Journal of Experimental Marine Biology and Ecology 366 (2008) 56–69

Contents lists available at ScienceDirect

Journal of Experimental Marine Biology and Ecology

journal homepage: www.elsevier.com/locate/jembe

Testing of null hypotheses in exploratory community analyses: similarity profiles and biota-environment linkage

K. Robert Clarke [a,b,*], Paul J. Somerfield [a], Raymond N. Gorley [b]

[a] Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK
[b] PRIMER-E Ltd, c/o Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK

ARTICLE INFO

Keywords:
Constrained clustering
Gradient studies
LINKTREE
Matrix correlation
Non-parametric multivariate
Permutation test
Selection bias
SIMPROF

ABSTRACT

Tests for null hypotheses of 'absence of structure' should play an important role in any exploratory study, to guard against interpretation of sample patterns that could have been obtained by chance, and two new tests of this type are described. In the multivariate analyses that arise in community ecology and many other environmental contexts, e.g. in linking assemblage patterns to forcing environmental variables (gradient analysis), the problem of chance associations is exacerbated by the large number of combinations of abiotic variables that can usually be examined. A test which allows for this selection bias is described (the global BEST test), which applies to any dissimilarity measure, utilises only rank dissimilarities, and operates by permutation, assuming no specific distributional form or parametric expression for the biotic to abiotic links. A second permutation procedure, the similarity profile routine (SIMPROF), tests for the presence of sample groups (or more continuous sample patterns) in *a priori un*structured sets of samples, for which an *a priori* structured test (e.g. the widely-used ANOSIM) is invalid. One context is in interpreting dendrograms from hierarchical cluster analyses: a series of SIMPROF tests provides objective stopping rules for ever-finer dissection into subgroups. Connecting these two tests is a third methodological strand, adapting De'ath's multivariate equivalent of univariate CART analysis (Classification And Regression Trees) to a non-parametric context. This produces a divisive, constrained, hierarchical cluster analysis of samples, based on their assemblage data, termed a linkage tree. The constraint is that each binary division of the tree corresponds to a threshold on one of the environmental variables and, consistently with related non-parametric routines, maximises the high-dimensional separation of the two groups, as measured by the ANOSIM R statistic. Such linkage trees therefore provide abiotic 'explanations' for each biotic subdivision of the samples but, as with unconstrained clustering, the LINKTREE routine requires objective stopping rules to avoid over-interpretation, these again being provided by a sequence of SIMPROF tests. The inter-connectedness of these three new developments is illustrated by data from the literature of marine ecology.

1. Introduction

Professor John Gray was a strong advocate of the insights obtainable from exploratory studies of gradients and did much to demonstrate their efficacy in the contexts of monitoring for pollution and studying biodiversity (see Gray et al., 1988; Gray et al., 1990; Ellingsen and Gray, 2002, amongst many others). For multivariate community analyses, however, gradient studies (broadly characterisable as adopting a regression approach) have sometimes suffered in comparison to studies involving factorial designs (broadly speaking, an analysis of variance approach) by their perceived lack of hypothesis testing for structure elucidated only *a posteriori*. Such criticism is often justified: for example, a search through large numbers of environmental variables, for combinations which 'explain' the among-sample structure of a biotic assemblage, is almost guaranteed to find a combination with some *apparent* explanatory power, even where there is no real linkage. The process of searching through many solutions for the one that optimises some criterion inevitably involves strong selection bias. At the least, what is required here is a formal test of the null hypothesis that there is no link between the sample patterns of biota and environment, adjusting for this selection bias. If the null can be decisively rejected then there is some objective basis for interpreting the observed correlative links.

In similar vein, application of hierarchical cluster analysis to a set of *a priori* unstructured samples of assemblage data yields a dendrogram, whether agglomerative or divisive, in which ever finer distinctions are drawn between groups of samples, ultimately terminating in each sample placed in a different group. Given that a cluster analysis will produce a grouping from data consisting entirely of random numbers, and thus with no meaningful sample structure, the question naturally arises as to what objective basis there is for interpreting particular groups or subgroups displayed by the dendrogram. Again, statistical

* Corresponding author. Plymouth Marine Laboratory, Prospect Place, West Hoe, Plymouth PL1 3DH, UK. Tel.: +44 1752 633100; fax: +44 1752 633101.
E-mail address: krc@pml.ac.uk (K.R. Clarke).

0022-0981/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jembe.2008.07.009

testing is needed, this time in the form of a series of null hypothesis tests that particular groups displayed in the dendrogram have no meaningful internal structure. Only if such a hypothesis can be rejected is it permissible to interpret a further subdivision of an existing group.

This paper describes such tests, for analysis of any similarity, distance or dissimilarity matrices (generically referred to as 'resemblance' measures, following Legendre and Legendre, 1998). It places these tests in the context of the non-parametric approach to analysing species-by-samples matrices described by Clarke (1993), which has been widely adopted in marine community ecology in particular, largely through availability of the PRIMER package (v6, Clarke and Warwick, 2001; Clarke and Gorley, 2006). A notable early step in the latter was Professor Gray's enthusiastic encouragement of development of these techniques through a series of workshops held under the auspices of the UNESCO/IOC Group of Experts on the Effects of Pollutants (Bayne et al., 1988) and the FAO/UNEP Mediterranean Pollution Programme. The core routines in this approach include non-metric MDS ordination of samples and ANOSIM tests of *a priori* factors defined on them, together with indirect gradient analyses linking biotic assemblage patterns to 'best' subsets of environmental variables, exhibiting matching sample structure (BEST routine). These routines are based on unconstrained choice of a resemblance matrix appropriate to the data type and question of interest, and the ANOSIM and BEST routines utilise only the rank values of the among-sample resemblances.

Within the existing framework, this paper adds, firstly, a 'global BEST test' which examines whether the highest rank correlation (ρ), obtainable between the biotic similarity matrix and the matching distance matrix from the optimal subset of environmental variables, exceeds values that would be expected by chance under the null hypothesis (of no real biota-environment link). Secondly, a 'similarity profile' (SIMPROF) test is described, in which the biotic similarities from a group of *a priori* unstructured samples are ordered from smallest to largest, plotted against their rank (the similarity profile), and this profile compared with that expected under a simple null hypothesis of no meaningful structure within that group. Repeated application of this test generates a stopping rule for *a posteriori* division of the samples into ever smaller subgroups, as in hierarchical cluster analysis. These two analytical strands converge in a third routine, a counterpart to the BEST procedure of matching environmental information to species patterns, which adapts the Multivariate Regression Trees of De'ath (2002) to the non-parametric framework in PRIMER. The LINKTREE procedure is a form of constrained cluster analysis involving a divisive partition of the biotic community samples into ever smaller groups, but in which each division has an 'explanation' in terms of a threshold on one of the environmental variables. As with agglomerative hierarchical clustering, such linkage trees also need stopping rules to avoid random sampling variation among samples from a single assemblage being interpreted as further sub-group structure. These are again provided by a series of similarity profile (SIMPROF) tests.

It should be borne in mind throughout that, though the above outline and the examples of this paper are couched in terms of tests on species assemblages and their relation to environmental variables, nothing in the formulation of the methods restricts their use to this context. The SIMPROF test will provide stopping rules for any *a posteriori* subdivision of a group of samples, based on multiple variables of taxa, physical environment, chemical water-quality, measures of diversity, biomarkers, distributions of particle sizes, etc. The global BEST test, rather than matching subsets of environmental variables to a fixed pattern of resemblances for whole communities, can be applied to testing whether subsets of species show a significantly matching pattern of samples to those of a fixed environmental gradient. Similarly, an optimal subset of biomarkers (or other metrics) can be tested for its match to an observed chemical gradient, or to manipulated levels of contaminants; patterns in subsets of one group of biotic variables can be tested for their ability to 'explain' patterns in another set (corals structuring assemblages of reef-fish, infaunal macrobenthos structuring meiobenthic communities etc); and many other 'linkage' problems could be formulated and tested in this way.

2. Methods

2.1. Global BEST test

The idea behind this test is outlined in the schematic diagram of Fig. 1. This shows, in normal typeface, the routine for linking of biota to environment (Bio-Env) described by Clarke and Ainsworth (1993). The data consist of two matrices (left-hand side), both referring to the same set of n samples (locations/times/treatments, or whatever context determines the sampling programme). For the biotic variables (top row), a triangular matrix of resemblances between samples is calculated for whatever combination of pre-treatment (standardisation, transformation, variable weighting) and (dis)similarity or distance coefficient is appropriate to the context (e.g. Clarke et al., 2006b). This is the core information (the 'fixed' resemblance matrix) which describes the precise biotic relationships among samples and which can be conveniently, though only approximately, displayed in a non-metric MDS ordination. For the 'explanatory' matrix (bottom row), all subsets of the p environmental variables for the same set of n samples are examined, one at a time (p single variables, $p(p\text{-}1)/2$ pairs

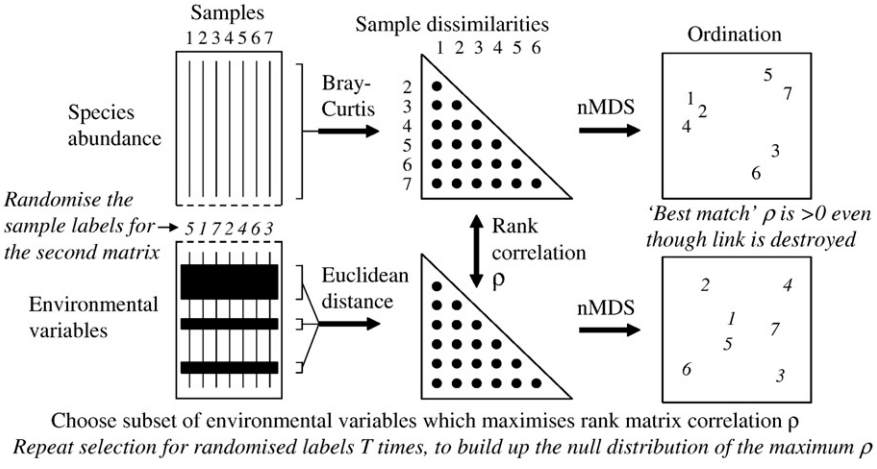

Fig. 1. Schematic diagram of the BEST matching procedure and the global BEST test.

of variables, $p(p\text{-}1)(p\text{-}2)/6$ triples, etc., totalling $(2^p - 1)$ possible combinations). For each subset, a resemblance matrix is calculated, again using any combination of transformation, normalisation, choice of distance measure etc, that is appropriate for the context. The extent to which the particular subset of environmental variables has 'captured' the pattern of samples from the biotic community is measured by correlating the matching entries of the two resemblance matrices. This matrix correlation (Mantel, 1967) could again be chosen in several ways, but in the context of the non-parametric framework outlined earlier, and noting that the biotic resemblances would typically be dissimilarities (range 0 to 100) and the abiotic resemblances Euclidean (range 0 to infinity), so that strictly linear relationships are not possible, a natural choice would be a rank correlation (ρ). Spearman, Kendall (Kendall, 1970) or the weighted Spearman measure given in Clarke and Ainsworth (1993) are among the possibilities here. The Bio-Env procedure then simply selects the combination of environmental variables which maximises ρ, i.e. 'best explains' the biotic assemblage structure. More generally, we shall refer to this as the BEST (*Bio-Env* + *ST*epwise) procedure, to include also the situations in which calculation of all $(2^p - 1)$ combinations of explanatory variables is prohibitive, and a stepwise algorithm, employing forward-stepping and backward-elimination as in step-wise regression (Draper and Smith, 1981), is used to search for the optimal ρ (Clarke and Warwick, 1998).

What is immediately clear is that the combination of explanatory variables selected by the BEST procedure is almost guaranteed to return a value of ρ greater than zero, even in cases where there is absolutely no link between the biotic and environmental data. That is, there is strong selection bias, and it is clearly not valid to carry out a Mantel-type test of the null hypothesis $\rho = 0$ by permutation (e.g. the RELATE routine in PRIMER, Clarke et al., 1993), between the biotic resemblances and the 'best' environmental distance matrix. A permutation test which allows for this selection bias is, however, provided by the global BEST test, illustrated in Fig. 1 (italic typeface). The key to constructing a valid null distribution for the test statistic in any permutation test is to replicate exactly the process used in generating that statistic, but under conditions appropriate to the null hypothesis. Here, the null hypothesis of no linkage of biotic to environmental variables can be recreated by permuting the sample labels for one of the data matrices, and Fig. 1 shows a random permutation of the columns of the environmental matrix in relation to those of the biota. (It does not matter whether the columns of one or both matrices are independently permuted in this way, the net effect always being to destroy any meaningful relationship between the two sets of samples). The entire BEST procedure is now repeated, i.e. selecting variables in the explanatory matrix to maximise the rank correlation between their resemblance matrix and the fixed biotic dissimilarities. However large the resulting ρ, this can only be an apparent relation, since any real linkage has been destroyed by the permutation, and it is clearly a value of ρ that it is possible to obtain, by chance, under the null hypothesis. Now, a second random permutation of the columns of the environmental matrix is performed and the entire BEST procedure repeated once more, to obtain a second possible value of ρ under the null hypothesis. This process is carried out for a total of $T = 999$ (say) random permutations of the labels, and a histogram formed of all the resulting optimal ρ values, the so-called null distribution.

As with any other permutation test, if only t of the T permuted values of optimal ρ are greater than or equal to the real optimal ρ, then one can reject the null hypothesis (of no link of biota to environment) at the $P < (t + 1)/(T + 1)$ level. This is a conservative P value (an upper limit), Hope (1968). As always, larger values of T will give more accuracy in defining P, and may allow a greater degree of significance to be quoted if the observed ρ exceeds that for any of the permutations ($t = 0$). In the latter case, however, rejection of the null hypothesis at $P < 0.001$, for $T = 999$, ought to be sufficient for most purposes, bearing in mind that this is a single (global) test, rather than any sort of multiple testing procedure. Note that it is seldom necessary to consider performing all possible permutations of the sample labels in one matrix in relation to the other, rather than just a large random subset of these permutations, since there are always a very large number ($n!$ for n samples) from which to choose T.

2.2. Similarity profiles

The schematic diagram in Fig. 2 illustrates the construction and testing of similarity profiles. For any specified set of samples (here 7 are shown), similarities are calculated between every pair, using whatever pre-treatments and resemblance measure are appropriate to the nature of the data and the question being asked. These sample similarities (21 of them here) are ranked from smallest to largest and plotted against their ranks (the numbers 1 to 21), giving the bold, continuous line in the SIMPROF plot of Fig. 2.

The informal principle on which the test works is that genuine structure within this set of samples will be evidenced by an excess of smaller and/or larger similarities than expected under the null hypothesis that all samples are drawn from the same species assemblage (or whatever the variables represent). For example, if samples 1, 2 and 3 form a cluster (group A), well differentiated from 4, 5 and 6 (group B), with sample 7 intermediate (group C), then one would expect 6 large similarities (within A and within B), 6 somewhat lesser values (C to A and C to B), and 9 small similarities (A to B), giving a relatively steep profile. If there is no such group structure, all similarities will vary randomly and relatively tightly round some common mean, leading to a much shallower curve.

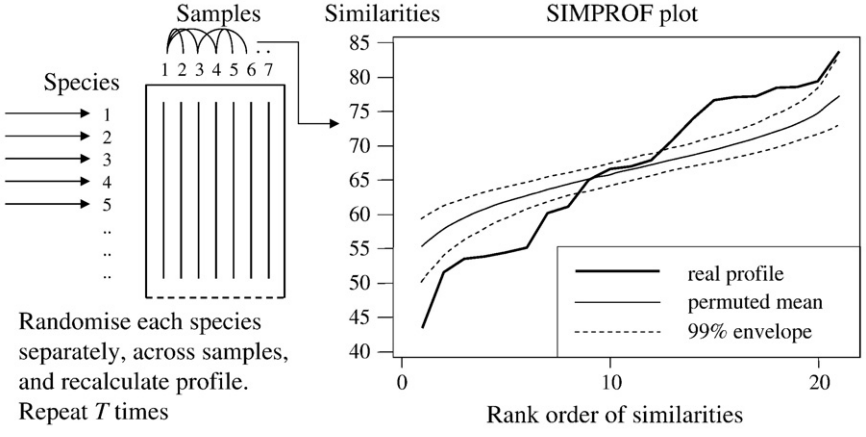

Fig. 2. Schematic diagram of similarity profiles and the SIMPROF test.

In order to compare the observed profile with an expectation under the null hypothesis (of no multivariate structure), permutation is again employed. The values in each row of the matrix (species) are permuted across all columns (samples), independently from row to row, and the similarity profile for samples recalculated. This recalculation is then repeated for another random and independent permutation of all values within each row of the matrix, and so on, a total of T times. Plotting all these permuted profiles allows a mean profile to be drawn, of the average values at each rank (the thin continuous line in the SIMPROF plot of Fig. 2), and also upper and lower bounds, corresponding to an envelope in which lie 99% (say) of the permuted similarities at each rank (the dashed lines in Fig. 2).

The permutation procedure employed is that categorised by Legendre and Legendre (1998) as Model 1, the 'environmental control model' (Whittaker, 1956). Its relevance and limitations here are discussed in more detail later, but it is intuitively clear that the permuted profiles can in no way correspond to interpretable multivariate structure in the assemblages, i.e. to *apparent* synergetic or antithetic relationships between species (whether intrinsic or extrinsically forced by environmental conditions). Such relationships are what drive the groupings (or more continuous changes) observed amongst a set of samples, and which multivariate methods set out to display and interpret. In other words, if the real profile falls within the channel represented by the 99% bounds, and thus looks identical to a profile that could have been created by a random redistribution of abundances across the samples, separately for each species, then there must be little basis for multivariate interpretation of the samples making up the real profile, e.g. by further clustering (or ordination). Note that the converse is not necessarily true. An observed profile that differs from those that can be generated under the above permutations may represent a rather trivial multivariate structure, of no great biological or practical significance, but at least a positive test of this sort provides a licence for further exploration.

A formal hypothesis test can be constructed from the statistic π, defined as the absolute deviation of the real profile from the mean of T permuted profiles, summed across all ranks on the x axis; T would typically be large so that the mean is well characterised (say, $T \geq 1000$). Another way of thinking of π is as the area between the real profile and the mean of the permuted profiles. The test statistic π needs to be compared with its null distribution, and the latter is obtained by generating a second set of S permuted profiles, in exactly the same way, where S would also be large (say 999). Each of those profiles from the second set is then compared to the mean profile from the first set of permutations, and its deviation π calculated. The observed value of π for the real profile is then referred to the null histogram of π values from the second set, and significance or otherwise ascribed exactly as for the earlier BEST test.

The SIMPROF test described above applies to a defined set of samples and is a means of stopping unwarranted further analysis of their sub-structure. The obvious application is therefore in the context of hierarchical cluster analyses, when the procedure becomes a series of SIMPROF tests. Starting at the top of an already computed hierarchy, progress to the first division of that dendrogram is only permitted if the null hypothesis (of no internal structure in the whole set of samples) is rejected. Similarly, progress to each succeeding partition only takes place if the current set of samples is deemed still to have internal structure. Once a non-significant test result is obtained, no tests are performed further down that branch and the samples below that point are regarded as homogeneous. Given the multiple testing inherent in this hierarchical approach, it might be desirable to use a more stringent P value (e.g. 0.01 or even 0.001) as the criterion for rejecting the null hypothesis of 'no structure' and progressing down to the next division. Too pedantic an approach to significance levels here, however, seems counter-productive, given the likely decline in power of the test with the shorter similarity profiles arising from smaller groups. This would make the process appropriately and fortuitously self-limiting, with a strong likelihood of detecting structure at the coarsest levels, but requiring clear evidence of heterogeneous similarities before permitting further division of already small numbers of samples.

A similar application of repeated SIMPROF tests to a hierarchical partition is found in the context of 'linkage trees', a clustering technique with the same general aim as the BEST procedure of section 2.1, of linking multivariate patterns (e.g. of biota) to explanatory variables, but taking a more piecemeal, predictive approach, as follows.

2.3. Linkage trees

De'ath (2002) introduced 'Multivariate Regression Trees (MRT)', a method for hierarchical, divisive clustering of samples which extends the technique of 'Classification And Regression Trees' (CART, Breiman et al., 1984), found in some major statistics packages such as S-PLUS and the routine rpart (recursive partitioning) in the R package. For the univariate case, say for a single diversity measure or biomarker, samples are initially placed in a single group which is then successively subdivided, in binary partitions, in such a way as to minimise the within-group variance (or, equivalently, maximise the between-group variance) for the two clusters formed at each step. This is not an unconstrained subdivision: the only partitions considered are ones that correspond to a threshold level for one of the associated environmental variables (or whatever the explanatory matrix represents). For example, 6 samples A-F of a sediment infaunal community might be split into groups A-D and E-F because this minimises the total within-group variance of Shannon diversity H', and the total hydrocarbon level (PAH) in the sediments at A-D is less than it is at either E or F (e.g. PAH < 20 for A-D and PAH > 30 for E-F). Other environmental variables may also permit the same division (e.g. grain size $\phi > 3$ for A-D, $\phi < 2.9$ for E-F), in which case multiple 'explanations' are given for that particular split. Note that the chosen division may not be the optimal unconstrained split: A-C and D-F may give smaller total within-group variance for H', but this option is not considered if no environmental variable takes values exclusively higher (or lower) in A-C than D-F. The two groups are then further subdivided, again commensurate with an inequality on at least one of the environmental variables, all of which can be used repeatedly at different threshold levels, the process ending up with a dendrogram with one or more environmental explanations for each binary partition.

De'ath (2002) shows that this idea can readily be extended to multivariate response variables, such as multi-species abundances, by minimising average distances or dissimilarities within groups. These are calculated using whatever resemblance coefficient is appropriate for the particular context, e.g. Bray-Curtis dissimilarity for species abundances, or normalised Euclidean distance for explaining biomarker responses to environmental contaminants.

Here, we take the process one step further, in keeping with our general non-parametric framework, by utilising only the ranks of the resemblances, and maximising the ANOSIM R statistic (Clarke and Green, 1988) between the two groups formed at each division. R is a difference of average rank dissimilarities between and within groups, scaled to take values over a fixed range, up to 1. As discussed by Clarke (1993), aside from its role in null-hypothesis testing (which is not the relevant usage here), R provides a general measure of the degree of separation of two groups in the high-dimensional space represented by the resemblance matrix for those samples. It reaches its maximum value of 1 when all dissimilarities between the two groups exceed any dissimilarity within either group, and its values may legitimately be compared irrespective of the group sizes. For each new subgroup of the tree, the resemblances for those samples are re-ranked and a further division into two groups is made, such that R is maximised, conditional on this being a split indicated by a threshold on one or more of the environmental variables, as described above. Note that

there should be no need to insist on equal-sized groups, indeed one of the subgroups could be a singleton sample. The completed tree therefore consists of a partition of the original set of samples, in terms of their biotic community structure (say), with each group having an 'explanation' in terms of a set of inequalities on associated environmental variables. In a hypothetical example, all samples in cluster A, the result of three binary divisions, might have: Total PAH < 10 or Total Cu < 30 (first split, R = 0.7), Water depth > 50 (second split, R = 1) and Total PAH < 1 or Salinity > 34 (third split, R = 0.8). This suggests that, in addition to explanation, there could be prediction of the group into which a future sample will fall, in terms of its community composition, given knowledge of the relevant environmental variables (though there could be many ambiguities, e.g. for abiotic values outside the range of previous experience).

De'ath (2002) represents the successive binary divisions of groups in his MRT displays by a tree diagram with equal steps, but there are advantages to displaying each division with a continuous scale on the y axis, representing the magnitude of that separation in relation to earlier or later subdivisions. (In univariate implementations, such as in S-PLUS or Breiman et al., 1984, this is usually the between-group sums of squares for the two groups created at that division.) In our non-parametric framework, the closest analogue for the y axis scale would seem to be the average of the rank dissimilarities between samples in the two groups, and there is another good reason for using this. As is clear from the simple example above, the values of the ANOSIM R statistic itself cannot be used for this y axis scale, since the ranks are rescaled for each new subset of samples and the R values will remain high throughout, for satisfactory partitions. Note, however, that since the ranks of a set of $M = n(n\text{-}1)/2$ dissimilarities from n samples are constrained to average $(M + 1)/2$, there is a simple relationship between the average rank dissimilarities between and within any two groups formed from these samples. Thus ANOSIM R has the following equivalent definitions:

$$R = \left[(av\ rank)_{between} - (av\ rank)_{within}\right]/(M/2)$$
$$= [2/(M - m_{between})].\left[(av\ rank)_{between} - \{(M+1)/2\}\right]$$

where m_{between} is the number of dissimilarities calculated between the two groups. That is, R *is* a simple (linear) function of the average of the rank dissimilarities between the groups. All that is necessary therefore to obtain a natural y axis scale for the tree (denoted B%) is that the *original* rank dissimilarities are used throughout (i.e. ignoring the re-ranking that takes place in recalculating R at each step). This average of the between-group rank dissimilarities is divided by the maximum value it can take for a perfect division (R = 1) on the first partition, and then multiplied by 100, to give a positive scale (B%) which never exceeds 100.

The simple interpretation of B% is that it measures how well separated the two groups of samples are, in the current division (relative to the maximum separation achievable at the first partition). One could choose to terminate the divisive clustering at a fixed level for B% (e.g. treating as unimportant any divisions which take place below B = 5%), or at the point at which a branch has less than a fixed number of samples (e.g. never attempting to divide a group with fewer than 4 samples). A more satisfactory stopping rule, however, is provided by a SIMPROF test, as intimated in the last section: never divide a group (and provide an environmental 'explanation' for this division) if there is no demonstrated multivariate structure in the biotic composition that needs explaining.

The terminology 'linkage tree' for the dendrogram produced by the above procedure is preferred to 'regression tree' because the latter seems inappropriate for a technique which sets out to avoid the 'regression to the mean' inherent in fitting parametric linear models, the source of the term 'regression' (Fisher, 1925, though the term 'regression to mediocrity' is believed to date from a presidential address on anthropology by Francis Galton in 1885). It is precisely the ability linkage trees have to sidestep the global modelling of response variable(s) in regression analyses, and focus instead on piecemeal, local structure and local explanations, that justifies this approach.

2.4. Data sets

The following literature data sets are used to illustrate the three inter-related techniques of global BEST tests, similarity profiles and linkage trees.

2.4.1. Clyde macrobenthos

Pearson and Blackstock (1984) report the results of biotic and abiotic sampling across the sewage-sludge dumpground at Garroch Head, Firth of Clyde, Scotland. Here, data from 1983 are examined at 12 sites along an E-W transect, the pooled contents of several grab samples for soft-sediment macrofauna at each site being identified, and biomass recorded for each of 84 taxa. Also recorded for each site was a suite of (mainly) contaminant variables, the metals Cu, Mn, Co, Ni, Zn, Cd, Pb, Cr, organics % Carbon, % Nitrogen, and the depth of the water column.

2.4.2. Messolongi lagoon diatoms

Danielidis (1991) studied the diatom assemblages in the lagoons of Messolongi, Aitoliko and Kleissova in Eastern Central Greece, encompassing hypersaline and elevated nutrient conditions. At each of 17 sites, abundances of 193 species of diatom were recorded, along with the physico-chemical environmental variables: temperature, salinity, DO_2, pH, PO_4, NH_3, NO_2, NO_3, Inorganic-N and SiO_2.

2.4.3. Exe estuary nematodes

Warwick (1971) describes assemblage data on 140 species of free-living nematode at 19 intertidal sites in the Exe estuary, UK. Analysed here are counts averaged over 6 bi-monthly sampling occasions in one year. Also recorded are 6 environmental variables for the sediments at those sites: median particle diameter, depth of the blackened H_2S layer, interstitial salinity, % organics, depth of the water table and height up the shore (position over the inter-tidal range).

2.4.4. Bristol Channel zooplankton

Collins and Williams (1982) describe data collected by double-oblique plankton net hauls from 57 sites in the Bristol Channel and Severn Estuary, UK. Sites were located on a regular grid covering the whole area and, for this exercise, broadly categorised into 4 salinity ranges, on the basis of seasonally-averaged salinity readings. An overall total of 24 zooplankton species were identified and enumerated at each site through the year, and a single seasonally-averaged sample for each site is analysed here.

2.4.5. Ekofisk oil-field macrobenthos

Professor Gray was instrumental in recognising the importance to statistical methodology and EIA development of the soft-sediment macrofaunal data being collected under licencing arrangements around oil rigs in the Norwegian sector of the N Sea (Gray et al., 1990), and in later papers he explored the implications for biodiversity studies of this unique resource of benthic marine data, taken over large spatial scales (e.g. Ellingsen and Gray, 2002). Here, we use only the data on 174 macrobenthic taxa recorded from the 1987 study of 39 sites in the Ekofisk oilfield. These are numbered by increasing distance from the oil-field centre, which ranges from 100m to 8km (in one case 30km) in a combination of 4- and 5-spoke radial transects. The assemblage data consist of total counts of each taxon from three Day-grab samples at each site.

2.5. Computations

The methods have been implemented in the PRIMER software (Clarke and Gorley, 2006), specifically as the routines SIMPROF (similarity

profiles), LINKTREE (linkage trees) and the permutation test in BEST (optimal matching of *B*iota to *E*nvironment, including *ST*epwise search). PRIMER v6.1.10 was used for the calculations and graphs of this paper.

3. Results and specific discussion

The analyses are not presented in the contexts and under the hypotheses of the original studies. It is not the purpose of this methodological paper to discuss and interpret the specific data in any detail. They are used merely to illustrate the techniques with realistic examples of possible outcomes. The associated discussion, similarly, focuses on general caveats and corollaries of the three methods and their inter-relationships.

3.1. Global BEST test

3.1.1. Clyde macrofauna

Fig. 3 (top row, left) displays the non-metric multi-dimensional scaling (MDS) plot for the 12 sites on the E-W transect, based on square-root transformed biomass of macrofaunal species and Bray-Curtis dissimilarities between sites. This is an example of a strong gradient of impact, with a complete turnover of the benthic macrofaunal assemblage between the end points of the transect (sites 1 and 12) and the dumpground centre (site 6). Fig. 3 (top, mid) shows the MDS plot for the three contaminant variables selected by the BEST procedure to provide the optimal match (largest Spearman ρ) between biotic and environmental resemblance matrices, using log transformed concentrations and normalised Euclidean distances for the abiotic analysis. Fig. 3 (top, right) is the permutation distribution of ρ under the null hypothesis of complete independence of biotic and environmental patterns. Note that the null distribution is centred at about ρ = 0.25 and not ρ = 0 (because of the inherent selection bias in the search process), and values as large as ρ = 0.7 could be obtained by chance for unrelated patterns. The true ρ value of 0.86, however, is larger than any of the 999 values obtained under random permutation of the site labels, so the null hypothesis of independence can be rejected at the P < 0.001 level (p < 0.1%, as a percentage).

3.1.2. Messolongi diatoms

Fig. 3 (mid row) displays the same three plots for the 17 Greek lagoon sites, based again on root-transformed densities and Bray-Curtis dissimilarities for the 193 diatom species, and normalised Euclidean distance on log-transformed concentrations for the nutrient variables (though temperature, salinity, DO_2 and pH readings are not transformed). The optimal match (ρ = 0.88) of abiotic to biotic patterns, based on 5 environmental variables, is again seen to be extremely good, and this time greatly exceeds the maximum values of around 0.5 observed for the 999 permutations (the null hypothesis of 'no linkage' is rejected at P < 0.001). Note, however, that the null distribution of ρ is again centred some way from zero, reflecting the large number of environmental combinations (1023 for 10 variables) that are trawled through for a spurious match.

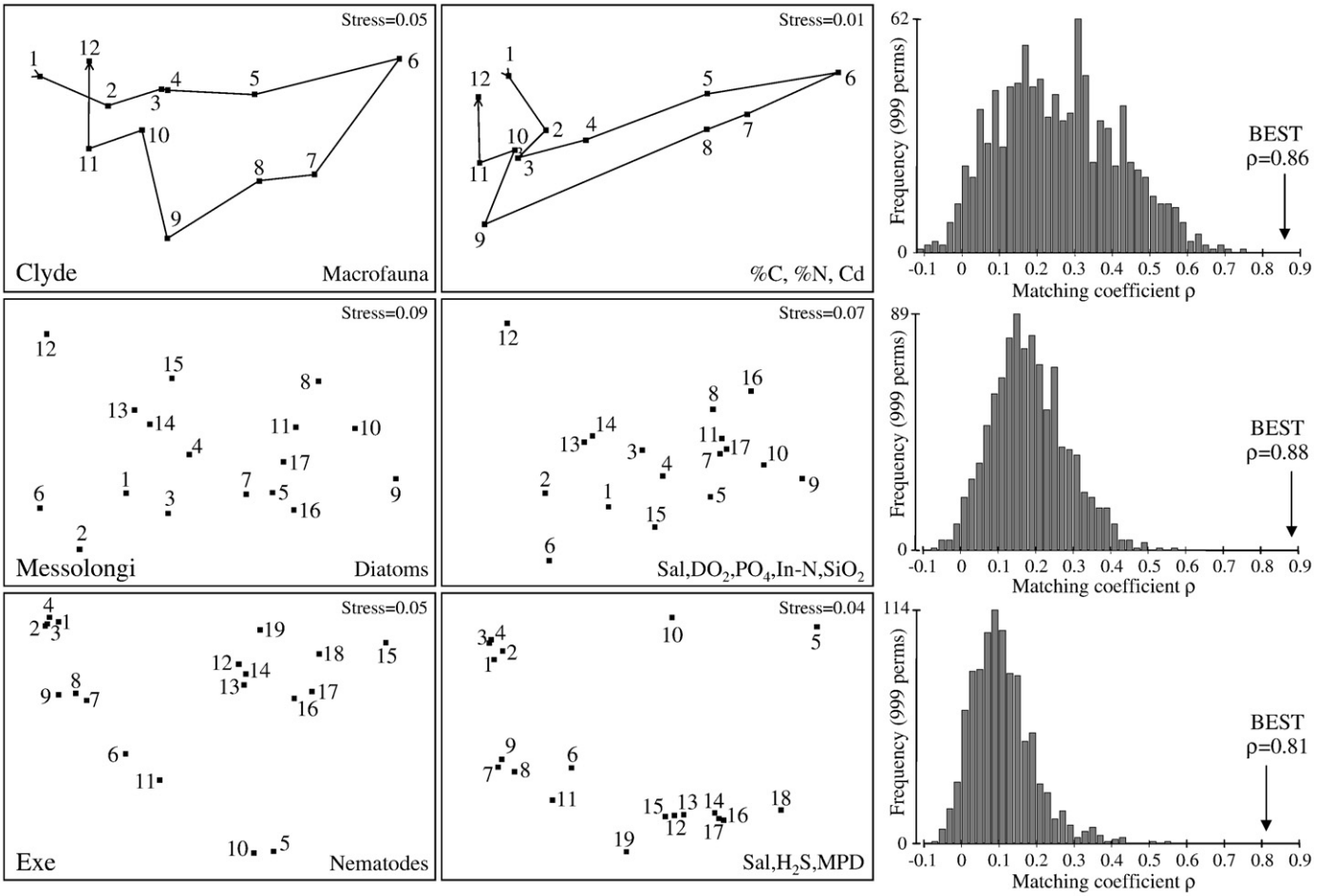

Fig. 3. Matching of biotic composition (MDS site plots, left column) by BEST choice of abiotic variables (MDS site plots, mid column), and permutation distributions (histograms, right column) of global BEST test statistic (Spearman ρ), under null hypothesis of no real biota-environment link. Also shown are real ρ values for displayed matches (P<0.001 in all cases). Data sets are: Clyde dumpground macrobenthos (top row), Messolongi lagoon diatoms (mid row), Exe estuary nematodes (bottom row).

3.1.3. Exe estuary nematodes

Similarly, Fig. 3 (bottom row) shows the MDS plots and null distribution under permutation for the 19 sites, though this time using a fourth-root transformation of the nematode counts and no transformation of the 6 environmental variables. The global BEST test again shows a large (ρ = 0.81) and significant ($P < 0.001$) link between biotic composition and three of the natural environmental factors. A notable difference here is that, because of the small number of abiotic combinations examined (63) and larger number of sites, displaying a clear pattern of biotic clustering in two dimensions, the likelihood of obtaining spuriously large correlations between biotic and abiotic resemblance matrices is substantially reduced, reflected in a lower mean value ($\rho \approx 0.1$) for the null distribution.

A final observation on Fig. 3 is that these three examples were selected because all allowed adequate display of the data in 2-d MDS plots, the stress values being low in all cases (see figure). The excellent description of biotic patterns by the selected environmental variables can readily be visualised therefore, but this is not a necessary condition for successful use of the method: the correlation ρ takes place between the underlying biotic and abiotic resemblance matrices, representing a match between the full-dimensional structure of each set of samples.

3.2. Similarity profiles

3.2.1. Exe estuary nematodes

Fig. 4 (top left) presents the MDS ordination of the 19 sites, based on their nematode communities as before (Fig. 3, bottom left), but now with superimposed symbols representing the group structure defined by hierarchical cluster analysis (Fig. 4, top right), on which similarity profile (SIMPROF) tests have been carried out at many (but not all) nodes of the tree. The dendrogram displays in dashed lines the group structure for which there is no evidence from a SIMPROF test, continuous lines being used for divisions for which SIMPROF rejects the null hypothesis (that samples in that group have no further structure to explore). Fig. 5 (top row) shows the result of the first such test, on all 19 sites. The true similarity profile (top left, bold line) departs markedly from that obtained under permutation, with a large excess of very high and very low Bray-Curtis similarities. The absolute deviation of the true profile from the mean permuted profile (test statistic π = 10.0) is so much larger than any value from the null distribution (Fig. 5, top right) that it will clearly be significant at whatever *a priori* level has been set. (Here only 999 permutations were carried out, so $P < 0.001$ is all that can be claimed, though clearly P would be infinitesimal if a vast number of permutations were used).

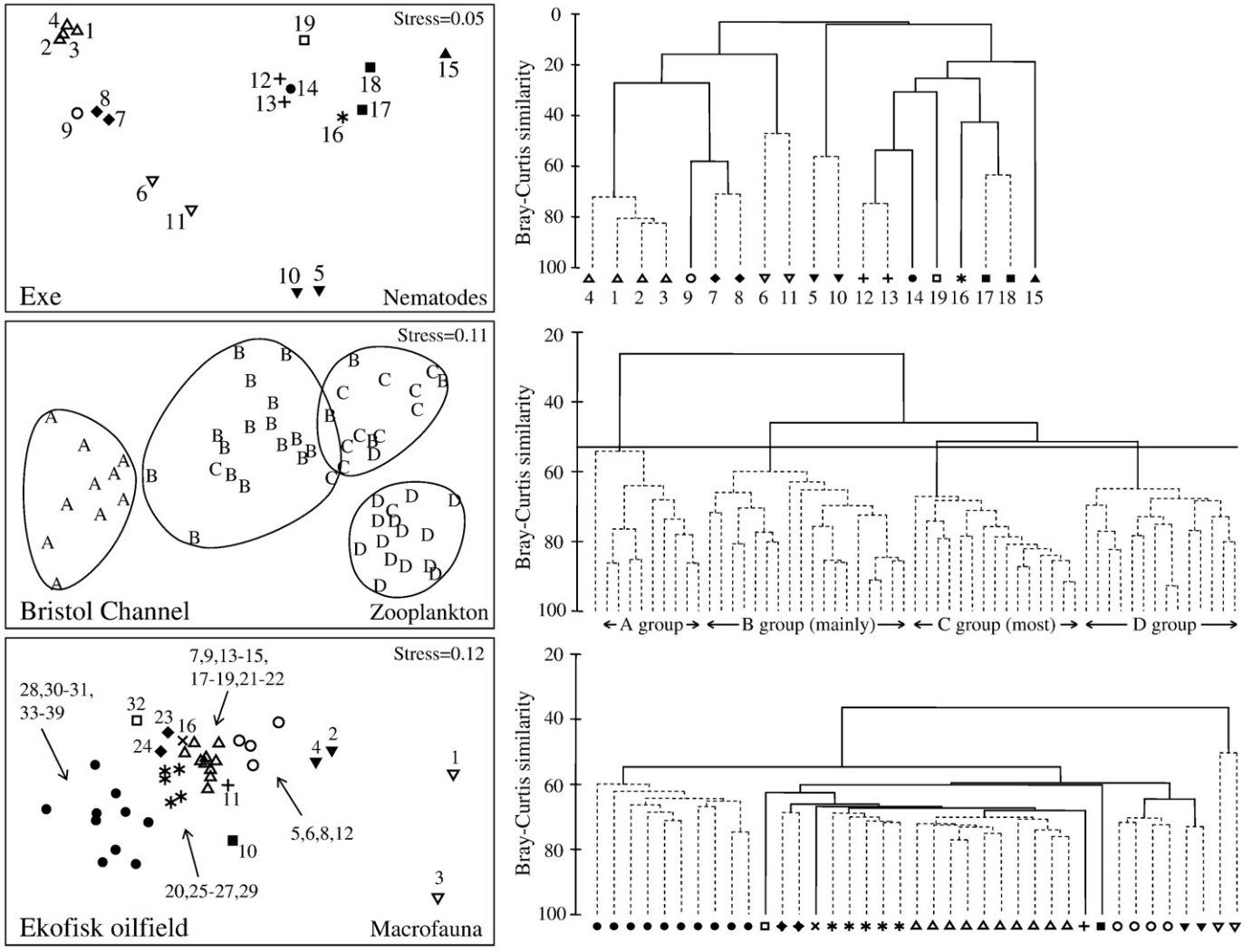

Fig. 4. MDS site plots (left column) showing groups, identified by symbols or boundaries, given by sequence of SIMPROF tests on dendrograms from standard hierarchical clustering (right column). Dashed lines indicate groups of samples not separated (at $P<0.05$) by SIMPROF. Data sets are: Exe estuary nematodes (top row, site numbers shown), Bristol Channel zooplankton (mid row, A-D are salinity ranges, low to high), Ekofisk macrobenthos (bottom row, sites numbered in order of increasing distance from oilfield centre).

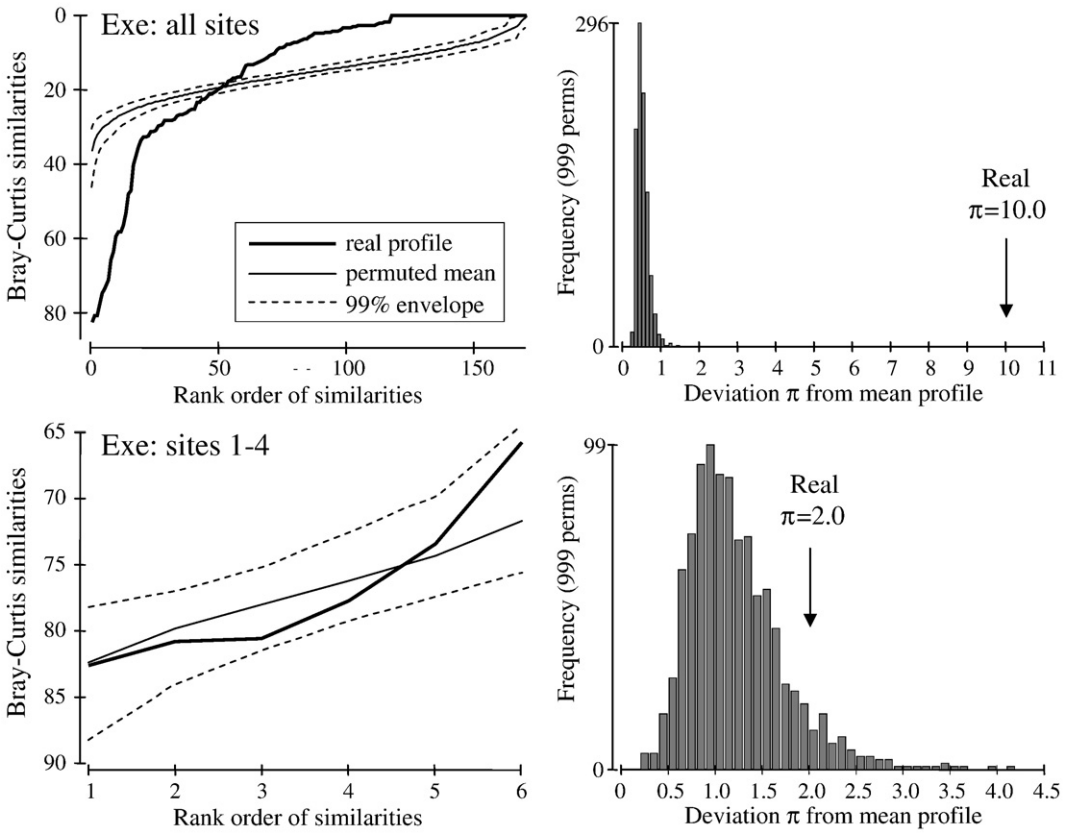

Fig. 5. Exe estuary nematodes: similarity profiles (left column) shown for all 19 sites (top row), and only sites 1-4 (bottom row). Shown are profiles (ordered similarities between sites plotted against their ranks) for real data (bold line), mean of 1000 permuted matrices (fine line) and 99% range for further set of 999 permuted matrices (dashed line). Histograms of absolute deviations π of the 999 permuted profiles from the mean permuted profile (null distributions, right column) are compared to π for the real profile.

The subgroups (5,10,12-19) and (1-4,6-9,11) are then separately tested for homogeneity of composition, leading to rejection in both cases, and so on down the dendrogram until a non-significant result is obtained on each branch. Fig. 5 (bottom row) shows the profiles and test for the group 1-4, for which the true similarity profile is seen to lie within the limits formed by 99% of the permuted profiles, and π = 2.0 ($P \approx 0.09$, insufficient to reject the null hypothesis). Note that the tests at each node were required to attain a significance level of $P < 0.05$ ($p < 5\%$) before further subdivision was made, though an informal Bonferroni-type correction for the multiple testing could have been used (say $P < 0.005$, there being about a dozen tests), in order to erect a larger deviation π that the similarity profile must 'jump over' before each group is deemed to have internal structure.

These data represent a set of samples for which there is much structure and little 'noise'. It always seemed likely from the tight groupings on the MDS plot that at least five different assemblages would be identified, namely sites (1-4), (5,10), (6,11), (7-9), (12-19), but the successive SIMPROF tests show that it is permissible to interpret further subdivision of (7-9) and a finer level of structure altogether in the group (12-19), down to perhaps 6 pairs or singletons: (12-13), (14), (15), (16-17), (18), (19). Two of these decisions are rather borderline and thus to be interpreted cautiously: 16 would not have been split from 17 and 18 at a more conservative significance level ($\pi = 6.6$, $P < 0.03$), and 9 would not have split from 7 and 8 ($\pi = 3.2$, $P < 0.04$). Note, however, that the group (12-14) still divides at $P < 0.001$ ($\pi = 6.9$), highlighting the fact that groups defined by SIMPROF do not always correspond to a slice through the dendrogram at a fixed similarity level, the group (16-18) being formed at a lower average similarity level than (12-14).

Note also one important limitation on the use of SIMPROF: it can never divide a group consisting only of two samples, e.g. sites (6,11) or (5,10), however low their similarity. The similarity profile in that case reduces to a single point (single similarity), and it is clear that the independent permutations of the counts of the species across these two samples can in no way change the value of that similarity, i.e. all the permuted profiles lie at the same point as the real profile.

3.2.2. Bristol Channel zooplankton

Fig. 4 (mid row) displays results for the 57 Bristol Channel sites sampled for zooplankton communities, and is in stark contrast to the previous example, showing a much larger degree of 'noise' and less structure in the SIMPROF tests. Here there is an *a priori* grouping of sites according to ranges in average salinity, denoted by the symbols on the MDS plot (left), groups denoted A to D. As it happens, these groups more or less correspond to a slice at a fixed level of similarity (around 52 or 53%) through the dendrogram (mid, right), as shown by the smoothed convex hulls of the points in each of these dendrogram-based groups (the smoothing uses an unpublished algorithm of one of the authors, RNG). Since the groups A to D are defined *a priori* on the basis of an environmental variable (salinity), it is valid to use an ANOSIM test (Clarke, 1993) and this duly demonstrates major and significant differences between them (global $R = 0.72$, $P < 0.001$). ANOSIM is clearly the better test when there is prior information that can be used to set up a null hypothesis of a specific group structure, but it is also valid to carry out a SIMPROF test on the full set of 57 samples, which also shows clear evidence of structure ($\pi = 6.5$, $P < 0.001$, with no permuted value of π greater than about 1.5). A SIMPROF test could be carried out where a (preferable) ANOSIM test is available but the converse is certainly not true: ANOSIM cannot be performed on groups defined by a cluster analysis of the same data, whatever SIMPROF says about the presence of structure in those samples.

The relevant hypotheses here, however, are about whether there is sub-structure within each of the four main groups, evident from the

cluster analysis and the main salinity divisions. If there is heterogeneity of similarities at this level then it is valid to search for an explanation of sub-structure in terms of finer divisions of the salinity scale or other environmental variables. In fact, the striking conclusion of the SIMPROF tests is of 'no further structure' within these four main groups, the statistics being π = 2.4, 1.0, 1.0, 0.9 ($P < 0.06$, 0.28, 0.34, 0.26) respectively, left to right, low salinity to high salinity groups, in Fig. 4 (mid right). The only borderline case here involves the separation of a single site from the other low salinity sites ($P < 0.06$ from the first test). The SIMPROF test therefore provides a useful check in this case to over-interpretation of the detailed structure seen in the dendrogram, much of which appears to be clustering of random variability.

3.2.3. Ekofisk macrofauna

Unlike the Bristol Channel zooplankton data, where the dendrogram indicates a strong overall grouping of samples, Fig. 4 (bottom row) illustrates the ability of SIMPROF to identify sample structure where change is subtle and continuous. The 39 soft-sediment sites at different distances from the Ekofisk oilfield do not show strong clustering of their macrobenthic communities (Fig. 4, bottom right). The most marked feature on the dendrogram is the separation of two of the sites within 100m of the oilfield, associated with high levels of total sediment PAH. However, SIMPROF tests do provide evidence for several groups in the dendrogram and the MDS ordination (bottom left) differentiates those groups by symbol. Labelling of sites is by their rank order of distance from the oilfield, showing the clear gradient of community change as sites approach the oilfield centre (low numbers). About half a dozen main stages are identified by SIMPROF tests carried out at $P < 0.05$, together with four singleton outliers. Most notable, however, is that 25% of the sites, including nearly all the sites greater than about 4km from the centre, are in a single group (the 'background' assemblage). These are identified as homogeneous by SIMPROF in spite of the large physical distances separating those sites (generally up to 16km, but with one reference site at 30km from the oilfield). This provides compelling evidence for macrobenthic community change out to nearly 4km from the oilfield centre.

3.3. Linkage trees

3.3.1. Messolongi diatoms

Fig. 6 (top) shows the LINKTREE analysis for the 17 sites in the Greek lagoons. This is a dendrogram from divisive clustering of the sites, based on diatom species, using the same Bray-Curtis resemblance matrix as underlies the MDS ordination of Fig. 3 (mid left), but with the partitions constrained by thresholds on three variables

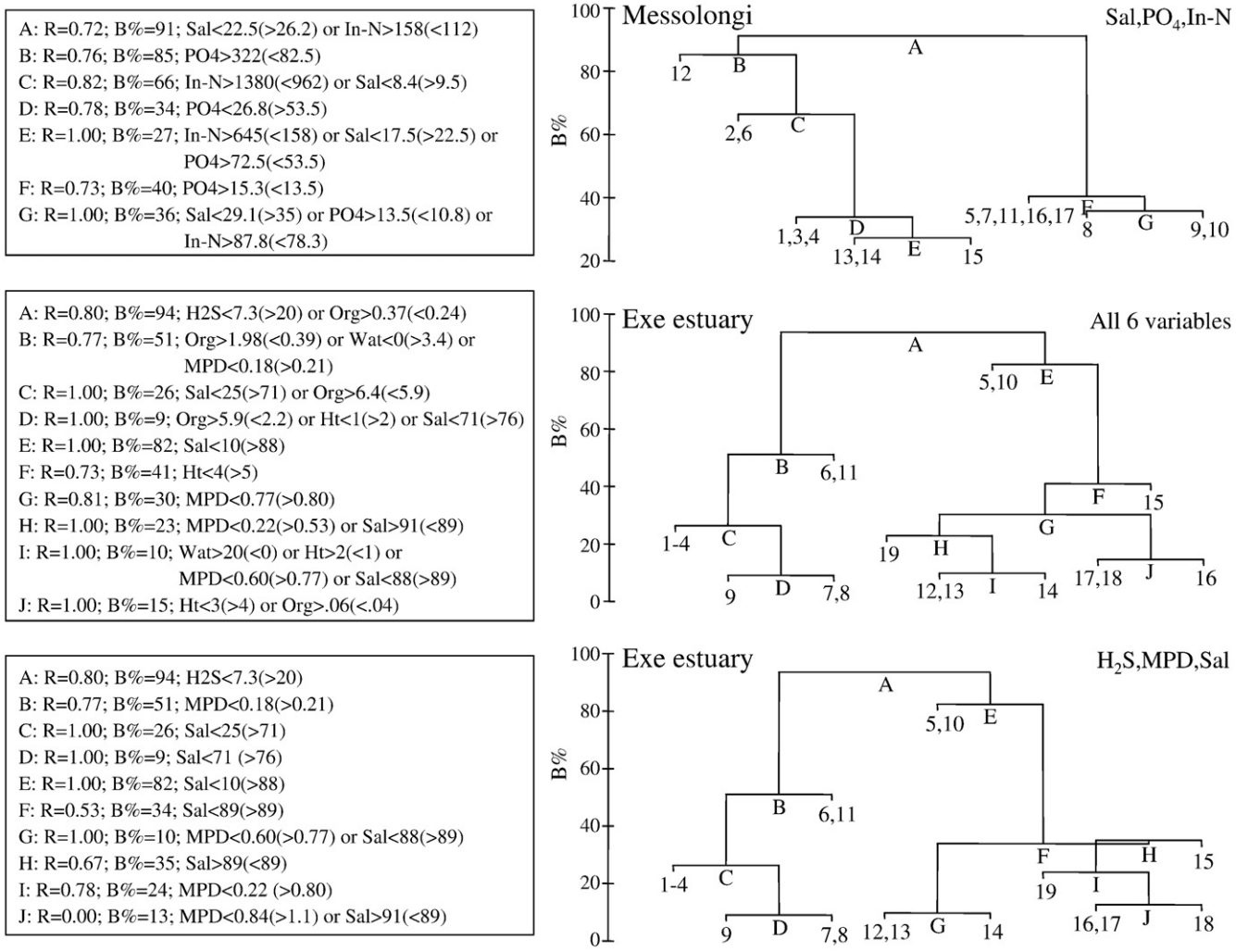

Fig. 6. Linkage tree analysis (LINKTREE), showing divisive clustering of sites from species compositions (right column), constrained by inequalities on one or more abiotic variables (left column). Given for each split is the optimal ANOSIM R value (relative subgroup separation) and B% (absolute subgroup separation, scaled to maximum for first division). For each binary partition (A, B, C,..), first inequality defines group to left, second inequality (in brackets) group to right. Data sets are: Messolongi diatoms linked to water column data (top row), Exe nematodes linked to all 6 (mid row) or only 3 (bottom row) sediment variables.

measured in the water-column: inorganic nitrogen (In-N), phosphate (PO_4) and salinity (Sal). These were selected by the BEST routine as the optimal three-variable combination giving the best match to the diatom assemblage ($\rho = 0.84$), a value almost as high as the overall optimum ($\rho = 0.88$), involving two further variables. (There is strong pressure for parsimony in choice of explanatory variables with a technique such as LINKTREE, because of the large and confusing number of piecemeal explanations that can be obtained with many abiotic variables, the scale for each of which can be sliced at n-1 distinct places, for n samples.)

The first split (A) in the divisive clustering (Fig. 6, top) is between (1-4,6,12-15) and the remaining samples, the sites in the first subdivision all having Sal < 22.5, with Sal > 26.2 for the second group (there are no samples between these two threshold levels). An alternative explanation is that sites in the first group all have In-N > 158 and for the second group In-N < 112. (Note the convention that the first inequality always describes samples to the left of the division and the second, in brackets, refers to samples on the right.) There is, of course, no way of choosing between those two alternative descriptors since both define the same split in the biotic samples, giving an optimal R of 0.72, with the division displayed on the y axis scale at B% = 91. The next division (split B) separates site 12 from the others in the first subgroup on the basis of its much higher PO_4 reading, and so on (see Fig. 6, top left, for a full listing of inequalities). Before each new division is attempted (including the first), a SIMPROF test is run on the diatom data for the current set of sites, to establish that there is heterogeneity in the similarities among those samples, which would justify seeking further subdivision. For example, for the full set of 17 sites, $\pi = 5.7$, $P < 0.001$; for the separation of site 12, $\pi = 4.1$, $P < 0.001$, and so on, down to the potential division of the group (1,3,4), which is not carried out because $\pi = 1.2$, $P = 0.56$, or to the group (5,7,11,16,17), which is not divided because $\pi = 1.0$, $P = 0.35$. A criterion of $P < 0.05$ was used in this case to indicate whether to carry on the subdivisions, and a fairly relaxed criterion like this seems advisable in most exploratory contexts, but note that a modest adjustment for multiple testing would have made no difference here, the least significant P value for a division shown in Fig. 6 being $P < 0.007$ for the separation of site 15 from (13,14). It is also true, in this case, that exactly the same dendrogram would have been obtained if all 5 abiotic variables selected by the BEST routine (Fig. 3) had been used to constrain the possible divisions. What changes is the multiplicity of alternative explanations for each of these splits, e.g. in split B, in addition to much higher phosphate, site 12 has higher SiO_2 and lower DO_2 than sites (1-4,6,13-15).

3.3.2. Exe estuary nematodes

Fig. 6 (mid row) shows the LINKTREE analysis for the nematode communities at 19 sites in the Exe estuary, utilising all 6 abiotic variables: depth of anoxic layer (H_2S), interstitial salinity (Sal), median particle diameter (MPD), %organics (%Org), height up the shore (Ht) and depth of the water table (Wat). Again a stopping rule of $P > 0.05$ was used in the SIMPROF routine, in order to be consistent with SIMPROF testing in the agglomerative clustering of Fig. 4 (top). In fact, precisely the same divisions result in this case, and LINKTREE now provides us with (multiple) 'explanations' for the observed biotic clusters. Sites (1-4,6-9,11) have a shallower H_2S layer, or higher % organics, than the remaining sites (split A), sites 5 and 10 have a much lower salinity than the group 12-19 (split E), etc. These patterns could have been ascertained in this case by multiple bubble plots (Fig. 7), displaying the environmental variables one at time as circles with sizes dictated by the abiotic values, placed at the relevant sites in the biotic MDS ordination. Thus, H_2S can be seen to 'explain' the left-right division in the MDS plot, MPD the distinction between site groups (6,11) and (1-4,7-9), and Sal clearly separates the (5,10) and (12-19) groups. Bubble plots are only of use, however, if the MDS plot has a sufficiently low stress to give an accurate representation of the community structure. This is true here (2-d stress = 0.05) but will often not be the case, LINKTREE then providing a more general alternative, not based on low-dimensional approximations but using the full-dimensional resemblance matrix.

Of course, though it is true for the Exe estuary plots of Fig. 4 (top) and Fig. 6 (mid), there is no guarantee that the specific unconstrained agglomerative (UPGMA) and constrained divisive (LINKTREE) clustering methods used here will always produce the same tree, even if starting from the same dissimilarity matrix and using the same critical P values in the SIMPROF tests. They utilise slightly different averages for example (respectively, dissimilarities between groups on original scales, and rank dissimilarities between groups, re-ranking at each division). The bigger difference between them, however, is that LINKTREE is constrained to consider only divisions which can be expressed as a threshold on an environmental variable. If this forces a less natural split of the samples

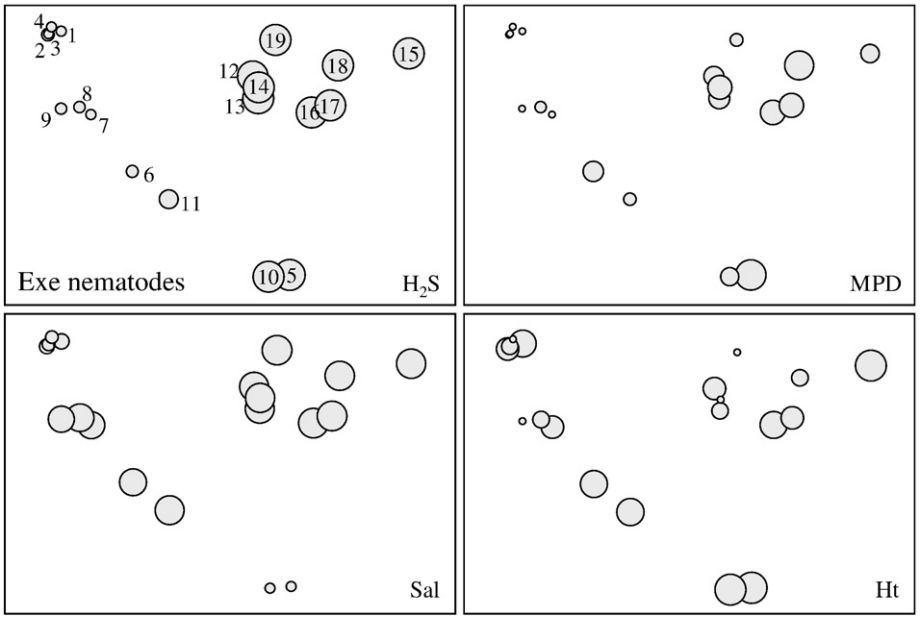

Fig. 7. Exe estuary nematodes: MDS 'bubble' plots of 19 sites with superimposed values of depth of H_2S layer (H_2S), median particle diameter (MPD), interstitial salinity (Sal) and height up the shore, intertidally (Ht), as circles of different sizes.

than would be obtained from unconstrained clustering of the community compositions then not only will a different dendrogram result, but the linkage tree might also exhibit 'reversals'. That is, two groups might be split at a lower level of B% (average rank dissimilarity between the groups on original ranks, appropriately scaled) than a subsequent subdivision of one of those groups, and the nodes of the tree will not fall in their natural order on the y axis of the LINKTREE plot. This is illustrated in Fig. 6 (bottom), where the LINKTREE procedure has been carried out with just the three environmental variables, H_2S, Sal, MPD, selected by the BEST routine (Fig. 3, bottom). Though the trees are broadly similar, the order in which the group (12-19) is now subdivided differs, its initial division taking place at lower B% (34), with separation first of sites (12-14) from (15-19), before 15 is divided from (16-19) at the slightly higher B% of 35 (a reversal). The reason here, as will usually (perhaps always) be the case, for the observed reversal is that the natural, unconstrained split of the assemblage structure is of 15 from the remaining sites in the group (12-19). The only abiotic variable that allows this split is Ht, site 15 being at the extreme of the inter-tidal range (see the bubble plot of Fig. 7), but this variable was excluded from the current LINKTREE analysis because the BEST procedure deemed it irrelevant to the characterisation of these sites. It did so because sites spanning a wide range of intertidal positions elsewhere, e.g. sites 1-4, have indistinguishable assemblage structures. In consequence, a less natural split (lower B%) is made prior to a more natural one (higher B%) and, given that there is evidence from SIMPROF of structure in the samples that requires explanation (rather than just random variability), such a reversal could indicate the absence of a needed explanatory variable.

4. General discussion

4.1. Global BEST test

In an indirect way, the BEST routine is trying to solve the same problem for multivariate community data as standard multiple regression does for single response variables (though more direct multivariate analogies of multiple regression are given by the dbRDA and DISTLM procedures of Legendre and Anderson, 1999, and McArdle and Anderson, 2001, since these employ explicit linear models). In multiple linear regression, 'all subset' regressions and basic stepwise selection methods (Efroymson, 1960; Draper and Smith, 1981) are well-known to provide outcomes in which the final F value reported for the statistical significance of the optimal model yields a biased, non-conservative test (Copas, 1983). Insufficient allowance is made for the repeated testing and multiplicity of parameters estimated in such linear models, and this has spawned a number of now widely-used modifications which, broadly speaking, add a 'penalty' to the function being minimised (some form of the residual sums of squares) which increases with the number of parameters in the model, thus balancing 'goodness of fit' with parsimony of the model. The genesis of these techniques (the Akaike criterion, 'Bayes' information criterion etc., Akaike, 1974; Schwarz, 1978) is parametric, large-sample, likelihood-ratio theory.

The BEST matching procedure cannot go down this route, since it does not invoke any form of parametric linear model. What it does have, however, is an overall significance test for the final subset of variables, which does allow fully, and robustly, for the selection bias inherent in examining many combinations of variables. This exploits the idea inherent in all permutation tests, that the fitting procedure (here, optimisation of a matrix correlation, ρ) is repeated many times on data permuted to represent conditions under the null hypothesis (of no genuine match). Of course, this general paradigm is equally applicable to the 'penalised fitting' routines above and, if computationally feasible, might also be worthwhile in the multivariate analogues of these linear models (McArdle and Anderson, 2001), as suggested in Anderson and Gorley (2008).

The large rank correlations ρ that could be obtained purely by chance in the Clyde macrofauna example make it very clear that the BEST test should always be carried out whenever the BEST matching routine is performed (on independently collected matrices) as a safeguard against over-interpretation. When there are many variables, a great many combinations of them are possible and the selection bias in examining them all can be extreme. If the explanatory variable set is to be reduced prior to the BEST analysis, to restrict the multiplicity of explanatory models, it should be on the basis of examining the abiotic set only. For example, it would be sensible to remove all but one of each set of variables that are very highly correlated with each other, and thus carrying essentially the same information (effective multicollinearity), see Clarke and Ainsworth (1993). Another way of achieving this pruning might be to use the BEST procedure to select a minimal-sized subset of abiotic variables that 'stands in' for the full abiotic set, in terms of replicating the full abiotic pattern of samples, with a matching coefficient ρ in excess of 0.95, say. It is clear that the global BEST test has no role in this latter context and should never be carried out on two matrices that involve some of the same data, since the null hypothesis of complete independence of patterns is never viable. Similar comments apply to the use of BEST to find a subset of species which 'stand in' for the full set, in terms of replicating the pattern of samples from the full biotic matrix (Clarke and Warwick, 1998).

The high ρ values obtainable by chance for the Clyde study (up to 0.7) are also a result of the low number of samples and their very simple multivariate structure. Samples are largely positioned on one single, strong gradient and under the permutation of labels (thus scrambling any true biotic-abiotic relationship) any chance combination of variables which tends to put a nominated half of the points to one side of the plot, and the other half to the other side, will result in a moderately large ρ, purely by chance. The analyses of Messolongi diatoms and Exe nematodes are more reassuring: there are a modestly greater number of sites (17 and 19, rather than 12), which more than doubles the number of similarities being matched, and the sample relationships for the biota are now more complex, though still largely 2-dimensional. The net result is to limit the ρ values that can be obtained by chance, to no more than 0.4 to 0.5. These are still relatively high values, however, compared with some that have been reported and interpreted in the scientific literature, on similar-sized sets of samples, so the importance of a global test of the null hypothesis of 'no relationship at all' is evident.

4.2. SIMPROF test of samples

On the mechanics of the SIMPROF test, it might be thought unnecessary to calculate two separate sets of permuted similarity profiles (of sizes $T = 1000$ and $S = 999$, say) and, indeed, utilising the same set of permutations to calculate the mean profile as are then used to calculate expected departures from the mean, would make very little difference to the outcome of the test. Re-use of the same permutations for both parts of the calculation, however, will very slightly bias the distribution of π in favour of lower values than expected under the null hypothesis, so is best avoided. (This can be seen by imagining that T is only 2 rather than 1000, the two permuted profiles then being demonstrably closer to the mean calculated from them than would be expected, for genuinely independent profiles representing the null condition.) At an even more technical level, the general maxim of permutation tests – that exactly the same procedure should be performed on each set of permuted data as was applied to the real profile – might suggest that a new mean profile (first set) should be calculated for comparison with each new individual profile (second set). However, for fixed total computing time this is readily seen to be inefficient in relation to the procedure adopted here. Of more practical importance to note is that, like all tests using random subsets of the full set of possible permutations, quoted P values are

only determined to a certain level of accuracy, so slightly different outcomes to a sequence of tests such as illustrated in Fig. 4 (right) might be obtained on different runs. Simple binomial calculations show that for tests which should give a P value in the region of 0.05, the actual P returned will vary over the range (0.035, 0.065), for S (and T) = 1000 permutations. Too much concern about precision of reported P values is unwarranted, but note that 10000 permutations would reduce this range to (0.045, 0.055).

It is worth emphasising that, as usual, a permutation procedure provides a realistic solution to the problem of generating conditions under the null hypothesis. An alternative might have been to simulate values of variables randomly and independently for each sample, drawn from some underlying parametric distribution, and to construct similarity profiles from these simulations (a Monte Carlo test procedure). This has the twin difficulties, however, of needing to justify a particular parametric distribution for species counts (say), a non-trivial problem, and estimating realistic values for those parameters across species. Some species are much more abundant than others, or present at greater frequency, and the counts for some species are much more spatially clumped than for others, leading to higher variance-to-mean ratios (e.g. Fisher et al., 1922; Greig-Smith, 1952; Clarke et al., 2006a). All of these differences, however, are exactly represented in the permutation procedure, because precisely the same set of counts are used for each species as observed in the original matrix but simply permuted across the samples, under the conditions of the null hypothesis.

Permutation tests have their limitations, naturally. It has already been noted that it is not possible to obtain a SIMPROF test of whether two samples have genuine multivariate structure (i.e. can be split into two groups or are homogeneous). It must also be true that, all else being equal, the power of the test to detect structure will tend to increase as the number of samples increases, so that the profile displays a richer set of similarities. (That it can, surprisingly perhaps, have power to split a group of three samples into a pair and a singleton is evident from the example on Exe estuary nematodes, where several such divisions are made.) Furthermore, it is clear that no SIMPROF test can be carried out for a single variable, all permutations of the single set of species values across samples then providing the same set of sample resemblances. The test does not work by exploring the spacing of samples along a line to infer the presence of a mixture of distributions, since any such technique would require modelling of the components making up the mixture and would be sensitive to these distributional assumptions, a well-known difficulty in the decomposition of univariate mixtures (Everitt and Hand, 1981). Instead, the SIMPROF test exploits the multivariate structure of the data, noting that a clustering of samples forces 'co-relationships' between variables (the term correlation being avoided here because species associations are sometimes better captured by other resemblance measures than standard correlation). Conversely, association amongst variables induces structure in the samples (though not necessarily group structure). At its simplest, this can readily be seen by envisaging a block-diagonal matrix in which groups of species are jointly present at a set of sites and jointly absent at another set of sites, forcing well-defined group structure on the samples. More subtly, smoothly varying relationships between species, synergistic and/or antagonistic, induce a smooth gradient in samples which will again be evidenced by an excess of high dissimilarities (from samples at opposite ends of the gradient) in relation to the similarity profiles under permutation. The test therefore has a capacity to detect gradients over samples, in addition to group structure, though whether a variation of the test statistic might have greater power when testing for the specific alternative hypothesis of seriated samples (Clarke et al., 1993) would bear further examination. (Note, however, that a formal definition of 'power' for such data is complex, given the difficulty in specifying alternative hypotheses in multivariate space. Realistic options are restricted to simulations on specific datasets, Somerfield et al., 2002).

It is therefore the duality between co-relationships amongst variables and structure in samples that is being exploited by the SIMPROF test. This has three corollaries: (i) a test is not possible for a single variable; (ii) the greater the number of variables the more powerful the test is likely to become, all else being equal; (iii) the sample structure identified by a significant SIMPROF test could be rather minimal, and not necessarily biologically important to interpret. What is unarguable, however, is that if the SIMPROF test fails to reject the null hypothesis of 'no structure', from at least a handful of samples and variables, then there can be little justification for attempting to explore sub-structure within these samples. There is an asymmetry here in the interpretation: the groups that SIMPROF identifies may be at too fine a level of detail for practical classification (e.g. for categories of water quality status, habitat classification etc). It is then entirely appropriate to define coarser groupings, e.g. of the sites categorised by a slice through the dendrogram at some arbitrary level of similarity, provided that the resulting clusters are always supersets of the SIMPROF groups. There would then be statistical evidence for such coarser groupings being interpretable, and not the product of random chance. What is harder to justify, and arguably not permissible at all, would be a finer-scale grouping than indicated by SIMPROF. In other words, SIMPROF erects a hurdle over which one must jump before further interpretation is pursued. It may be argued that this is a very low hurdle but the surprise from practical experience is how often one is tripped up! For example, for the 57 sites in the Bristol Channel zooplankton data of Fig. 4 (middle row), SIMPROF identifies only four main clusters of sites, with no substructure indicated within any of those (even using a non-conservative $P < 0.05$ criterion for statistical significance of each of the multiple tests). The test is an effective deterrent in this case to over-interpretation. There is no need to seek a supplementary environmental variable to explain the variation shown in the MDS plot (Fig. 4, mid left), running orthogonally to the main salinity gradient, since SIMPROF suggests that this is purely random spatial and sampling variability.

4.3. SIMPROF for similarities amongst variables

Note that the SIMPROF procedure could be used to examine similarity profiles among *variables*, rather than the more usual similarity matrices calculated for multivariate data, which are among samples. This is not simply a transposition of the axes of the matrix in the SIMPROF routine, since randomisation would logically again be based on permuting the entries for each variable, e.g. species, across all samples. (Conditions under which it would make sense to permute entries for each sample across variables must be rather rare, since this corresponds to an assumption of equally abundant species, or equally frequently-occurring species in the case of presence/absence data. For environmental data, or metrics or biomarkers, with their different measurement scales, it seems even more unlikely that there exist meaningful null hypotheses permitting permutation across variables). Independent permutation of the values of species across samples would give profiles for 'species similarities' which address the issues of species intercorrelation, either direct dependency through competitive interactions or synergies or, more likely for most observational data sets, through a common or antithetic response to particular environmental gradients. The choice of appropriate pre-treatment (standardisation, transformation etc) and similarity measure, and the precise definition of the hypotheses consequently being tested, merits further study.

What is clear, however, is that applying the SIMPROF procedure to correlation matrices computed over a wide set of physico-chemical variables (or diversity metrics, biomarkers etc) does provide a global test of whether there are *any* statistically significant correlations among these variables. This specific suggestion can be found for estuarine physical variables in Potter et al. (2001), termed 'correlation profiles'. It is often the case that particular pairwise correlations appear significant when referred to standard tables for Pearson r or

Spearman ρ (or other coefficient), but there may be many other non-significant correlations in a triangular correlation matrix, and the question naturally arises as to whether the larger observed correlations could have been obtained by chance, under conditions in which there are no real correlations present. This is a problem of multiple testing but is particularly acute here because the set of computed correlation values are highly interdependent, the same values of each variable being used many times over (correlation of variable 1 with 2, 1 with 3, 1 with 4, 2 with 3, 2 with 4, 3 with 4 etc). A simple Bonferroni correction for the number of tests performed will therefore be entirely inadequate to this task. A SIMPROF test, on the other hand, neatly sidesteps the multiple testing by calculating only one test statistic (π) and one P value for its global test of the null hypothesis of no correlations between any of the variables. Furthermore, it uses a permutation technique to derive the null distribution and thus conditions on the observed (marginal) distributions of each variable, without needing any parametric (e.g. normality) assumptions, as described above. It is therefore suggested that it may often be desirable to to use these permutation P values for bivariate data, i.e. single correlation coefficients, in preference to standard correlation tables. (Note that version 6 of PRIMER does implement the SIMPROF permutation procedure on both sample and variable similarities, including a range of correlation coefficients.) Unlike permutation of variables to examine a single similarity between two samples, permuting variables to examine a single similarity (correlation) between two *variables* will lead to a useable test. Each permutation does provide a different (null) value of the correlation between variables. The 'similarity profile' is just the single observed correlation, and the SIMPROF π statistic is the absolute value of that correlation (at least for ≥ 10000 permutations, say, so that the mean under the null hypothesis is effectively zero). The permutation distribution of the π statistic can therefore be used to give an exact P value for 2-tailed testing of the null hypothesis that the correlation between two variables is zero, without the need for any distributional assumptions.

4.4. Linkage trees

The contrast between the LINKTREE analyses carried out for the Exe estuary data in Fig. 6 (mid and bottom) and the BEST procedure for the same data in Fig. 3 (bottom) highlights a general point of some importance. The BEST procedure identifies three environmental variables (H_2S, Sal, MPD) as providing the best description of the nematode communities, but these three variables produce a LINKTREE dendrogram (Fig. 6, bottom) which does not convincingly account for the initial separation of site 15 from the remaining sites in the group (12-19), observed in the unconstrained cluster analysis of Fig. 4 (top). This is also seen in the 'reversal' in the linkage tree involving site 15, a clear indication of disparity between an unconstrained division of the nematode assemblages and one which is constrained by the current set of abiotic variables. The problem arises here because the 'height up the shore' variable (Ht) was not included in the explanatory set. When it is (Fig. 6, mid) the unconstrained and constrained trees are identical, and an abiotic 'explanation' is provided for the initial separation of site 15. (Whether this is the true explanation is, of course, an entirely different matter! Nothing in such observational studies can ever demonstrate causality of the correlative links found between biotic and abiotic patterns).

The general point is this: the Ht variable was not selected by the BEST routine because BEST is global in application and implicitly assumes additive effects of the environmental variables across the full range of sites. If Ht seems unimportant in structuring the relationships between nematode assemblages for most of the sites (and it does span a wide range of values for sites 1-4, for example, for which the communities are indistinguishable), but then appears important in drawing a distinction in another part of the space, then BEST will understandably be in some confusion about the relevance of this variable. Informally, the matching process works on the basis that if two sites have differing values for the suite of driving abiotic variables then they will exhibit differing community structure, and for the same suite of values they will have similar communities (Clarke and Ainsworth, 1993). There is an implicit assumption here of additivity. BEST will not respond well to situations in which abiotic variables strongly interact, a feature which is also true of the way the so-called 'canonical' methods are commonly used. As De'ath (2002) points out, the canonical techniques for describing assemblage patterns by environmental variables are based on specific models of response, e.g. linear for canonical correlation (Krzanowski, 2000) and unimodal Gaussian in canonical correspondence (ter Braak, 1986), and both of these latter techniques employ linear, additive combinations of the submitted environmental variables. The more informal BEST procedure is less constrained, in that linearity is clearly not necessary, but additivity is assumed, implicitly.

For the Exe estuary data, BEST does identify three abiotic variables with very good predictive properties for the broad structure of these nematode communities ($\rho = 0.81$), so their effects must be largely additive, but LINKTREE, with its purely local focus, has the ability to mediate that interpretation within specific subgroups of sites. It follows that there will also be situations in which the global BEST procedure fails to provide a convincing explanation for an assemblage pattern at all, even though the driving environmental variables have been recorded, because the way the abiotic variables combine to produce their effects may be highly interactive. Some progress might be made in this case, whilst retaining a largely holistic approach, by applying the BEST procedure, and the BEST test, separately within groups (and sub-groups) of samples. This would allow for the possibility of large-scale interactions across groups but assume local additivity within groups.

A radically different alternative is provided by LINKTREE and related techniques (De'ath, 2002), which could make some sense of those situations where other methods fail, precisely because they seek only local explanations. The piecemeal approach is specifically designed to cope with interactions: if 'height up the shore' (Ht) has an effect on community structure for coarse sediments but no effect at all in fine sediments, even though there are sites with coarse and fine sediments having the same range of shore heights, then this is an interaction of Ht and median particle diameter (MPD). LINKTREE can handle this structure straightforwardly because it has no need to invoke mention of Ht when subdividing groups within the fine-sediment sites, but can utilise it in separating sites with high MPD. Furthermore, the inequalities mean that it 'models' the response as a threshold effect rather than a steady linear or monotonic change, further increasing the flexibility.

Two other attractive features of the CART/MRT approach (Breiman et al., 1984; De'ath, 2002), which apply equally to LINKTREE, are as follows. Firstly, monotonic transformations of the explanatory variables make no difference whatsoever to the outcome, a result of the use of thresholds. For example, if $x>y$ then $\log(x)>\log(y)$, so the n sites are divided into exactly the same n-1 binary divisions by thresholds on salinity (Sal) as they are by thresholds on log(Sal). This makes the procedure nicely non-parametric but it would normally be convenient to describe the outcome in terms of the original abiotic scales of measurement. Secondly, there is some tolerance to missing data in the explanatory variables for some sites, because of the local mode of operation in LINKTREE. If a variable is unavailable for one or more of the sites currently under division, then it cannot be used, but one of the resulting subgroups may have complete data for this variable so it can be reinstated in choosing constraints for further subdivisions.

There is a price to pay for all this generality, however, and it is often a heavy one, namely that of overly detailed and 'intimate' explanations, in great multiplicity, especially when using moderately large numbers of explanatory variables and relatively few samples. Many of these variables will be able to replicate each other in explaining

particular subdivisions of the samples, especially at the finer levels of the hierarchy, involving small numbers of samples. This can already be seen with the 6 abiotic variables for the Exe estuary study, subdividing 19 samples (Fig. 6, mid). Most of these divisions now have multiple explanations, in contrast with the simpler thresholds, but somewhat less satisfactory clustering, for 3 abiotic variables (Fig. 6, bottom).

In fact, there is a real danger here of 'chasing the noise' in the detailed structure of the ordinations rather than standing back to look at the global structure, as BEST and the canonical methods do. To a large extent, however, this temptation to over-interpret is counteracted by the testing structure provided by similarity profiles, which give stopping rules for the divisive clustering process. A new binary subdivision of an existing group is only made if the null hypothesis for the SIMPROF test is rejected, indicating the presence of at least some structure, however marginal, in the group of sites under consideration.

4.5. Concluding remark

The perennial tension between analyses which are confirmatory (or, more properly speaking, falsification-based) and those which are exploratory, can be characterised as that of whether a paper ends with a significance test or starts with it. In the spirit of the latter, we suggest that the SIMPROF and global BEST tests described here may have some small use in stiffening exploratory studies of observed gradients with a backbone of null-hypothesis testing, and feel sure that this is an aim of which John Gray would have approved.

Acknowledgments

We thank the referees and the guest editor (RMW) for their most helpful and positive comments. This work is a contribution to the biodiversity component of the Plymouth Marine Laboratory's core strategic research programme. It was supported by the UK Natural Environment Research Council (NERC) and the UK Department for Environment, Food and Rural Affairs (DEFRA) through the AMBLE project ME3109. KRC acknowledges his position as honorary fellow of the Plymouth Marine Laboratory and of the Marine Biological Association of the UK, and would like to record the far-reaching influence that John Gray had on his early research career and scientific interests and aspirations. **[SS]**

References

Akaike, H., 1974. New look at statistical model identification. IEEE Trans. Autom. Contr. 19, 716–723.
Anderson, M.J., Gorley, R.N., 2008. PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods. PRIMER-E, Plymouth.
Bayne, B.L., Clarke, K.R., Gray, J.S. (Eds.), 1988. Biological Effects of Pollutants: The Results of a Practical Workshop. Mar. Ecol. Prog. Ser, 46, p. 278.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont, California, p. 358.
Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18, 117–143.
Clarke, K.R., Green, R.H., 1988. Statistical design and analysis for a 'biological effects' study. Mar. Ecol. Prog. Ser. 46, 213–226.
Clarke, K.R., Ainsworth, M., 1993. A method of linking multivariate community structure to environmental variables. Mar. Ecol. Prog. Ser. 92, 205–219.
Clarke, K.R., Warwick, R.M., 1998. Quantifying structural redundancy in ecological communities. Oecologia 113, 278–289.
Clarke, K.R., Warwick, R.M., 2001. Change in marine communities: an approach to statistical analysis and interpretation, 2nd edn. PRIMER-E, Plymouth.
Clarke, K.R., Gorley, R.N., 2006. PRIMER v6: User Manual/Tutorial. PRIMER-E, Plymouth.
Clarke, K.R., Warwick, R.M., Brown, B.E., 1993. An index showing breakdown of seriation, related to disturbance, in a coral-reef assemblage. Mar. Ecol. Prog. Ser. 102, 153–160.
Clarke, K.R., Chapman, M.G., Somerfield, P.J., Needham, H.R., 2006a. Dispersion-based weighting of species counts in assemblage analyses. Mar. Ecol. Prog. Ser. 320, 11–27.
Clarke, K.R., Somerfield, P.J., Chapman, M.G., 2006b. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. J. Exp. Mar. Biol. Ecol. 330, 55–80.
Collins, N.R., Williams, R., 1982. Zooplankton communities in the Bristol Channel and Severn Estuary. Mar. Ecol. Prog. Ser. 9, 1–11.
Copas, J.B., 1983. Regression, prediction and shrinkage. J. Roy. Statist. Soc. B 45, 311–354.
Danielidis, D.B., 1991. A systematic and ecological study of diatoms in the lagoons of Messolongi, Aitoliko and Kleissova. Ph.D. dissertation, University of Athens.
De'ath, G., 2002. Multivariate regression trees: a new technique for modeling species environment relationships. Ecology 83, 1105–1117.
Draper, N., Smith, H., 1981. Applied Regression Analysis, 2nd ed. Wiley, New York.
Efroymson, M.A., 1960. Multiple regression analysis. In: Ralston, A., Wilf, H.S. (Eds.), Mathematical Methods for Digital Computers. Wiley, New York.
Ellingsen, K.E., Gray, J.S., 2002. Spatial patterns of benthic diversity - is there a latitudinal gradient along the Norwegian continental shelf? J. Anim. Ecol. 71, 373–389.
Everitt, B.S., Hand, D.J., 1981. Finite mixture distributions. Chapman and Hall, London. 143 pp.
Fisher, R.A., 1925. The goodness of fit of regression formulae, and the distribution of regression coefficients. J. Roy. Statist. Soc. 85, 597–612.
Fisher, R.A., Thornton, H.G., MacKenzie, W.A., 1922. The accuracy of the plating methods of estimating the density of bacterial populations, with particular reference to the use of Thornton's agar medium with soil samples. Ann. Appl. Bot. 9, 325–359.
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.
Gray, J.S., Clarke, K.R., Warwick, R.M., Hobbs, G., 1990. Detection of initial effects of pollution on marine benthos: an example from the Ekofisk and Eldfisk oilfields, North Sea. Mar. Ecol. Prog. Ser. 66, 285–299.
Greig-Smith, P., 1952. The use of random and contiguous quadrats in the study of the structure of plant communities. Ann. Bot. 16, 293–316.
Hope, A.C.A., 1968. A simplified Monte Carlo significance test procedure. J. R. Stat. Soc. Ser. B 30, 582–598.
Kendall, M.G., 1970. Rank correlation methods. Griffin, London.
Krzanowski, W.J., 2000. Principles of multivariate analysis: a user's perspective, revised edn. Oxford University Press, Oxford.
Legendre, P., Legendre, L., 1998. Numerical ecology, 2nd English ed. Elsevier, Amsterdam.
Legendre, P., Anderson, M.J., 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecol. Monogr. 69, 1–24.
Mantel, N., 1967. The detection of disease clustering and a generalized regression approach. Cancer Res. 27, 209–220.
McArdle, B.H., Anderson, M.J., 2001. Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82, 290–297.
Pearson, T.H., Blackstock, J., 1984. Garroch Head sludge dumping ground survey. Final contract report: Dunstaffnage Marine Research Laboratory, Oban.
Potter, I.C., Bird, D.J., Claridge, P.N., Clarke, K.R., Hyndes, G.A., Newton, L.C., 2001. Fish fauna of the Severn Estuary. Are there long-term changes in abundance and species composition and are the recruitment patterns of the main marine species correlated? J. Exp. Mar. Biol. Ecol. 258, 15–37.
Schwarz, G., 1978. Estimating dimension of a model. Ann. Statist. 6, 461–464.
Somerfield, P.J., Clarke, K.R., Olsgard, F., 2002. A comparison of the power of categorical and correlational tests applied to community ecology data from gradient studies. J Anim Ecol 71, 581–593.
ter Braak, C.F.J., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67, 1167–1179.
Warwick, R.M., 1971. Nematode associations in the Exe estuary. J. Mar. Biol. Assoc. U.K. 51, 439–454.
Whittaker, R.H., 1956. Vegetation of the Great Smoky Mountains. Ecol. Monogr. 26, 1–80.