Vol. 46: 213–226, 1988 | MARINE ECOLOGY – PROGRESS SERIES
Mar. Ecol. Prog. Ser. | Published June 30

Statistical design and analysis for a 'biological effects' study

K. R. Clarke[1], R. H. Green[2]

[1] Plymouth Marine Laboratory (West Hoe), Prospect Place, The Hoe, Plymouth PL1 3DH, United Kingdom
[2] Department of Zoology, University of Western Ontario, London, Ontario, Canada N6A 5B7

ABSTRACT: Statistical aspects of 'biological effects' field surveys are discussed, with particular reference to the GEEP Workshop. Recommendations are made on design criteria, for example, selection of sites and samples, and replication strategies (including formulae for sample size determination). The role of transformations is discussed, both for univariate sub-lethal response data and the multivariate data arising from benthic community studies. Statistical analysis is categorised into testing methods, for establishing biological differences between field sites, and descriptive techniques, for representation of those differences. The former includes a non-parametric randomisation test for use with site-species arrays and the latter a survey of various multivariate ordination and clustering methods. A final section outlines a procedure for comparison of different pollution indices, combining their power to detect specific contaminant inputs with their associated 'costs'

INTRODUCTION

A large number of statistical issues were raised in the planning of the GEEP Workshop, ranging from questions of sampling design for pollution studies, through methods of univariate statistical analysis on the resulting sub-lethal stress responses, and multivariate analyses of benthic community change, to techniques for comparison of the various pollution indicators. The intention of this paper is to describe some of the thinking behind the statistical design and analysis for the workshop and, more importantly, to elaborate on aspects relevant to future impact assessment programmes. Thus, while illustrative results are drawn both from the data of the Frierfjord/Langesundfjord survey and the Solbergstrand mesocosm study, the structure of the paper reflects statistical aspects of field studies rather than laboratory experiments.

Though much of what follows is relevant to any field study, attention is restricted here to a putative spatial pollution gradient examined at one point in time (as at the workshop), rather than to time series of observations where the significance of an impact is assessed in relation to temporal controls. The sections of the paper refer to the main statistical stages in such a study:

(a) survey design – the criteria for selection of samples, and extent and type of replication;

(b) pre-processing – the initial examination and possible transformation of data for conformity with the assumptions of statistical analyses;

(c) tests of null hypotheses of 'no biological differences' between sites on a contaminant gradient;

(d) descriptive and explanatory analyses, displaying the relationships between responses at each site and relating those changes to the contaminant gradient;

(e) retrospective assessment of the sensitivity of various response measures and analyses employed, as part of the continuous cycle of improving subsequent design.

A further categorisation needs to be made clear at the outset. Methods examined at the workshop fell broadly into 4 groups: biochemical, cellular, physiological and community studies (see sections of this MEPS SPECIAL). However, for the statistical discussion, only a broad dichotomy is needed, distinguishing the benthic faunal community analyses from the individual organism studies. This distinction is reflected both in the constraints on sampling design and in the different analyses required – community studies principally employ *multivariate* statistics whilst the sublethal stress responses are primarily *univariate*.

0171-8630/88/0046/0213/$ 03.00

DESIGN

In selecting sample areas it is usually crucial to choose one or more sites which are spatial 'controls', i.e. relatively unimpacted (reference) sites for which comparison can be made with the contaminant-affected site(s). It is true that certain measures of sub-lethal stress in individual organisms have been the subject of sufficient reseach for their values to be interpretable in an absolute rather than comparative sense. For example, 'scope for growth' in mussels (e.g. Widdows & Johnson 1988), a net energy balance arrived at by measurement of energy intake and losses of individual animals, takes values over a well-defined range, in which small or negative values are usually indicative of stressed populations. Similarly, for some methods of examining benthic faunal communities, it may be possible to detect stress in samples from a single site, rather than by comparison with a reference site. For example, Warwick (1986) proposes a comparison of the *k*-dominance curves for species abundance and species biomass at a single site, different relative positions of the two curves indicating disturbed or undisturbed communities (see also Gray et al. 1988). However, even in these cases, credibility would be greatly enhanced by demonstration of statistically significant differences in the chosen indicator (predicted a priori), between impacted and reference sites or between sites differing in degree of impact. Discussion of spatial (and temporal) controls in observational studies can be found in Green (1979).

Four other major design features are considered below, separately for individual organism responses and benthic community data:

(1) the desirability of selecting sites and faunistic samples such that 'nuisance' physical and biological variables are controlled within set limits (where it is known that variation within these ranges has little effect on the biological measures);

(2) the importance of proper replication at each site (with appropriate randomisation in sampling of the faunistic material);

(3) the importance of background data (preferably the collection of pilot samples) for selecting appropriate sites/samples, erecting suitable hypotheses about changes in biological measures and choosing the right level of replication to ensure these are adequately testable;

(4) the need to perform analyses 'blind' in order to minimise the dangers of self-fulfilling predictions.

Univariate data from sub-lethal responses

A simple maxim in experimental studies is to hold constant the values of any (nuisance) variables that are not of relevance to the treatment differences being investigated, thus increasing precision in the measured response. This can apply equally to field studies so that, for example, it would be advantageous to collect individuals of the target species within the same narrow size range from all sites. Other variables may not be possible to control, for example the changes in salinity that may accompany sites on a decreasing pollution gradient down an estuary; prior experimental and observational evidence plays an important role here in deciding whether such confounded variables can be discounted (e.g. scope for growth in mussels has been shown not to be sensitive to modest salinity changes, Widdows 1985).

It helps in the definition of what constitutes a 'site', and how to collect animals from it, to define the 'target population' which the sample animals are intended to represent. The objectives should make this clear; the intention is to demonstrate whether this defined geographical location is more impacted than a control location, by like-with-like comparison of a very specific biological effects measure. Thus, it is quite legitimate to postulate a narrow target population, for example all mussels in the size range (3.5,4.5) cm, of one sex, located at MLW etc. However, spatial definition of a site must remain broad, along 100 m (say) of shoreline. If the 'sampled population' is spatially more restricted than this, e.g. clumps of mussels are all taken from a single rock within 1 or 2 m of each other, some strong (and possibly unjustifiable) assumptions are needed to equate sampled and target populations. The risk here is obvious; all one may succeed in demonstrating is that mussels on a certain rock are significantly more impacted than those on another rock several kilometres away, leaving open the possibility that such differences could have been seen for a nearby rock also. A better strategy is therefore to collect individual animals (of desired size) across the full spatial extent of the site. Formal random selection is impractical, and largely unnecessary since the spatially stochastic distribution of populations will probably generate adequate randomness from evenly-spaced selection along the shoreline. Where randomisation can be used to good effect is when it is required to sub-sample from the full set of animals collected at a site, perhaps for separate biological measures or for chemical analysis of tissues. Interpretability is always improved by performing biological and chemical measures on the same individuals; whilst this is possible for larger organisms (e.g. fish) it may be impractical for other commonly used species (e.g. bivalves). Randomisation in the division of animals for different analyses at least ensures that they all unbiasedly address the same 'sampled' (and thus 'target') population.

The importance of the right amount and *kind* of

replication cannot be overstressed. It is usually right to aim for a balanced design, with equal numbers of replicate animals analysed at each site. Replicate readings can also be taken at different hierarchical levels and it is important to allocate effort across the levels efficiently. An example from the workshop concerns stereological measurement of cell tissue structure (Lowe 1988). On sections of mussel mantle tissue the volume fraction of nutrient storage (VCT) cells is determined by counting the proportion of points on an eyepiece graticule that fall on VCT cells (Lowe et al. 1982). The hierarchical levels here would be: the number of graticule points to count per field, the number of fields to observe per section, the number of sections to cut per animal and, finally, the number of animals to sample per field site. In a nested structure of this sort, the significance of an effect at any level is judged by comparison of the variance at that level with the variance at the level below (nested ANOVA); thus the presence of significant heterogeneity across sections within an animal, across animals within a site etc. can all be tested for, if there is enough replication at the lower levels. But this is largely irrelevant (there will almost always be significant animal-to-animal variation); the main purpose of the analysis is to examine variation between sites, and the appropriate replication variance to compare this with is that between animals. With typical levels of biological variability, no amount of effort in examining many sections, and many fields per section, will make up for a design in which there are only 2 animals per site; the top level ANOVA *F*-test will be based on few residual degrees of freedom and will lack power to detect site-to-site changes. In this context, Gundersen & Osterby (1981) discuss optimal allocation of effort.

In fact, the choice is often more complex since some techniques, particularly the biochemical and chemical analyses, involve replicate determinations on pools of animals rather than individuals; optimal selection of pool size is then also required. This too is tractable, as follows, at least for the case of a 2-level hierarchy in which several pools of animal tissue (taken at each site) are subsampled to give replicate assays.

For any particular field site (or experimental condition) define:

n = number of pools analysed,

p = number of animals in a pool,

r = number of replicate determinations on each pool,

σ_1^2 = variance from pool to pool (within a site) of the *true* pool means, and

σ_2^2 = variance of replicate determinations within a pool.

Then, the observed mean value $\bar{y}$ of the response variable, averaged over all pools and replicates at that site, has variance:

$$\mathrm{var}(\bar{y}) = [(\sigma_1^2/p) + (\sigma_2^2/r)]/n \quad (1)$$

The precision likely to result from proposed p and r values can be seen, for example bearing out the above comments about a law of diminishing returns as r tends to ∞. For evaluation, σ^2 values need to be estimated from current data. Defining V_1 as the *observed* variance between pool means and V_2 as the observed variance of replicate determinations within a pool (and noting that V_1 reflects contributions from both σ_1^2 and σ_2^2), these estimates are:

$$\hat{\sigma}_1^2 = V_1 - (V_2/r_0),\ \hat{\sigma}_2^2 = V_2 \quad (2)$$

where r_0 = number of replicate determinations in the current data.

Optimal future balance of effort between p and r depends on the structure of the 'cost' function. A reasonable description would be:

$$\text{Cost (man-hours)} = d_0 + d_1 n + d_2 np + d_3 nr \quad (3)$$

where d_3 = marginal cost (time) incurred to do an extra replicate determination, d_2 = cost of an extra animal dissection, d_1 = cost of preparing an extra pool, d_0 = fixed costs. In fact, only the ratio d_2/d_3 needs to be determined. A little algebra shows that, if the technique currently uses a pool size of p_0 animals, and r_0 replicate determinations are made, then a switch to p' and r' given by:

$$r' = [r_0 + (d_2/d_3)p_0][1 - \sqrt{K}]/[1 - K],$$
$$K = (d_2\sigma_1^2)/(d_3\sigma_2^2), \quad (4)$$
$$p' = p_0 + (r_0 - r')(d_3/d_2)$$

will reduce the $\mathrm{var}(\bar{y})$ to a minimum, for the same fixed number of pools *(n)* and the same total cost (effort).

An example from the Workshop is provided by cytochrome P-450 assays on mesocosm mussels (Livingstone 1988). Here, pool size was $p_0 = 6$ and $r_0 = 2$ replicate determinations were made. Estimated variances were $V_1 = 169$, $V_2 = 227$; time (d_2) per mussel dissection was about 0.6 min and operator time (d_3) for an additional replicate assay about 10 min, giving $r' = 1.9$, $p' = 7.1$. The closest integer solution, preserving fixed total 'cost', is $r' = 2$, $p' = 6$, demonstrating that the current balance of effort is optimal.

Returning to cellular-level responses, but retaining the simple two-level hierarchy, pool size is in effect constrained to $p = 1$, so optimal balance is now possible between n (number of animals) and r (number of measurements on each animal). This is again subject to fixed total effort, structured as:

$$\text{Cost (man hours)} = c_0 + c_1 n + c_2 nr \quad (5)$$

where again all that is needed is the ratio c_1/c_2 of the time for one extra animal dissection, and histological preparation, to the time for an extra replicate measurement on the resulting slide. If the current technique uses n_0 animals and r_0 readings, then it is optimal to switch to:

$$r^* = \sqrt{(c_1\sigma_2^2)/(c_2\sigma_1^2)} \quad (6)$$
$$n^* = n_0[1 + (c_2 r_0/c_1)]/[1 + (c_2 r^*/c_1)].$$

In the measurement of VCT volume fraction on a sample of Solbergstrand mussels (data taken prior to the workshop), $n_0=10$ mussels were examined and $r_0=5$ fields viewed for each individual (1 section only taken per mussel). These gave $V_1=30$, $V_2=25.5$; c_1 and c_2 were about 4 and 1.5 min respectively, giving $r^*=1.6$, $n^*=18$. The suggestion is that it would pay to increase the number of mussels at the expense of reducing the number of fields viewed, the best integer solution of $r^*=2$ and $n^*=16$ reducing overall variance by 20 %, for the same total effort.

The above formulae are concerned with correct allocation within an overall fixed effort; the question remains as to whether the total effort employed is sufficient to meet the objectives of the study. The choice of an appropriate n (number of mussels or pools at the top replication level) is a function of (a) how large a change (δ) in the biological response one wishes to detect, by comparison with the control, (b) the probability *(P)* with which the study should detect that level of change (the 'power' of the test, usually set at 0.9 or 0.95), and (c) the variance (σ^2) of the response for a single mussel/pool. (If a variance-stabilising transformation is required – see the next section – all these values should be defined on the transformed response scale). For a test of significance level $p < 0.05$, n should be chosen to satisfy:

$$n > (k+1) + \tfrac{1}{4}(k^2+1),\ k = (\sigma/\delta)^2|2 + \Phi^{-1}(P)|^2 \quad (7)$$

where Φ^{-1} = inverse of the unit normal distribution function, which is widely tabulated (e.g. for $P = 0.95$, $\Phi^{-1}[P] = 1.64$). Note that this is only an approximation to the correct result, the latter involving the more complex non-central t-distribution (less widely tabulated). However, the approximation is acceptable for $n > 4$ or 5, though an alternative is to read the precise values from a set of power curves (e.g. Bayne et al. 1981).

Good design therefore requires prior knowledge of the likely behaviour of the biological response (e.g. its sampling variance) for the field conditions and contaminant gradient expected. Prior information will also be needed to judge availability of organisms, and the biological and physical variables that it may be important to control (e.g. size range, degree of exposure to wave action etc.) and so define properly the sample sites and target populations. A pilot sampling programme is often essential and always desirable.

Finally, it is usually desirable for analyses to be carried out 'blind', i.e. such that the experimenter performs the preparation and measurement phases without knowledge of the source of his material. This was true of nearly all benchwork performed at the workshop, on both field and mesocosm samples, and was not difficult to arrange in practice. It is recommended as a simple safeguard against (unconscious) biases on the part of the experimenter. However objective the protocols for a technique may be, there are usually stages in its execution that may call for some judgement (e.g. the decision to reject a replicate because of a suspected analytical failure, the selection of microscope fields in histological and histochemical work, the relative effort put into taxonomic identification in different benthic faunal samples etc.). It is wise not to risk selection biases where these can be simply avoided by recoding of material before analysis.

Multivariate data from benthic community studies

Much of the above discussion is equally apposite to the sampling of benthic faunal communities. Thus it is important to have replication at each site, those replicates to be properly representative of the 'target population' (i.e. community) of the area of interest. As with the littoral populations, this area should not be defined to be too spatially compact. If it is, the replication variance between sample cores (or grabs) represents no more than local sampling fluctuation (termed 'pseudo-replication' by Hurlbert 1984) rather than the relevant level of between-area variation within that site. There is a strong analogy with the earlier discussion on different levels of replication in measurements of cellular response. Tests of between-site differences rely on adequate 'top level' replication.

Control of 'nuisance' physical variables, in selection of sites to be compared, is another important aspect. The experience of the Workshop suggests that this can be more difficult for benthic sampling than individual organism studies. For example, sediment grain size and water depth are known to be important determinants of community structure; ideally all selected sites should be chosen to have the same narrow range of median particle size and depth. Where no choice is possible of the impacted areas to be studied, and the depth or particle size differs in a way that is likely to be important between them, then separate control sites may need to be selected for each (matched as closely as possible in terms of physical variables). Obviously, careful pilot sampling is called for here.

An alternative scenario is that sediment structure is not grossly different in mean value between sites but varies in an important way within the boundaries of each site. The variable cannot be controlled by selection, so must be controlled by statistical means (Cochran 1983, Chapter 1), i.e. by regressing out its effect with an analysis of covariance or its multivariate equivalent. This is discussed further under 'Testing for structure' but the implication for design is that values of all relevant variables need to be determined from each replicate sample, thus matching the biological, physical and chemical data as closely as possible.

PRE-PROCESSING OF DATA

Both univariate data on sub-lethal responses and multivariate species counts from benthic communities may require some 'pre-processing', i.e. transformation or selection/pooling of variables. However, there can be a number of different motives for employing transformations.

Univariate responses

For univariate measures, the aim is usually to allow interval estimation and tests (*t*-tests, ANOVA etc.) to take place under the standard assumptions of approximate normality and of equality of variance between responses at different sites. A common violation is for the variance to increase with the mean, often associated with a right-skewed distribution of replicates at a single site. A simple power transform $(y^{\lambda}-1)/\lambda$, $0 \leq \lambda \leq 1$, might then be appropriate (Box & Cox 1964, 1982). λ controls the severity of the transform, from no transform at all ($\lambda=1$), through square root and 4th root ($\lambda=0.5$ and 0.25) to logarithmic ($\lambda \rightarrow 0$). Optimal choice of λ is possible but this is usually unduly precise; for example, the 'best' transformation for a specific response variable would then vary slightly from one data set to another. Usually adequate is a simple plot of standard deviation against mean, a straight line suggesting (from 'Taylor's power law') that variances will be stabilised by the log(y) transform, whereas a straight line for variance against mean indicates the less severe $\sqrt{y}$. Equivalently, rounding the slope of a plot of log(standard deviation) against log(mean) to the values (0, 0.5, 0.75, 1) suggests transformations (none, $\sqrt{}$, $\sqrt[4]{}$, log) respectively.

Fig. 1 displays 2 examples from the workshop, involving scope for growth determinations in mussels (Widdows & Johnson 1988) and induction of the enzyme EROD in flounder (Addison & Edwards 1988). The crosses denote replicate readings (n = 16 ind. for the former and n = 11 or 12 for the latter) for Field sites 1 to 4 along the Langesundfjord gradient, from reference to contaminated sites. Note that for scope for growth the variance is stable over quite a wide range of mean response and no transformation is needed. By contrast, EROD activities have standard deviation closely proportional to the mean, together with a right-skewed error distribution; a log transform succeeds well in inducing normality and stabilising the variance.

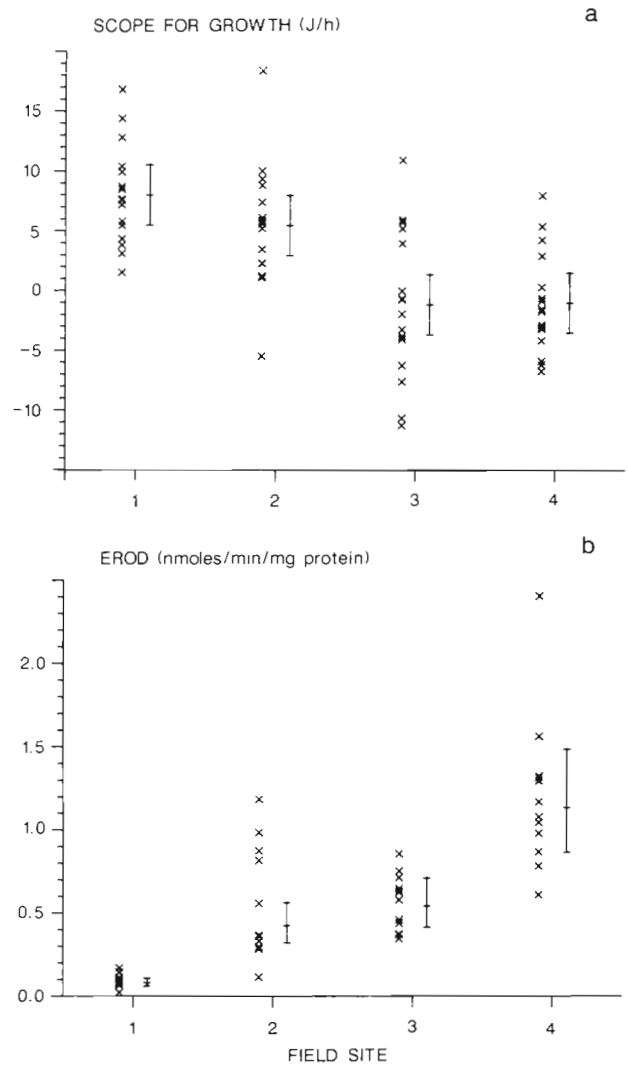

Fig. 1. (a) 'Scope for growth' determinations in mussels *Mytilus edulis*, Widdows & Johnson (1988), (b) EROD activity in flounder *Platichthys flesus*, Addison & Edwards (1988). Data from Field sites 1 to 4 in Langesundfjord; total concentration of selected PAHs in whole mussel tissue is 2.2, 5.9, 11.4 and 15.5 $\mu g\ g^{-1}$ respectively. Crosses: replicate mussels; bars: 95 % confidence intervals for mean response based on pooled SD (back-transformed, for EROD only, from log scales)

Multivariate data

For benthic species abundance arrays, hypothesis testing methods which are multivariate extensions of classical univariate methods (e.g. MANOVA) usually assume errors that are independent, normal and homogeneous. The models are usually linear and additive. There is a consensus that univariate methods are fairly robust to violations of normality (though less so to gross differences in variability, especially where replicate numbers are not balanced across sites). Much less is known about robustness in the multivariate case, though what work there is (e.g. Mardia et al. 1979) suggests a similar pattern: tests concerning differences in means are reasonably robust to non-normality, those concerning variances and covariances are not. Only

proper sampling design can ensure independent errors, whereas it may be possible to achieve normally distributed, homogeneous errors and linear additive relationships by applying an appropriate transformation. Tests for detecting departures from multivariate normality exist (e.g. Mardia et al. 1979); whilst it is theoretically not sufficient to demonstrate 'marginal normality' (the separate component variables are univariate normal), this is a good start in practice. A multivariate extension of the Box & Cox (1964) procedure for choosing suitable power (or logarithmic) transformations is given by Andrews et al. (1971).

Logarithmic transformations are widely used and there are good reasons for this: all variables are put onto a common scale of variation (percentage variation) regardless of the original units of measurement, and it is true that population density does tend to vary spatially and temporally on a percentage rate of change basis. Variance in $\log(y)$ corresponds to coefficient of variation in y so that significant heterogeneity of covariance matrices, for example in a MANOVA, can be simply interpreted as differing percentage variation. These appealing properties theoretically disappear when using the transformation $\log(1+y)$ rather than $\log(y)$, as one is often forced to do for species abundance data, in which zeros are almost always present. Further, $\log(1+y)$ can be slightly unsatisfactory when y is not species abundance but, say, species biomass. The modified transformation is affected by a change of scale (biomass cm^{-2} of surface rather than biomass m^{-2}). Of course, one would normally change the location shift parameter (using $\log[0.001+y]$ say) but there is a degree of arbitrariness in its choice. For this sort of reason a comparable power transformation, like $\sqrt{\ }$, is sometimes advocated (Field et al. 1982), though in practice the two transformations are rarely distinguishable.

A further important consideration is that, if all species are included, species abundances or biomass arrays are usually very sparse, the predominant entry being zero. Thus, approximate multivariate normality can only be attained by a transformation coupled with a substantial reduction in species considered, to the most abundant ones. This is also usually required in order that there is some chance of validity of the 'asymptotic' distributions of the relevant multivariate test statistics; the total number of observations (n samples $\times$ p species) should be 'large' in relation to the number of mean, variance and covariance parameters that need to be estimated ($2p + p[p-1]/2$). Alternatively, one might achieve reduction by pooling species into higher taxonomic categories. The GEEP Workshop results contain examples where counting at higher taxonomic levels does not significantly degrade the ability to discriminate field sites (Warwick 1988). There may therefore be advantages in a pooling strategy both from a statistical and practical viewpoint, bearing in mind the lower levels of taxonomic expertise required.

Analyses employing 'classical' hypotheses tests, under assumptions of multivariate normality, are by no means commonly used. Much more widespread are a variety of descriptive clustering and ordination techniques which are not based on underlying distributional assumptions (and largely lack a framework for hypothesis testing, in consequence). Field et al. (1982) describe a strategy of this sort for species abundance data, based on hierarchical, agglomerative, group-average clustering and non-metric multi-dimensional scaling (MDS). In spite of the lack of model-based assumptions, transformations still play an important role, that of determining the relative weight given to rare and common species in assessing differences between samples. Either explicitly or implicitly many of these 'ad-hoc' multivariate methods take as their starting point a triangular matrix of similarities (or dissimilarities) in species abundances between every pair of samples. These can be correlation-based measures, but often other measures are appropriate. In sparse matrices, where many of the p species are jointly absent from any 2 samples whose similarity is being calculated, correlation can be unsatisfactory; it is arguably counter-intuitive for similarity between 2 samples to be increased by addition to the species list of a species not present in either sample. One coefficient which avoids this problem is the 'Bray-Curtis' dissimilarity (Bray & Curtis 1957), defined as the absolute differences between the (possibly transformed) species count for two samples, summed over all species, and then divided by the total count over both samples and all species.

The effects of the power transformations discussed earlier are particularly clear for this coefficient (though broadly similar conclusions will hold for a wide range of measures). No transformation ($\lambda = 1$) will generally mean that only the few most numerically dominant species will contribute anything at all. Whilst this is likely to make it easy for an ordination to reflect faithfully in a 2-dimensional plot all the information in the dissimilarity matrix, it is certain to be very susceptible to the typically large absolute fluctuations in counts of the numerical dominants, and will have no chance of eliciting a structure where differences are in the medium-abundance or rare species. A mild transform such as the square root ($\lambda = 0.5$) will place most emphasis on the numerical dominants whilst not ignoring the medium-abundance species, whereas the 4th root ($\lambda = 0.25$) and log ($\lambda \rightarrow 0$) will further reduce the differential effects of dominant in relation to less dominant species and begin to differentiate between sites with many and few rare species. The logical

endpoint of this process is to consider for each species only whether it is present or absent in a sample (this is simply a transformation like any other). In fact, the need to use the $\log(1+y)$ transform rather than $\log(y)$ slightly distorts this transformation sequence, the general effect of $\log(1+y)$ being intermediate between $\sqrt{}$ and presence/absence for moderate or large counts but less severe than $\sqrt{}$ in accentuating the difference between counts of 0 and 1.

So, as the severity of transformation increases, more species come into play in determining dissimilarity, thereby tending to increase the dimensionality of the ordination space in which the samples can be placed in 'true' relation to each other. It should therefore be expected that the difficulty of ordinating the samples in a reduced space (usually 2-dimensions) will increase through a transformation sequence. This is borne out by meiofaunal analyses from the workshop mesocosm experiments (e.g. Fig. 1 of Warwick et al. 1988), where such a sequence gave rise to steadily increasing 'stress values' for the MDS ordinations.

TESTING FOR STRUCTURE

The objective of a biological effects study of this type is to describe response differences observed between sites and then attempt to relate these to some measured or inferred contaminant gradient. However, there is an obligatory first stage, that of demonstrating that differences of some sort genuinely exist before setting out to describe them. Descriptive multivariate methods in particular can easily fall into the trap of ignoring this stage; a hierarchical cluster analysis will always find clusters at some level of similarity, even from a set of random numbers! Significance testing thus plays an important role in this preliminary stage.

Univariate responses

For a sublethal response measured on a number of replicate animals at each of a number of sites, a global significance test for site differences is just the standard 1-way analysis of variance (ANOVA), provided a transformation has succeeded in (roughly) stabilising the variance across sites and (very roughly) inducing normality in the response. Where this is not possible, there is an equivalent non-parametric test – the Kruskal-Wallis test for a 1-way layout (e.g. Siegel 1956). Notice the implication that if these global tests fail to reject the hypothesis of no significant difference between sites (at the $p < 0.05$ level say) then no further analyses should be done. Such a policy diminishes the risk of the spurious conclusions which can occur when pairs of sites are selected a posteriori, for comparison by a standard t-test, the Type I error not being controlled at 0.05 because of the large number of pairwise comparisons that have been performed, either explicitly or implicitly.

There are a number of 'multiple comparison' tests which attempt to control the Type I error rate to 0.05 over all such a posteriori comparisons, e.g. the Tukey T and Scheffé S tests (Scheffé 1959). The latter, which can be used when there are unequal numbers of replicates at each site, also has the attractive property that it will detect at least one significant difference between the site means (in some combination) if *and only if* the overall ANOVA F-test rejects the hypothesis of equality of site responses. The Tukey T-test (in its simple form) is restricted only to pairwise comparison of sites and to balanced numbers of replicates. However, it does generally have greater power than the Scheffé S test so that a situation can occasionally arise in which the global F-test fails to find differences but the follow-up T-test would. This is only paradoxical if one (mistakenly) regards significance testing as an exact science, instead of a general guide to the approximate truth of hypotheses. Perhaps the most consistent approach to take in this case is that suggested above: regard a non-significant ANOVA F-test as a 'red light', stopping progress to further tests.

For the data of Fig. 1, follow-up tests verify the clear picture of no differences in scope for growth between Field sites 1 and 2, or between Sites 3 and 4, but a large difference between the 2 sets. For EROD data, Sites 2 and 3 are the only two not to differ significantly in pairwise comparisons.

Multivariate data

The multivariate case, arising from species abundance matrices (or from simultaneous examination of multiple sub-lethal responses), must again be divided into whether or not the data can be transformed to approximate multivariate normality, with sufficiently many samples by comparison with the number of species to validate classical theory. If so, there exists a 1-way MANOVA analysis (e.g. Mardia et al. 1979) for testing differences between sites, by comparison with variation and covariation observed between replicates within a site, exactly analogous to the univariate 1-way ANOVA. The test corresponding to the ANOVA F-test is known as Wilk's lambda; an alternative statistic likely to have greater power for detecting a gradation of sites is 'Roy's greatest root' (Seber 1984). Pairwise differences between sites can be tested by Mahalanobis' distances (Seber 1984). (Note that the latter are not 'multiple comparison' tests and do not

control the overall Type I error rate.) If the null hypothesis that sites have the same species composition is rejected, then canonical discriminant analysis (CDA) can provide a reduced-space description of species' contributions to site differences. Examples of the use of these test statistics at the workshop can be found in Gray et al. (1988) and Warwick et al. (1988).

An alternative parametric approach sometimes suggested for matrices of species counts is the log-linear model (Fienberg 1970). The assumptions are that counts of the number of individuals of any particular species found in replicate cores (or grabs) at one site will have a Poisson distribution, and that the variation of these counts from core to core within a site will be independent of the variation for any other species (i.e. the data is not genuinely multivariate, only multi-category). Provided that rarer species are again deleted, so that 'large sample' likelihood theory can be invoked, the log-linear model provides a test of whether the species composition changes across sites, by testing for a significant interaction in the 2-way layout. If no change is detected in composition, a test of the site main effects will examine whether there are changes in absolute numbers. However, there is considerable empirical evidence that the Poisson model is often inadequate. This can be examined by calculating the among-core variances for any particular species at each site; for a Poisson distribution these should equal the respective among-core means. In fact, this will only happen if the individuals of a species are distributed randomly and independently throughout the area represented by that site (technically, if they form a spatial Poisson Process, e.g. Diggle 1983). It is much more common for species to be spatially clustered or for their mean density to be locally variable because of small-scale environmental variation. Replicate field macrofauna data from the workshop bear this out. The variance-to-mean ratio of the four replicate grab counts at each site/species combination was almost always in excess of 1, its median value being 4.1 (quartiles 2.6 and 8.9) over the 50 site/species combinations displaying greatest total abundance. Log-linear models were therefore not used at the workshop.

Most descriptive multivariate analyses, for example hierarchical clustering, principal co-ordinate analysis, multi-dimensional scaling (MDS) etc., make no parametric assumptions at all. What are required here are tests for the presence of structure which make similarly few model assumptions. This is a neglected area, though one of us (KRC) has advocated such a test based on the principles of permutation and randomisation tests (Hope 1968). It operates on the triangular matrix (described in the last section) whose entries are the similarities or dissimilarities in species abundance or biomass calculated between every pair of benthic samples. Such matrices are the starting point for many descriptive analyses, both in clustering and ordination. Even the 'classical' technique of correlation-based principal components involves an implicit dissimilarity matrix of simple Euclidean distances between replicates (after normalisation), thus allowing an alternative non-parametric test for site differences to the normality-based MANOVA tests discussed above (see Gray et al. 1988, for a comparison).

The test is very simple, in concept. Assume that the n samples consist of r replicate cores at each of k sites. Under the null hypothesis of no between-site differences, one could allocate the r 'labels' for Site 1 at random to any of the $k \times r$ cores, Site 2 labels to a further r cores at random (without replacement) and so on. The similarity matrix then constructed between the newly labelled cores will clearly be some permutation of the entries in the original matrix. One can then construct a test statistic likely to reflect the joint similarity of replicates within a site, contrasted with the similarities between sites, and calculate this statistic for the original data and each of a large number of random relabellings. If it is more extreme for the original data than for (say) 95 % of the random relabellings then the null hypothesis is rejected by a $p < 0.05$ randomisation test. (Alternatively, if the number of sites and replicates is small, one might be able to enumerate all possible relabellings and so construct an equivalent permutation test).

In order to make the test as non-parametric as possible, and bearing in mind that non-metric MDS, one of the most powerful ordination methods (Kruskal & Wish 1978), relies only on the rank order of similarities in the original matrix, it is desirable that the test statistic should be a function only of these ranks. A natural choice is the difference between the average rank similarity 'between replicates within a treatment' and 'between replicates in different treatments'. It can be standardised so that a value of 0 reflects the null hypothesis of no site differences, and +1 corresponds to an alternative in which all replicate cores within a site are more similar to each other than any replicates across sites. The test (termed '1-way ANOSIM', by analogy with 1-way ANOVA) is straightforwardly generalised to the case where there are unequal numbers of replicates at each site. If global differences are found, it can be followed by pairwise comparisons of sites, using precisely the same randomisation/permutation principles, provided there are sufficient replicates at each site (4, in the balanced case) to generate a large enough set of possible permutations. (Note that these are the analogue of pairwise t-tests in the univariate case and suffer from the same dangers of repeated comparisons.)

Benthic data from the workshop can be used to

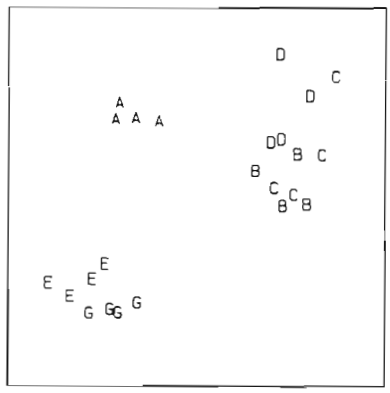

Fig. 2. Non-metric multi-dimensional scaling (MDS) plot in 2 dimensions for benthic macrofaunal data (Gray et al. 1988) from 4 replicate grabs at Sites A to E, G in Langesundfjord/ Frierfjord. Species abundances were $\sqrt[4]{\ }$-transformed and between-sample similarities calculated with the Bray-Curtis coefficient. 'Stress' for the MDS is low, at 0.10

illustrate the test. Fig. 2 displays an MDS ordination for macrofaunal abundances from the Frierfjord Sites A to E, G (Gray et al. 1988) with 4 replicate grabs per site. The data was 4th root transformed and Bray-Curtis similarities calculated. An ANOSIM test on the whole data is unnecessary – 3 groups of sites stand out as clearly different. However, the similarities for Sites B to D can be tested on their own; the statistic takes the value 0.45 and the hypothesis of 'no site differences' is rejected by a $p < 0.01$ randomisation test. Pairwise tests show that Site D differs from the other 2 sites, B and C being indistinguishable. (The pairwise comparisons are permutation tests, with minimum attainable significance level 0.03, there being only 35 distinct possible relabellings).

Effects of 'nuisance' variables

So far, discussion has concerned testing for any structure in the biological responses for different sites, however induced. In order to relate effects more closely to contaminant levels, it may first be desirable to remove effects of physical or biological 'nuisance' variables not controlled in the design phase. For example, in collecting certain species it may not be possible to select individuals from a narrow weight range. However, weight-induced differences in the biological response can be corrected for by a standard analysis of covariance (ANCOVA). This consists of linear regressions (possibly after transformation) of the response on the weight, usually fitting common slope but different intercepts across the sites. A test of equality of intercepts is then the global test of any between-site biological differences.

This univariate ANCOVA approach fails if each site exhibits a different, and narrow, range of animal weights; between-site differences are then totally confounded with weight differences. The confounding can only be removed by inputting strong prior information, e.g. a known value for the common slope in the response-versus-weight regressions. This simple illustration carries over to the multivariate case. If classical multivariate normal assumptions are permissible for species abundance data (appropriately reduced and transformed) then it may be possible to exploit the analogous multivariate analysis of covariance (MANCOVA) to remove any between-site differences caused by uncontrolled differences in, for example, sediment grain size. However, the univariate analogy demonstrates that this will only be possible if (a) physical data is available for each replicate core (or grab), and (b) variation in the physical data is large enough for there to be overlapping values between sites.

When multivariate normal assumptions are not valid, and description is to be via clustering or ordination techniques, no equivalent structure exists for neutralising an uncontrolled nuisance variable, though its effect may be seen as an axis in an ordination plot, and informally distinguished from contaminant-induced effects along other axes. There is therefore strong motivation for controlling important nuisance variables. Of course, in some environments, a major physical determinant of community type (e.g. sediment structure or water depth) will change systematically in line with an anticipated contaminant gradient, with little variation between replicates at a site. The confounding is then total and convincing statistical analysis impossible under any assumptions; such designs should be avoided.

DESCRIPTION OF STRUCTURE

Univariate responses

When a univariate sub-lethal response is demonstated (by ANOVA) to vary significantly between sites, the standard procedure is to compute 95 % confidence intervals (CIs) for the mean response at each site, based on a pooled estimate of standard deviation across all sites; these means and CIs could usefully be plotted (as y) against an x axis representing the contaminant gradient in some form. If a transformation has been performed prior to the calculation of means and the ANOVA, it is often desirable to back-transform to the original y scales for this plot. For a transform $y^* = \log(y)$, the reverse transform for the mean $\bar{y}^*$ will be $\exp(\bar{y}^*)$, and similarly back-transforming the endpoints of the 95 % CI on the log scale will give a 95 % CI on the original scale. This interval will not be symmetric about the back-transformed mean but this is

to be expected – a transform would have been required, in part, because of the lack of symmetry of the variation on the original scales. Fig. 1 illustrates this in its display of 95 % CIs for mean 'scope for growth' and EROD responses, in the latter case back-transforming from a log scale. The reason for emphasising back-transformation is that, in this descriptive phase, the statistical significance of a between-site difference is less important than its biological significance. It may often be easier to think in terms of original scales when assessing the practical significance of changes.

Linking to the contaminant gradient. In relating observed biological responses to chemical causes, it must be accepted that field studies of a contaminant gradient down an estuary (for example), will always involve a large number of chemical compounds; most of these will covary very precisely, as point contaminant inputs are steadily diluted. There will be little prospect of discriminating the effects of particular contaminants by purely statistical means (such as a multiple regression of biological response on a suite of chemical data). This is compounded by the fact that, typically, the biological responses are generalised sub-lethal stress measures, responsive to a wide variety of pollutants. Which variable(s) to display as the x-axis contaminant gradient in the above plots of response means (and CIs) is therefore a decision for the biologist, but the choice would usually reflect the closest possible coupling of contaminant availability to the organism with observed biological effect. Thus, scope for growth in mussels might be related to tissue concentrations of some total of polyaromatic hydrocarbons (= 2.2, 5.9, 11.4, 15.5 $\mu g\ g^{-1}$ for the 4 sites in Fig. 1a, see Appendix 1, Table 4). If there are a minimum of 4 or 5 sites along a well-spaced contaminant gradient x-axis, then it may be possible to fit a dose-response curve. This could be approximately linear or, perhaps over a wider range of doses, a sigmoidal curve such as the 4-parameter logistic:

$$y = \epsilon + (\delta - \epsilon)/[1 + \exp(-\alpha - \beta x)], \qquad (8)$$

(or its converse), where x is often taken as log(dose). This model is non-linear in the 4 parameters, so fitting would be by some iterative non-linear least squares technique, such as the modified Marquadt alogorithm (Nash 1979). Here, δ and ϵ represent the maximum and minimum responses, with the other 2 parameters controlling the dose at which 50 % of maximal response is achieved (the equivalent of the LC50 in lethal toxicity studies) and the dose range over which the response effectively changes.

Multivariate data

Many ways have been proposed of visualising the structure in a species abundance/biomass matrix. Three areas are distinguished here: ordination, clustering and a set of important special techniques. Broadly speaking, an ordination is an attempt to present a picture of the relationship between the samples, in terms of their similarity of species abundance or biomass (in either absolute or compositional form). In this picture, preferably 2-dimensional, the relative distance apart of any pair of samples is intended to reflect their relative dissimilarity. By contrast, cluster analysis attempts to form discrete groupings of samples, where samples within a group have a more similar species composition than those in separate groups. Ordination and clustering are not methods in competition with each other, though this has been a common assumption in the past. It is often a good strategy to do both and then plot the samples in the ordination space, with cluster membership indicated appropriately – see for example Fig. 8 of Warwick et al. (1988). In general, neither ordination nor clustering techniques use knowledge about the source of each sample. Thus, one can examine the outcome for evidence that samples within a site are placed nearer to each other in the ordination, or cluster more often in the same groups, than would be expected by chance; of course, this is what the ANOSIM test of the previous section does, more formally.

No attempt will be made here even to list, much less discuss, all the possible ordination and clustering methods, though some of the more useful techniques will be outlined.

Ordination. This can be defined as an analysis of an n samples by p species matrix whereby a new set of variables is found, numbering less (usually much less) than p, which optimally predicts the structure in the relationships among the original p variables. Ordination methods differ from each other in (a) the optimality criterion and (b) how the ordination algorithm 'finds' the new axes which represent the new variables. Principal components analysis (PCA, Seber 1984) maximizes the amount of variation accounted for by the new axes, and proceeds by way of an eigenanalysis on the p-by-p correlation (or covariance) matrix. The new axes are uncorrelated. PCA is simple to perform and does a good job within its limitations, though the new axes are rarely interpretable as simple environmental factors 'causing' the structure in the species-abundance data. Principal coordinates analysis (PCoA, Gower 1966) starts with an n-by-n matrix of 'distances' (dissimilarities) among the samples, and then proceeds as in the PCA. The same result as for the PCA can be obtained if the distance is defined appropriately There are 2 potential advantages of PCoA over PCA, that the eigenanalysis is easier to do if $n < p$, and that one is free to choose any of a large number of possible distance measures, using one appropriate to the data and the objectives.

Multi-dimensional scaling (MDS) finds a specified number of new axes which attempt to preserve some relationship among the between-sample distances, in the case of non-metric MDS their rank order (Kruskal & Wish 1978). The latter was mentioned in the previous section, and its application to the macrofaunal samples from Frierfjord/Langesundfjord is illustrated in Fig. 2. The attraction of non-metric MDS lies in its dependence on rank rather than quantitative values in the between-samples dissimilarity matrix; it uses only statements of the form 'Sample A is more similar to Sample B than it is to Sample C' and constructs a 'map' of the samples, in 2 dimensions say, which attempts to satisfy all such conditions. The extent to which this has been achieved is given by a 'stress' statistic, low values (< 0.1, say) indicating success. It is clear from the type of input that the final plot will have arbitrary orientation and scale. Non-metric MDS is an iterative procedure and more computationally demanding than PCA or PCoA (much more than 100 samples is prohibitive).

Finally, reciprocal averaging (RA, Hill 1973) and correspondence analysis (CA) are ordinations of count data, proposed initially for 'contingency tables' obeying the assumptions of a Poisson error distribution (see the discussion on log-linear models in the previous section). Various algorithms exist, including one for 'detrended correspondence analysis' (DECORANA, Hill & Gauch 1980), but a simple RA-CA type solution can be obtained by doubly-standardizing (both rows and columns) an n-by-p matrix of counts, and then subjecting it to a PCA. The endpoint of a correspondence analysis is a simultaneous display of both the relation of the samples to each other (in terms of their species composition) and the relation of the species to each other (in terms of their degree of co-occurrence in these samples). Most of the other ordination (and clustering) methods have some associated way of identifying which species have the major responsibility for the observed pattern of the samples.

A related technique is that of 'best variable subset' selection (Orloci 1978). This attempts to select the subset of species having the maximum predictive information about the full species set (for any subset of that size). The subset will tend to have low redundancy, i.e. to contain species which have low correlations with each other. The result is often similar to that of a PCA, except that each component will be represented by a single species, which aids interpretation. The disadvantages are that the components will not be completely independent, and the amount of structure accounted for by a given number of components will not be as high as in PCA.

Though the above discussion concentrates on the use of multivariate techniques on species abundance/biomass arrays, the methods are obviously more widely applicable. For workshop data, they were used to describe the field and mesocosm contaminant gradients, via a site PCA in which species counts were replaced by values of chemical variables (metals, PAHs); see Figs. 14 and 18 of Gray et al. (1988). ANOSIM (or MANOVA) tests can then determine objectively whether a contaminant gradient is present (which proved not to be the case for the mesocosm sediments, in spite of the dosing). Gray et al. (1988) also employ PCA to summarise the information in the different diversity indices that can be computed from the site/species arrays. Here again a site ordination is performed but with the different diversity values replacing species counts (Fig. 9 of Gray et al. 1988). This can be a useful counterweight to the dubious practice of calculating a large number of different diversity indices; here a few simple indices define the diversity relationship between the sites, and this can be adequately displayed in 2 dimensions. (It is less clear cut than the relationship defined by species counts, Fig. 2 of this paper). Adding additional diversity measures to the PCA did not force a higher dimensional solution nor alter the 2-dimensional picture in any way.

Clustering. This can be defined as an analysis on an n-by-p data matrix whereby a partitioning of the n samples into subsets is found, numbering less (usually much less) than n, such that the relationships among the subsets optimally predict the relationships among the original samples. Clustering methods differ from each other in (a) the optimality criterion and (b) how the clustering algorithm 'finds' the sample subsets, but clustering strategies are more diffuse than for ordination and less easily summarized. They are best defined by a number of dichotomous choices regarding strategy. A method may be hierarchical or non-hierarchical. If the former, then the algorithm may be agglomerative (repeatedly fuses samples or groups until all samples are in one group) or divisive (repeatedly divides what is initially one group containing all the samples). The fusions or divisions are based, respectively, on some n-by-n distance matrix or some p-by-p similarity (e.g. correlation) matrix. Divisive methods are usually monothetic (the next division is based on a statistic related to one variable), whereas agglomerative methods are usually polythetic (the next fusion is based on a statistic calculated from all variables). In hierarchical methods, samples or clusters of samples are compared via a particular linkage strategy: nearest neighbour, furthest neighbour, group average of cluster members, etc. The clusters produced can be overlapping or discrete; usually hierarchical methods produce discrete clusters and non-hierarchical methods overlapping clusters. Finally, clustering can be constrained by external criteria (e.g. that any cluster must contain samples contiguous in space), or it can be

unconstrained and then evaluated by mapping the clusters (and seeing whether samples are contiguous). More recent developments in clustering methods combine strategy choices that do not usually go together, for example polythetic divisive clustering (Lefkovitch 1979). For an introductory text on clustering see Everitt (1974).

Other methods. There are a large number of other techniques for comparing and summarising the information in benthic faunal samples, chiefly characterised by their exploitation of some aspect of community structure which is independent of the particular species involved. The summaries are either univariate statistics, such as the diversity indices mentioned earlier (e.g. Pielou 1975), or some estimated probability density (or distribution) function such as the sample 'species abundance distribution' (e.g. Engen 1978). This latter is usually presented as a histogram of the numbers of different species represented by x individuals in the sample, the x scale being conventionally grouped into $\log_2$ classes (see Fig. 5 of Gray et al. 1988). Other structural aspects of communities studied include the 'species biomass distribution' (a similar construct but with the x axis being log weight or size classes) and 'biomass-size spectra' (the x axis again consists of increasing size classes but the y axis is total biomass of all organisms in each size class, of whatever species), see Figs. 1 to 3 of Schwinghamer (1988). Another graphical indicator of community change, the 'ABC method' (Warwick 1986), contrasts the pattern of biomass-dominants and numerical-dominants in a sample (see Fig. 13 of Gray et al. 1988).

There is a large literature on the mathematics of this subject, particularly on diversity indices and 'species abundance distributions'. Nonetheless, the derivation of sampling properties of some of the graphically-based community measures – a necessary prerequisite for their use in categorising pollution-induced change – is still very much an area of current statistical research and will not be considered further in this paper.

Linking to the contaminant gradient. Much of the previous discussion on relating sub-lethal responses to specific chemical causes applies equally to benthic community data. (Of course, when the community structure is summarised in a univariate measure, such as a single diversity index, precisely the same arguments apply). Typically there will be a large number of closely correlated chemical compounds involved in any sediment contaminant gradient and little discriminating power is to be expected in distinguishing specific chemical causes from biological effects; rather, one would select 2 or 3 classes of contaminants to represent the gradient: perhaps total PAH, total PCB and 1 heavy metal (or, as in Fig. 15 of Gray et al. 1988, the combination of metal levels given by the first principal component from a PCA on the site/metals array). Even this would be over-ambitious if these 2 or 3 representatives were highly correlated along the gradient.

An ordination then provides a good means for displaying the relations between the biological pattern and the chemistry, by superimposing symbols of different sizes, representing chemical values, at the sample positions on the faunal ordination plot, e.g. Field et al. (1982). (This technique was used for workshop data, Fig. 15 of Gray et al. 1988, though its usefulness was limited by the lack of replicate chemical values matched to the replicate faunal samples.) At best, one may be able to distinguish a clear axis of an increasing chemical gradient, possibly 2 axes if 2 types of contaminant change in different ways along the gradient and induce different community changes. (More likely a second axis could reflect an important but uncontrolled physical variable, not confounded with the chemical gradient). This visual approach could be formalised by multiple regression of the chemical variable on the ordination axes (or bivariate multiple regression for 2 chemical/physical variables). When multivariate normal assumptions are justifiable for the species abundance matrix, a wider range of inference is available, in terms of multivariate ANCOVA and multivariate correlation analysis (Canonical Correlation). The discussion on confounding effects of physical variables, in the section on testing structure, is equally relevant here.

COMPARISON OF METHODS

Whilst it is important to examine differing biological effects measures in combination (particularly across differing levels of biological organisation) it can be important in some cases to establish the relative effectiveness of each of a set of measures, in detecting the type and degree of pollution impact present in a particular field study. Questions of comparative sensitivity can be addressed at varying levels of statistical sophistication. For illustration, assume that the problem is to discriminate 2 sites, an impacted area and a reference site, at each of which n replicates of a particular biological response are available. The 2-sample t-test of 'no between-site differences' gives (possibly after transformation) the usual Student statistic t, a standardised difference of the mean response at the 2 sites (this is just the ANOVA F statistic in a different guise). At the simplest level, if this is non-significant then the biological variable has no *demonstrated* sensitivity to that impact (it may have had some sensitivity if more replicates had been taken, but this suggestion cannot be examined without more data). At the next level, significant responses could be ordered – as a measure of relative performance – by their t (or p) values. How-

ever, this could be misleading, since no account is taken of how the discriminating ability of a test fluctuates with changes in the sample size n (i.e. changes in experimental effort). By doubling its sample size an 'inferior' test could become 'superior', perhaps still with less experimental effort than for the initially 'superior' test.

A better basis for a comparative study is to define sensitivity to a specific impact as the *power* to detect that impact (in a standard, 2-tailed, 0.05 significance level test), expressed as a function of sample size. More conveniently, one can then ascertain the critical sample size, $N^\bullet$, necessary to ensure that the power P (the probability of detecting the impact) is at least 0.95 (say). $N^\bullet$ can then be converted to more readily comparable units of 'cost'; this might initially be man-hours, though some consideration of technical sophistication costs may also be involved. Where there is difficulty in constructing 'costs', at least the suggested power functions will allow comparisons of the form 'measure A needs to treble its replicate numbers to match the sensitivity of B, for this impact', and a qualitative knowledge of the operating requirements for the 2 techniques might then suffice.

Assuming that the *observed* difference between the response means at the reference and polluted sites represents the *true* impact level, then $N^\bullet$ can be computed exactly using the non-central t distribution (e.g. Scheffé 1959) or, as suggested in the design section of this paper, by the approximation:

$$N^\bullet > (k+1) + \tfrac{1}{4}(k^2+1), \qquad (9)$$
$$k = (0.5n)\{2 + \Phi^{-1}(P)\}^2/t^2.$$

For example, when power $P = 0.95$, $k = 6.6n/t^2$. The approximation will be adequate for $N^\bullet \geq 4$ and $P > 0.5$. Of course, the *true* difference in mean response between reference and impacted areas is only known to within certain limits, and these can be transformed into a 95 % confidence interval for the critical sample size, given by the above equation for $N^\bullet$ but with

$$k = (0.5n)\{2 + \Phi^{-1}(P)\}^2/(t+2)^2 \text{ (lower limit)}, \qquad (10)$$
$$k = (0.5n)\{2 + \Phi^{-1}(P)\}^2/(t-2)^2 \text{ (upper limit)}.$$

The procedure is illustrated on the Fig. 1 data and a further set from the workshop (Addison & Edwards 1988, Widdows & Johnson 1988). Comparing the endpoint sites of the field gradient, $t = 5.0$ ($n = 16$) for mussel 'scope for growth', and $t = 13.2$ ($n = 11$) for log(EROD) in flounder, whereas an activity measure of a further enzyme, log (BPH), from the flounder MFO system gave $t = 2.2$ ($n = 11$). Clearly, the impact is sufficiently large on the first 2 measures for it to be detected (with 95 % certainty) from modest numbers of replicates, the estimates of critical sample size $N^\bullet$ being 10 and 4 respectively; by contrast $N^\bullet = 31$ for the third measure. The 95 % confidence intervals for $N^\bullet$ are (6,24), (4,4) and (10,2500) respectively. The exercise can be repeated for intermediate sites on the gradient, comparing against both reference and other impacted sites, so that sensitivity can be assessed over different ranges of the 'dose' scale. (The idea extends straightforwardly to the situation where a dose-response curve has been fitted to the data.)

Benthic community responses are harder to fit into a framework of this sort, unless they lead to univariate measures of community change (e.g. diversity indices) calculated on a number of replicate samples from each site. If they do, there is no theoretical difficulty in comparing their 'sensitivity' directly with that of sub-lethal stress responses for the same sites, using exactly the above procedures. However, as the GEEP Workshop demonstrated, multivariate and graphical methods of description and testing are more sensitive than diversity indices, and usually to be preferred. Although, for example, the ANOSIM test will improve its ability to discriminate between sites as the number of replicate cores increases, nothing at all is known about the formal relation between power and sample size in this case. Some progress might be possible with 1-dimensional ordination solutions by, for example, examining the positions of replicates from each site on the first PC axis or the first Canonical Discriminant axis. However, there will be selection biases here which cannot easily be compensated for (such axes are automatically chosen to reflect the direction of greatest *observed* change).

Returning to the sub-lethal response examples above, the exercise was carried to the point of determining man-hours per replicate and converting $N^\bullet$ to an equivalent man-hour total. (The relationships between total effort and numbers of replicates were approximately linear). In fact, the comparison between scope for growth, EROD and BPH is not changed by this additional information, the relative 'critical man-hours' of laboratory time necessary to demonstrate (with 95 % certainty) a difference between the 2 endpoint sites being approximately 15, 6 and 50 h respectively (with a very wide confidence interval for the latter of course). Other relevant 'costs' include those for consumables (e.g. < \$ 1 per replicate for the physiology but \$ 5 to 10 per replicate for the biochemistry) and, of course, more major issues of capital equipment requirements and other 'level of sophistication' costs. However, it is unrealistic to expect to produce a univariate measure which places these 'costs' on a common footing for a full cost-benefit analysis; the objectives and background conditions for any future study will always dictate a different weighting of the factors involved.

Acknowledgements. We are grateful to many GEEP Workshop participants for helpful discussions and for permission to use their data in illustration of statistical points; also to John Hall, for implementing the ANOSIM test in FORTRAN 77 code.

LITERATURE CITED

Addison, R. F., Edwards, A. J. (1988). Hepatic microsomal mono-oxygenase activity in flounder *Platichthys flesus* from polluted sites in Langesundfjord and from mesocosms experimentally dosed with diesel oil and copper. Mar Ecol. Prog. Ser. 46: 51–54

Andrews, D. F., Gnanadesikan, R., Warner, J. L. (1971). Transformations of multivariate data. Biometrics 27: 825–840

Bayne, B. L., Clarke, K. R., Moore, M. N. (1981). Some practical considerations in the measurement of pollution effects on bivalve molluscs, and some possible ecological consequences. Aquat. Toxicol. 1. 159–174

Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. J. R. statist. Soc. Ser. B 26: 211–243

Box, G. E. P., Cox, D. R. (1982). An analysis of transformations revisited. J. Am. statist. Ass. 77: 209–210

Bray, J. R., Curtis, J. T. (1957). An ordination of the upland forest communities of Southern Wisconsin. Ecol. Monogr. 27: 325–349

Cochran, W. G. (1983). Planning and analysis of observational studies. Wiley, New York

Diggle, P. J. (1983). Statistical analysis of spatial point patterns. Academic Press, London

Engen, S. (1978). Stochastic abundance models. Chapman and Hall, London

Everitt, B. (1974). Cluster analysis. Heinemann, London

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

Fienberg, S. E. (1970). The analysis of multidimensional contingency tables. Ecology 51: 419–433

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–38

Gray, J. S., Aschan, M., Carr, M. R., Clarke, K. R., Green, R. H., Pearson, T H., Rosenberg R., Warwick, R. M. (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. (1979). Sampling design and statistical methods for environmental biologists. Wiley, New York

Gundersen, H. J. G., Osterby, R. (1981). Optimizing sampling efficiency of stereological studies in biology· or 'Do more less well!' J. Microsc. 121: 65–73

Hill, M. O. (1973). Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61 237–249

Hill, M. O., Gauch, H. G. (1980). Detrended correspondence analysis, an improved ordination technique. Vegetatio 42:47–58

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. J. R. statist. Soc. Ser. B 30: 582–598

Hurlbert, S H. (1984). Pseudoreplication and the design of ecological field experiments. Ecol. Monogr. 84: 187–211

Kruskal, J. B., Wish M. (1978). Multidimensional scaling. Sage Publications, Beverley Hills, California

Lefkovitch, L. P. (1979). Hierarchical clustering from principal coordinates: an efficient method for small to very large numbers of objects. Math. Biosci. 31: 157–174

Livingstone, D. R. (1988). Responses of microsomal NADPH-cytochrome c reductase activity and cytochrome P-450 in digestive glands of *Mytilus edulis* and *Littorina littorea* to environmental and experimental exposure to pollutants. Mar. Ecol. Prog. Ser. 46: 37–43

Lowe, D. M. (1988). Alterations in cellular structure of *Mytilus edulis* resulting from exposure to environmental contaminants under field and experimental conditions. Mar. Ecol. Prog. Ser. 46: 91–100

Lowe, D. M., Moore, M. N., Bayne, B. L. (1982). Aspects of gametogenesis in the marine mussel *Mytilus edulis* L. J. mar. biol. Ass. U.K. 62: 133–45

Mardia, K. V., Kent, J. T., Bibby, J. M. (1979). Multivariate analysis. Academic Press, London

Nash, J. C. (1979). Compact numerical methods for computers. Adam Hilger, Bristol

Orloci, L. (1978). Multivariate analysis in vegetation research, 2nd edn. Junk, The Hague

Pielou, E. C. (1975). Ecological diversity. Wiley, New York

Scheffé, H. (1959). The analysis of variance. Wiley, New York

Schwinghamer, P. (1988). Influence of pollution along a natural gradient and in a mesocosm experiment on biomass-size spectra of benthic communities. Mar. Ecol. Prog. Ser. 46: 199–206

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

Siegel S. (1956). Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York

Warwick, R. M. (1986). A new method for detecting pollution effects on marine macrobenthic communities. Mar. Biol. 92: 557–562

Warwick, R. M. (1988). Analysis of community attributes of the macrobenthos of Frierfjord/Langesundfjord at taxonomic levels higher than species. Mar. Ecol. Prog. Ser. 46: 167–170

Warwick, R. M., Carr, M. R., Clarke, K. R., Gee J. M., Green, R. H. (1988). A mesocosm experiment on the effects of hydrocarbon and copper pollution on a sublittoral soft-sediment meiobenthic community. Mar. Ecol. Prog. Ser. 46: 181–191

Widdows, J. (1985). The effects of fluctuating and abrupt changes in salinity on the physiology of *Mytilus edulis*. In: Gray, J. S., Christiansen, M. E. (eds.) Marine biology of polar regions and effects of stress on marine organisms. Wiley, Chichester, p. 555–566

Widdows, J., Johnson, D. (1988). Physiological energetics of *Mytilus edulis*: Scope for Growth. Mar. Ecol. Prog. Ser. 46: 113–121