kennicutt, R.H. Green, P.A. Montagna, D.E. Harper, E.N. Powell, and P.F. Roscigno. 1996. Ecological consequences of environmental perturbations associated with offshore hydrocarbon production: a perspective on long-term exposures in the Gulf of Mexico. Can. J. Fish. Aquat. Sci., *53*: 2637-2654.

Phillips, D.J.H. 1976. Common mussel *Mytilus edulis* as an indicator of pollution by zinc, cadmium, lead and copper. 2. Relationship of metals in mussel to those discharged by industry. Mar. Biol., *38*: 71-80.

Pielou, E.C. 1984. The interpretation of ecological data. Wiley, New York.

Pimentel, R.A. 1978. Morphometrics: the Multivariate Analysis of Biological Data. Kendall-Hunt, Dubuque, Iowa.

Reynoldson, T.B., R.C. Bailey, K.E. Day, and R.H. Norris 1995. Biological guidelines for freshwater sediment based on BEnthic Assessment of SedimenT (the BEAST) using a multivariate approach for predicting biological state. Austr. J. Ecol., *20*: 198-219.

Schonemann, P.H. and R.M. Carroll 1970. Fitting one matrix to another under choice of a central dilation and a rigid motion. Psychometrika, *35(2)*: 245-255.

Seber, G.A.F. 1984. Multivariate observations. Wiley, New York.

Smith, A.L., R.H. Green, and A. Lutz 1975. Uptake of mercury by freshwater clams (Family Unionidae). J. Fish. Res. Board Can., *32*: 1297-1303.

Somerfield, P.J., J.M. Gee, and R.M. Warwick 1994a. Soft sediment meiofaunal community structure in relation to a long-term heavy metal gradient in the Fal estuary system. Mar. Ecol. Prog. Ser., *105*: 79-88.

Somerfield, P.J., J.M. Gee, and R.M. Warwick 1994b. Benthic community structure in relation to an instantaneous discharge of waste water from a tin mine. Mar. Poll. Bull., *28*: 363-369.

Sprules, W.G. 1977. Crustacean zooplankton communities as indicators of limnological conditions: an approach using principal component analysis. J. Fish. Res. Board Can., *34*: 962-975.

Stewart-Oaten, A., W.W. Murdoch, and K.R. Parker. 1986. Environmental impact assessment: pseudoreplication in time? Ecology, *67*: 929-940.

Taylor, B.R. and R.C. Bailey. 1997. Technical Evaluation on Methods for Benthic Invertebrate Data Analysis and Interpretation. Prepared for Natural Resources Canada.

ter Braak, C.J.F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology, *67*: 1167-1179.

Underwood, A.J. 1981. Techniques of analysis of variance in experimental marine biology and ecology. Ann. Rev. Oceanogr. Mar. Biol., *19*: 513-605.

Underwood, A.J. 1992. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. J. Exp. Mar. Biol. Ecol., *161*: 145-178.

Underwood, A.J. 1993. The mechanics of spatially replicated sampling programmes to detect environmental impacts in a variable world. Austr. J. Ecol., *18*: 99-116.

Underwood, A.J. 1997. Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, Cambridge, UK.

Warwick, R.M. 1993. Environmental impact studies on marine communities: pragmatical considerations. Austr. J. Ecol., *18*: 63-80.

Warwick, R.M. and K.R. Clarke 1993. Increased variability as a symptom of stress in marine communities. J. Exp. Mar. Biol. Ecol., *172*: 215-226.

Wright, J.F., D. Moss, P. Armitage, and M.T. Furse 1984. A preliminary classification of running-water sites in Great Britain based on macro-invertebrate species and the prediction of community type using environmental data. Freshw. Biol., *14*: 221-256.

Wright, J.F. 1995. Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. Austr. J. Ecol., *20*: 181-197.

Yan, N.D. and R. Strus 1980. Crustacean zooplankton communities in acidic, metal-contaminated lakes near Sudbury, Ontario. Can. J. Fish. Aquat. Sci., *37*: 2282-2293.

Yan, N.D., W. Keller, K.M. Somers, T.W. Pawson, and R.E. Girard. 1996. Recovery of crustacean zooplankton communities from acid and metal contamination: comparing manipulated and reference lakes. Can. J. Fish. Aquat. Sci., *53*: 1301-1321.