Austral ECOLOGY A Journal of ecology in the Southern Hemisphere

Austral Ecology (2021) **46**, 911–926

Analysis of similarities (ANOSIM) for 2-way layouts using a generalised ANOSIM statistic, with comparative notes on Permutational Multivariate Analysis of Variance (PERMANOVA)

PAUL J. SOMERFIELD,*[1] K. ROBERT CLARKE[1,2] AND RAY N. GORLEY[2]
[1]*Plymouth Marine Laboratory, Prospect Place, Plymouth, PL1 3DH, UK (Email: pjso@pml.ac.uk); and* [2]*Primer-E Ltd, c/o Plymouth Marine Laboratory, Plymouth, UK*

Abstract ANOSIM (Analysis of Similarities) is a robust non-parametric hypothesis testing framework for differences in resemblances among groups of samples. The generalised ANOSIM statistic R^O is defined as the slope of the linear regression of ranked resemblances from observations against ranked distances in a model describing the unordered or ordered distances among samples under an alternative to the null hypothesis. In the absence of ordering, this becomes the standard ANOSIM R statistic. The construction of 2-way tests using the generalised statistic in various nested and crossed designs, with and without ordered factors, and with or without replication, is described. Examples are given of 2-way tests with ordered factors in marine ecological studies: 1. phytal meiofaunal communities in species of macroalgae with increasing physical complexity, among islands in the Isles of Scilly; 2. coral community composition across intertidal flats in Thailand, sampled in different years; 3. macrofauna inhabiting kelp holdfasts from different places in response to an oil spill; 4. experimental effects of salinity stress and food restriction on nematode communities. ANOSIM is fully non-parametric and thus cannot, for two-way crossed designs, decompose factors into (metric-based) main effects and interactions; this requires at least semi-parametric modelling, such as provided by PERMANOVA. The two approaches therefore test very different hypotheses: ANOSIM gives a robust, comparable and globally interpretable measure of magnitude of overall community change associated with each factor, having excised any possible effect from the factor(s) it is crossed with, irrespective of whether the factors interact or not. PERMANOVA cannot do this because the presence of interactions will compromise (sometimes totally) any overall measures of the main effects of each factor. Conversely, PERMANOVA can test for interactions involving directional (but non-magnitudinal) community change, which are entirely invisible to ANOSIM. The two methods are therefore seen as complementary, rather than as alternatives.

Key words: distance matrix, hypothesis test, interactions, multivariate analysis, non-parametric, ordered factors.

INTRODUCTION

Data with numerous variables (such as abundances or biomasses of different species) in samples can often be difficult to analyse with traditional statistical approaches. Field *et al.* (1982) described a robust non-parametric multivariate strategy for the analysis of such ecological data which was expanded and clarified by Clarke (1993) and continues to evolve (see Clarke *et al.* 2014, Somerfield *et al.* 2021 and references therein).

A key formal hypothesis test within the framework is ANOSIM (Analysis of Similarities), a special form of Mantel (1967) test originally described for one-way layouts by Clarke and Green (1988), which performs a permutation test of the null hypothesis of no differences among *a priori* defined groups of samples, based on the ranks of the sample dissimilarity matrix. Clarke (1988, 1993) described how ANOSIM can be extended to two-way nested and crossed layouts with replication, and Clarke and Warwick (1994) consider the special case of crossed layouts without replication. Somerfield *et al.* (2021) extend one-way ANOSIM to cater for testing serially ordered factors, such as in space, time or experimental treatment levels.

The current paper demonstrates how ordered and unordered factors, with or without replication, may be tested and compared in a generalised ANOSIM framework for 1- and 2-factor designs, with the approach exemplified by four studies in marine ecology. Consideration of crossed designs for 2- and higher-way layouts raises the issue of interactions, where the effects of a factor vary across the levels of the factor(s) with which it is crossed. In classical univariate ANOVA, decomposition of the influence of two factors into their main

*Corresponding author.
Accepted for publication April 2021.

doi:10.1111/aec.13059

effects and interaction can be a strong function of the measurement scale, with some interactions defined in this way appearing or disappearing under simple transformation of the units of measurement (e.g. log or square root, which cannot change the rank order). Fully non-parametric techniques such as ANOSIM could not therefore be expected to provide a parallel decomposition into tests for main and interaction effects. There are, however, other forms of interaction which *are* testable within a non-parametric framework (Clarke *et al.* 2006) and, as will be seen later, interactions involving differing magnitudes of community change in a factor, over different levels of a factor with which it is crossed, can be indirectly measured by ANOSIM statistics.

Taking the measurement scale of the dissimilarities seriously, a Principal Co-ordinate Analysis (PCO) of the community data is able to construct a (complex) metric space within which linear models can be fitted, as in the semi-parametric Permutational Multivariate Analysis of Variance method (PERMANOVA, Anderson 2001a, 2017; Anderson *et al.* 2008). PERMANOVA and ANOSIM approaches are compatible, since they can be based on the same dissimilarity matrix, and are able to supplement each other in the interpretation of community change. Using simulated and real data sets, the later discussion illustrates the similarities and differences between ANOSIM and PERMANOVA in the context of 2-way crossed designs, highlighting the strengths and different emphases of the two methods.

MATERIAL AND METHODS

Background

One-way ANOSIM (Clarke & Green 1988) tests the null hypothesis of no difference between the factor levels, using a statistic which is the scaled difference between the average ranks of sample dissimilarities between and within groups: $R = (\bar{r}_B - \bar{r}_W)/c$. When there are no group differences R is centred at zero, and the scaling constant $c = n(n\text{-}1)/4$ (where n is the total number of samples) is chosen so that $R \leq 1$, a limit it only attains if all sample dissimilarities between groups are larger than any within groups. Testing recalculates R under all (or a random subset of) permutations of the sample labels to the factor levels.

Two-way ANOSIM tests (Clarke ,1993, 1998) divide into 2-way crossed and 2-way nested cases. An example of a 2-way crossed layout might be samples collected from the same range of tidal heights on a set of different shores (see ecological examples in Table 1). In a 2-way crossed test (denoted A × B), with replication, the effect of factor B on factor A may be entirely removed by calculating the ANOSIM statistic R_A within each level of factor B and then averaging R_A across all levels of B to give $\overline{R}_A$. The significance of the observed $\overline{R}_A$ is then tested by permuting the sample labels and recalculating $\overline{R}_A$ while constraining permutations within levels of B. As the design is crossed, two complementary hypotheses may be tested, namely whether there is an effect of factor A having removed any effect of factor B, or whether there is an effect of factor B having removed any effect of factor A.

When there is no replication, the classical R statistic is undefined as there are no within-group dissimilarities. Instead, the effect of factor A is determined by assessing whether there is evidence of a common pattern among the different levels of A when examined for each of the levels of B (Clarke & Warwick 1994). For every pair of levels of B, a correlation (ρ) is computed between the corresponding elements of the two rank dissimilarity matrices for the A samples. These matrix correlations for all pairs of levels of B are then averaged to give the test statistic ρ_{av}. If there are differences between the levels of A, consistently enough across the levels of B to generate some commonality of patterns, then $\rho_{av} > 0$. Its significant departure from zero can be tested by permuting the A labels separately for each B level, as before. Note that there must be at least three levels in A for this to be a possible test, and then only if there are several levels in B. With only two B levels, A will need five or more levels to provide a rich enough structure for any meaningful computation of ρ. Designing the sampling to have genuine replication is always to be preferred to this matching procedure based on ρ_{av}, since replication will permit follow-up pairwise tests and better interpretation of magnitude of effects in the presence of interaction.

For a 2-way nested analysis with B nested in A (denoted B(A)), such as a number of samples collected from each of a set of polluted beaches and a different set of unpolluted beaches, so the factor 'beaches' is nested in 'pollution' (see ecological examples in Table 1), two null hypotheses are tested sequentially (Clarke 1988, 1993):

H_{0B}: there are no differences among levels of factor B within each level of A;
H_{0A}: there are no differences between levels of factor A.

H_{0B} is examined by calculating $\overline{R}_B$ among levels of B within levels of A. Constrained permutations within levels of A are then used to recalculate possible values of $\overline{R}_B$ representative of the null hypothesis of no difference among levels of B, given that there may be differences among levels of A. The approach to testing H_{0A}, which will usually be the more interesting of the two hypotheses, might depend on the outcome of testing H_{0B}. If H_{0B} is rejected, then the individual samples within levels of B cannot be used as replicates for testing differences among levels of A, and samples need to be pooled in some way to give a single replicate for each level of B within the different levels of A. Consistent with the overall strategy, that tests should only be dependent on the rank similarities in the original resemblance matrix, this may be done by averaging over the appropriate ranks to obtain a reduced matrix, which is then re-ranked. H_{0A} may now be tested by conducting a 1-way ANOSIM test on the reduced re-ranked matrix. If H_{0B} is not rejected, the conservative option is to proceed as if it had been, as failing to demonstrate an effect is not the same as demonstrating that an effect does not exist. Alternatively, all samples within levels of B could be considered as replicates in a 1-way

doi:10.1111/aec.13059

Table 1. 1-way and 2-way ANOSIM (global) test statistics, for crossed and nested designs, with unordered or ordered factors, and with or without replication at the lowest level of the design. Also given are the possibility (or not) of pairwise tests, details of the test constructions and examples of contexts in which they might be employed

No.	Type of design	Factor(s)	Factor level ordering	Replicates?	Statistics used	Pairwise test?[†]	Construction of statistic	Examples
1a	1-way	A	Unordered	Yes	R	Yes	A: Standard 1-way ANOSIM statistic[‡]	A: sites, with replicates in each
1b	1-way	A	Unordered	No	-	-	A: No basis for a test	-
1c	1-way	A	Ordered	Yes	R^{Oc}	Yes	A: ANOSIM form of seriation statistic for ordered categories[§]	A: impact levels, expecting monotonic response
1d	1-way	A	Ordered	No	R^{Os}	No	A: ANOSIM form of simple seriation statistic (no replicates)[§]	A: inter-annual trend or positions along a transect
2a	2-way crossed	AxB	A unordered	Yes	A: $\overline{R}$	Yes	A: Average of 1-way R for testing A across separate levels of B	A: shores, B: treatment types (several applications), or
			B unordered		B: $\overline{R}$	Yes	B: Average of 1-way R for testing B across separate levels of A	A: locations, B: habitats (sites as replicates)
2b	2-way crossed	AxB	A unordered	No	A: ρ_{av}	No	A: Average of ρ among resemblance matrices (of A) across levels of B[¶]	As 2a but each treatment only once on each shore, or A: sites,
			B unordered		B: ρ_{av}	No	B: Average of ρ among resemblance matrices (of B) across levels of A[¶]	B: times, each site visited once at each time
2c	2-way crossed	AxB	A unordered	Yes	A: $\overline{R}$	Yes	A: As test 2a	A: shores, B: increasing treatment impact levels, or
			B ordered		B: $\overline{R}^{Oc}$	Yes	B: Average of 1-way R^{Oc} for testing B across separate levels of A	A: locations, B: water depths (sites as replicates)
2d	2-way crossed	AxB	A unordered	No	A: ρ_{av}	No	A: As test 2b	A: site, B: tidal height (transect down shore) or
			B ordered		B: $\overline{R}^{Os}$	No	B: Average of 1-way R^{Os} for testing B across separate levels of A	A: patch reefs, B: inter-annual trend
2e	2-way crossed	AxB	A ordered	Yes	A: $\overline{R}^{Oc}$	Yes	A: Average of R^{Oc} for testing A across B levels (i.e. 2c, switching A and B)	A: shores on latitudinal gradient,
			B ordered		B: $\overline{R}^{Oc}$	Yes	B: As 2c	B: coarseness of sediment classes, replicate sites in each combination
2f	2-way crossed	AxB	A ordered	No	A: $\overline{R}^{Os}$	No	A: Average of R^{Oc} for testing A across B levels (i.e. 2d, switching A and B)	A: transect of sites along shore and
			B ordered		B: $\overline{R}^{Os}$	No	B: As 2d	B: depth transect at each site, sampling (once) the same set of depths
2g	2-way nested (B within A)	B(A)	A unordered	Yes	A: R	Yes	A: As test 1a, but with levels of B as replicates (averaging within those)**	A: protected/not protected areas,
			B unordered		B: $\overline{R}$	No	B: As test 2a, but without pairwise tests**	B: sites within each type (replicates are trawls within each site)
2h	2-way nested	B(A)	A unordered	No	A: R	Yes	A: As test 1a, but this time the sole levels of B are the only replicates	A: location,
			B unordered		B: -	-	B: No basis for a test	B: site (e.g. time-averaged to give one sample for each site)
2i	2-way nested	B(A)	A ordered	Yes	A: R^{Oc}	Yes	A: As test 1c, but with levels of B as replicates (averaging within those)**	A: depth bands,
			B unordered		B: $\overline{R}$	No	B: As test 2g	B: random sites in each depth band, replicate grab samples at each site

doi:10.1111/aec.13059

Table 1. *Continued*

No.	Type of design	Factor(s)	Factor level ordering	Replicates?	Statistics used	Pairwise test?[†]	Construction of statistic	Examples
2j	2-way nested	B(A)	A ordered	No	A: R^{Oc}	Yes	A: As test 1c, but this time the sole levels of B are the only replicates	A: distance from outfall,
			B unordered		B: -	-	B: No basis for a test	B: random sites at each distance, and 'pseudo-replicates' (e.g. multicorer) pooled
2k	2-way nested	B(A)	A unordered	Yes	A: R	Yes	A: As test 2g (ordered levels of B assumed representative as replicates)[**]	A: dry/wet season,
			B ordered		B: $\overline{R}^{Oc}$	No	B: As test 2c, but without pairwise tests [**]	B: months (replicates as random days in month)
2l	2-way nested	B(A)	A unordered	No	A: R	Yes	A: As test 2h (ordered levels of B assumed representative as replicates)[**]	A: site,
			B ordered		B: $\overline{R}^{Os}$	No	B: As test 2d	B: points along transect (one transect at each site, randomly oriented and located)
2m	2-way nested	B(A)	A ordered	Yes	A: R^{Oc}	Yes	A: As test 2i (ordered levels of B assumed representative as replicates)	A: region, latitudinally arranged,
			B ordered		B: $\overline{R}^{Oc}$	No	B: As test 2k	B: transect of sites in each region (all at same depth), replicates within
2n	2-way nested	B(A)	A ordered	No	A: R^{Oc}	Yes	A: As test 2j (ordered levels of B assumed representative as replicates)	A: seamounts in different depth classes,
			B ordered		B: $\overline{R}^{Os}$	No	B: As test 2l	B: distance along single random transect on each seamount

[†]All pairwise tests are unordered, by definition.

[‡]$R = (\overline{r}_B - \overline{r}_W)/(M/2)$, equivalently the slope of a linear regression of ranks of the biotic resemblances against ranks from a (0,1) model for levels of A.

[§]R^{Oc} is the slope from a linear regression of ranks of biotic resemblances against ranks from a 'seriation with replication' model matrix and R^{Os} against a simple seriation model without replication; they are the (asymmetric) ANOSIM R forms of the (symmetric) RELATE Spearman ρ statistic. The distinction between ordered categories (R^{Oc}) and simple seriation (R^{Os}) is not crucial for calculation purposes (thus R^{O}).

[¶]Matrix correlation (Spearman rank ρ) calculated between all pairs of biotic resemblance matrices (for levels of A) within levels of B, and then ρ averaged over the separate B levels to give ρ_{av} for A (vice-versa for B).

[**]Ranked resemblances are averaged within levels of B(A), and for all pairs across levels of B(A); the resulting averaged matrix is re-ranked and input to 1-way ANOSIM for levels of A, using B levels as replicates. The same is done for each of the pairwise tests, first selecting only resemblances for the requisite pair of A levels, then ranking, averaging and re-ranking before inputting the two levels to 1-way ANOSIM.

[††]The global test is the same as the crossed case but here the levels of B, even if similarly denoted (by 1, 2, ... say) have nothing in common across the levels of A, so a pairwise test of B1 v B2 (say) is meaningless.

[‡‡]A nested factor might typically be a randomly located site (B) in a region (A). Ordered sites might come from transects of sites across each region (randomly directed so transect points are nested not crossed with region). If representative of the region's extent, transect sites could still be considered suitable replicates for a test of region, the 'randomness' coming from the stochastic nature of the environment being sampled.

test for differences among levels of A. The choice here is parallel to that of whether 'to pool' or 'not to pool' in forming the residual for the analogous univariate 2-way nested ANOVA. The second option is likely to have greater power, resulting from the much larger number of possible permutations, but runs a real risk of being invalid (Clarke, 1993).

Somerfield *et al.* (2002) showed that if groups of replicates are serially ordered then a non-parametric Mantel test (RELATE) of the matrix correlation ρ between the ranked dissimilarities from observations and a simple model matrix characterising the ranked distances between groups under the (ordered) alternative hypothesis will have more statistical power to detect differences than the

doi:10.1111/aec.13059

ANOSIM test for (unordered) differences among groups. Somerfield *et al*. (2021) discuss ANOSIM for 1-way (single factor) designs, showing how the traditional ANOSIM statistic R may be generalised. They redefined it as the slope of the linear regression of ranked resemblances from observations against ranked distances in a model matrix. When the model represents an unordered factor, the standard R statistic is obtained, and for a matrix characterising serial ordering of the factor levels, the regression slope is termed the generalised ANOSIM statistic, R^O (superscript 'O' for ordered). Two variants are described, namely R^{Oc} ('c' for categories), where there are replicates within groups, and R^{Os} ('s' for singles) where there are only single samples (no replicates) within groups (Somerfield *et al*. 2021).

2-way ANOSIM designs with ordered factors

The principle of 2-way ANOSIM tests, and their permutation procedures (Clarke 1993), remains largely unchanged when A or B (or both factors) are ordered. Previously, the test for B under the nested B(A) model averaged the 1-way R_B statistic for each level of A, denoted $\overline{R}_B$, and the same form of averaged statistic was used for testing B under the crossed A × B model with replicates. Without replicates, the crossed test used the special (and less powerful) construction in which the test statistic was the pairwise averaged matrix correlation, ρ_{av}. There was no test for B in the nested model, in the absence of replicates for B. If B is now ordered, R is replaced by R_B^{Oc} where there are replicates (becoming $\overline{R}_B^{Os}$ when averaged across the levels of A), or by R^{Os} where there are not (becoming $\overline{R}^{Os}$) and there is no longer any necessity to invoke the special form of test based on ρ_{av} when the B factor is ordered. The same substitutions then happen for the test of A in the crossed case; if it too is ordered then $\overline{R}_A$ and ρ_{av} are replaced by $\overline{R}_A^{Oc}$ and $\overline{R}_A^{Os}$. If A is not ordered, any ordering in B does not change the way the tests for A are carried out, for example for A × B, the A test is still constructed by calculating the appropriate 1-way statistic for A, separately for each level of B, and then averaging those statistics.

Table 1 lists all the many possible combinations of 1- and 2-way design, factor ordering (or not) and presence (or absence) of replicates, giving the test statistic and its method of construction, listing whether pairwise tests are feasible and desirable or not, and then giving examples of marine studies in which the factors would have the right structure for such a test. Four examples follow.

Example data sets

Isles of Scilly phytal meiofauna

Gee and Warwick (1994a,b) collected macroalgae and the fauna within them from eight sites on three islands (St Agnes, St Marys and St Martins) in the Isles of Scilly, UK. The meiofauna were extracted and identified to 99 taxa, predominantly harpacticoid copepods. Here, a subset of samples from three species of algae is analysed. For animals in the meiofaunal size range, these algae represent a gradient of increasing habitat complexity in the order *Chondrus crispus* < *Lomentaria articulata* < *Cladophora rupestris*. This is a 2-factor crossed design, with one ordered factor (Fractal) representing the ordering of the species of algae by their measured fractal dimension, crossed with a second factor Island. As quantitative control of sample size was difficult (equivalent volumes of algae of different complexity provide different amounts of living space for the fauna), abundances were standardised prior to a fourth-root transformation.

Ko Phuket corals

In many (though not all) years since 1983, coral assemblages have been sampled along three permanent transects across intertidal flats on the south-east tip of the island of Kho Phuket, Thailand, in the Andaman Sea (Brown *et al*. 2019). The transect considered here, Site A, was sampled on each occasion by twelve '10m plotless line samples', perpendicular to the main transect and spaced at about 10 m. Percentage cover of each line sample by each of 53 coral taxa was recorded. The data therefore represent a sequence of 12 un-replicated samples from the upper to the lower shore within each sampled year. Although various events have influenced coral communities in different years, data from 7 years collected over the period 1988 to 1997, representing 'normal' conditions, are used here. Inter-sample Bray–Curtis resemblances were calculated using square-root transformed percentage cover data. This is a two-factor crossed un-replicated design, with one spatial factor (Position on the shore) and one temporal factor (Year), with the spatial factor clearly ordered and the temporal factor capable of being analysed either as unordered or ordered, depending on whether the test is for non-specific inter-annual variation or for a trend in time.

Sea Empress holdfast fauna

On the 15th February 1996, the 147,000 t oil tanker Sea Empress ran aground whilst trying to enter Milford Haven, south-west Wales, UK, releasing 72,000 t of light-grade crude oil over the ensuing six days. The resulting slick contaminated 500 km^2 of sea surface and over 100 km of coastline (Law & Kelly 2004). A field programme was initiated to monitor the effects of the oil and subsequent recovery using the fauna inhabiting the holdfasts of kelp on rocky shores (Somerfield & Warwick 1999). Four sites were selected within each of four regions. Region A was in the near-vicinity of the wreck, region B was located within the surface extent of the spilled oil, whilst regions C and D were control areas immediately outside the oil-affected area on both the Welsh (C, to the north) and North Devon (D, to the south) coasts. Initial sampling took place over low spring tides in March 1996, one month after the accident. At each site, five samples of kelp holdfasts were collected as far down the shore as possible around the time of low water (± 1 h). Samples were preserved in formalin. They were subsequently broken up in the laboratory and the fauna was extracted using a 0.5 mm mesh, sorted and identified (Somerfield & Warwick 1999). As sample volumes varied, the abundances were standardised and then square-root transformed prior to calculating Bray–Curtis similarities.

doi:10.1111/aec.13059

This is a 2-way nested design with sites nested in regions. Regions may be ordered into zones (A >B>C=D) reflecting the potential influence of the oil (case 2i in Table 1), or analysed as ordered (or unordered, case 2g in Table 1) regions reflecting a spatial gradient from north to south (C>B>A>D). There are therefore three competing alternatives to the null hypothesis H_0: A=B=C=D for the top level test of regions, namely H_1: A,B,C,D differ (in ways unspecified), H_2: A>B>C=D and H_3: C>B>A>D, each generating a different ANOSIM statistic and associated test. The values of the resulting R or R^{Oc} statistics are directly comparable and, if the null hypothesis can be rejected, the largest R can be considered to give the greatest credence to its associated alternative hypothesis.

Microcosm nematodes

Austen (1989) examined the effects on a free-living estuarine nematode community (45 species observed) after 48 weeks of exposure in microcosms, subject to Normal or Limited feeding and control (C, 25ppt)/medium (M, 15ppt))/high (H, 5ppt) levels of stress from reduced salinity, with a design of 8 replicates in each of the 2 × 3 experimental conditions. Abundances were fourth-root transformed prior to calculating Bray–Curtis similarities. This is a further example of a 2-factor crossed design with an ordered factor (salinity stress) and provides a convenient framework for discussing the various strengths and weaknesses of the non-parametric ANOSIM methods of this paper with those of the semi-parametric partitioning models of PERMANOVA (Anderson 2001a, 2017; Anderson *et al.* 2008).

Data analyses

All the analyses were undertaken with PRIMER v7 (Clarke & Gorley, 2015), with the PERMANOVA+ add-on (Anderson *et al.* 2008). Testing utilised the ANOSIM routine with 9999 randomly selected permutations, where it was not feasible to evaluate all the potentially distinct ones (see Appendix S1). Tests using PERMANOVA also used 9999 randomly selected permutations under a reduced model (Freedman & Lane 1983). Inter-sample Bray–Curtis resemblances calculated from pre-treated data were ordinated using non-metric or metric multidimensional scaling, nMDS or mMDS (e.g. Kruskal & Wish 1978; Clarke *et al.* 2014).

RESULTS

Isle of Scilly phytal meiofauna

Ordination of the samples (Fig. 1) clearly suggests that meiofaunal communities change along a gradient of increasing structural complexity of the macroalgae. The appropriate 2-way test (case 2c of Table 1) calculating R^{Oc} for differences among ordered algal species within each level of the island factor, and then averaging them, gives $\overline{R}^{Oc}$= 0.76 for ordered differences between algae, removing the effects of differences among islands. This value is greater than any of 9999 random selections from the possible 35,083,125 constrained permutations (see Appendix S1), so $p < 0.01\%$.

Pairwise tests (for which, of course, assumptions of ordered or unordered levels can make no difference) give $\overline{R} = 0.56$ ($p = 1.9\%$) for the difference between samples from *Chondrus* and *Lomentaria*, 0.84 ($p = 0.3\%$) for the difference between *Lomentaria* and *Cladophora*, and the highest value of 0.97 ($p = 0.3\%$) for those species that the model places furthest apart, namely *Chondrus* and *Cladophora*. In comparison with the ordered global statistic $\overline{R}^{Oc}$ = 0.76, the unordered global test gives a lower value of $\overline{R} = 0.67$. Although there are fewer possible permutations for the unordered test (1,299,375; see Appendix S1) this observed value is again greater than that from 9999 random selections, so $p < 0.01\%$.

An unordered test for differences among islands, calculating R separately for each algal species and averaging over these, gives $\overline{R} = 0.15$. Appendix S1 shows that there are 9,261,000 possible permutations for this design (which is unbalanced within the strata of the differing algae), and a randomly drawn set of 9999 permutations gives a non-significant result ($p = 15.3\%$) An ordered alternative in which islands are thought of as ordered from North to South (case 2e in Table 1) gives a higher value of $\overline{R}^{Oc} = 0.19$, but this is still not significant ($p = 10.4\%$).

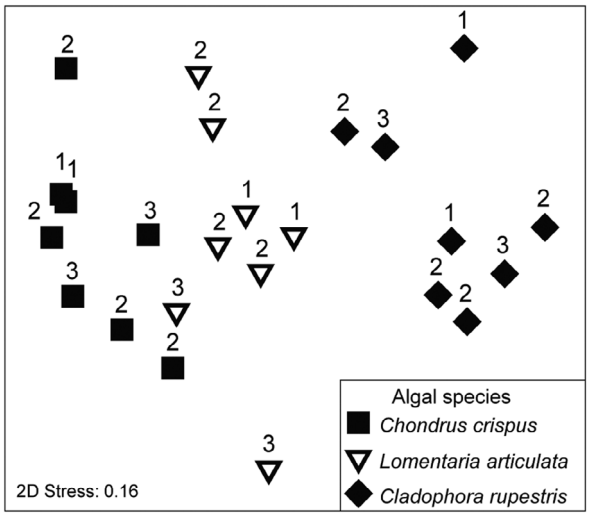

Fig. 1. Non-metric MDS of Bray–Curtis similarities derived from fourth-root transformed standardised abundances of meiofaunal taxa in three species of macroalgae with increasing complexity from three islands in the Isles of Scilly, UK. Labels indicate islands as 1, St Martins; 2, St Marys; 3, St Agnes.

doi:10.1111/aec.13059

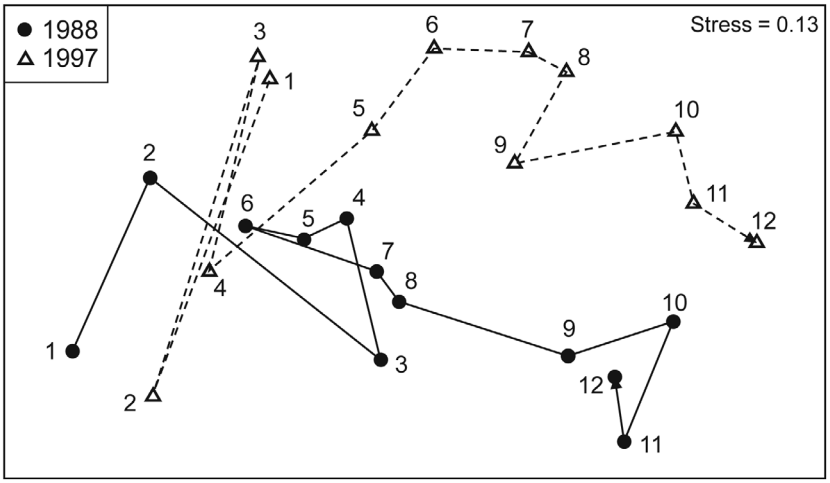

Fig. 2. Ordination by nMDS of samples based on Bray–Cutis similarities among square-root transformed percent cover of 53 coral taxa in just two (1988, 1997) of the full set of 7 years analysed, at 12 positions along an inshore-offshore transect on the shore of Ko Phuket.

Ko Phuket corals

For clarity, Fig. 2 shows an nMDS ordination of the 12 positions down the shore (inshore to offshore, 1 to 12), in just the first and last years of the selected time period. The other five years have similarly clear spatial trends, so it is not surprising that the ordered ANOSIM test for Position (the B factor in case 2d of Table 1), which uses the un-replicated $\overline{R}^{Os}$ statistic, an average of the separate R^{Os} statistics over 7 years, returns the high value of 0.68 ($p < 0.01\%$). In spite of the absence of replication, separate analyses of the Position factor for each year are now possible, using a 1-way ordered ANOSIM without replication (case 1d of Table 1). For example, the spatial trends seen in Fig. 2 for 1988 and 1997 have $R^{Os} = 0.65$ and 0.73 (both $p < 0.01\%$), respectively.

With no replication and no ordering, the general test for the Year factor (7 levels between 1988 and 1997, factor A in case 2d of Table 1) looks for any evidence of common patterns of annual change for the separate shore positions. The submatrices representing the ranked differences among years at each shore position are pairwise rank-correlated with each other and these matrix correlations (ρ) averaged over all pairs of positions. The resulting average, $\rho_{av} = 0.02$ ($p = 28\%$), indicates that there is no commonality of inter-annual patterns across the transect sites, thus no detectable year effect. A more directed test of an inter-annual *trend* for the seven years (case 2f in Table 1), based on the R^{Os} statistic computed through the years separately for each transect position, and then averaged ($\overline{R}^{Os}$) over transect positions, also gives a low and non-significant value of 0.08 ($p = 11\%$), indicating no year effect.

Sea Empress holdfast fauna

Holdfasts from unordered sites of a similar type (i.e. within each of the zones influenced by differing exposure to oil) have differing community composition ($\overline{R} = 0.66$, $p < 0.01\%$). Averaging the holdfasts within sites (Fig. 3), the ordered test for differences among zones influenced by oil (A>B>C=D, the alternative hypothesis H_2), a 1-way ANOSIM test using the ordered category statistic, gives $R^{Oc} = 0.28$, $p = 0.9\%$. The corresponding unordered (3-level) test gives $R = 0.195$, a lower value (for which p is only 5.6%) demonstrating the additional power of an ordered test in a situation where the data are genuinely ordered. There is therefore some evidence here that, relative to spatially more widely defined control regions, the holdfast communities differ between zones subject to differing levels of potential impact.

An alternative test is for ordered regions (north to south, C>B>A>D, the alternative hypothesis H_3). This gives an ordered test with $R^{Oc} = 0.27$, $p = 0.8\%$, but the corresponding unordered (4-level) test now gives $R = 0.47$, $p = 0.03\%$. Values of the various R statistics are comparable, and their values from the different tests are measures of the relative strength of the match between the data and each of the competing alternative hypotheses. The overall conclusion, therefore, is that the data show a pattern which is most consistent with the alternative hypothesis H_1, namely that there are unordered differences between the four regions. There is a serial change in holdfast community structure related to the potential effects of oil coming ashore, but this should be viewed in the context of large differences that were

doi:10.1111/aec.13059

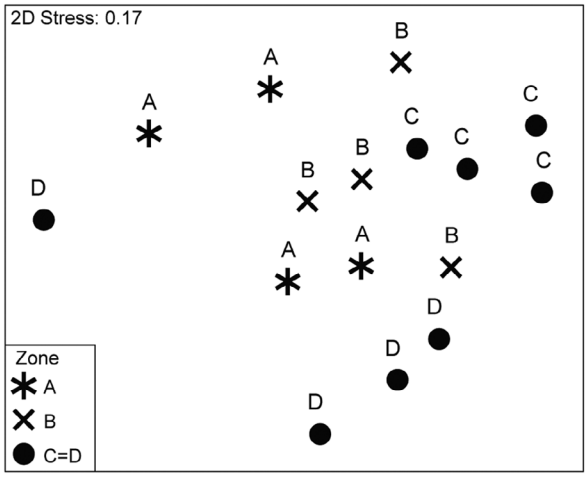

Fig. 3. Ordination by nMDS of Bray–Curtis similarities calculated from square-root transformed standardised abundances of 146 macrofaunal taxa inhabiting kelp holdfasts, averaged within each of 16 sites. Ordered zones are structured to reflect the potential impact of oil from the Sea Empress tanker accident, with A close to the wreck, B within the area affected by floating oil, and C = D including control regions to the north (C) and south (D) of the affected area.

also observed between the two control regions. It might be concluded, therefore, that although there is some evidence for the effects of the oil on holdfast communities, this is within the range of natural geographical variation.

Microcosm nematodes

The final data set is another 2-way crossed design, well-replicated with ordering in one of the factors and effects evident in both, as seen in the non-metric MDS of nematode communities (Fig. 4a), after 48 weeks of a microcosm experiment. A 2-way crossed ANOSIM test of the ordered factor representing salinity stress (C: control 25ppt, M: medium 15ppt, and H: high stress 5ppt) confirms the strong group differences $\overline{R}^{Oc} = 0.749$, an average of consistent values $R^{Oc} = 0.765$ and 0.733 from the k (= 2) strata of different feeding levels (respectively, Normal: closed symbols; Food limited: open symbols in Fig. 4). From Appendix S1 equations (A3) and (A4), with $g = 3$ groups and $n = 8$ replicates per group, the $\log_{10}$ of the total number of potentially distinct permutations for the $\overline{R}^{Oc}$ test would be a very large $2\times\log_{10}(T)$ (≈ 19.4); unsurprisingly, $p < 0.01\%$ from a random sample of 9999. (In fact, one of the 8 replicates from the M salinity group under Normal feeding was lost, but the change to T is minor of course, and lack of exact balance in replication is not

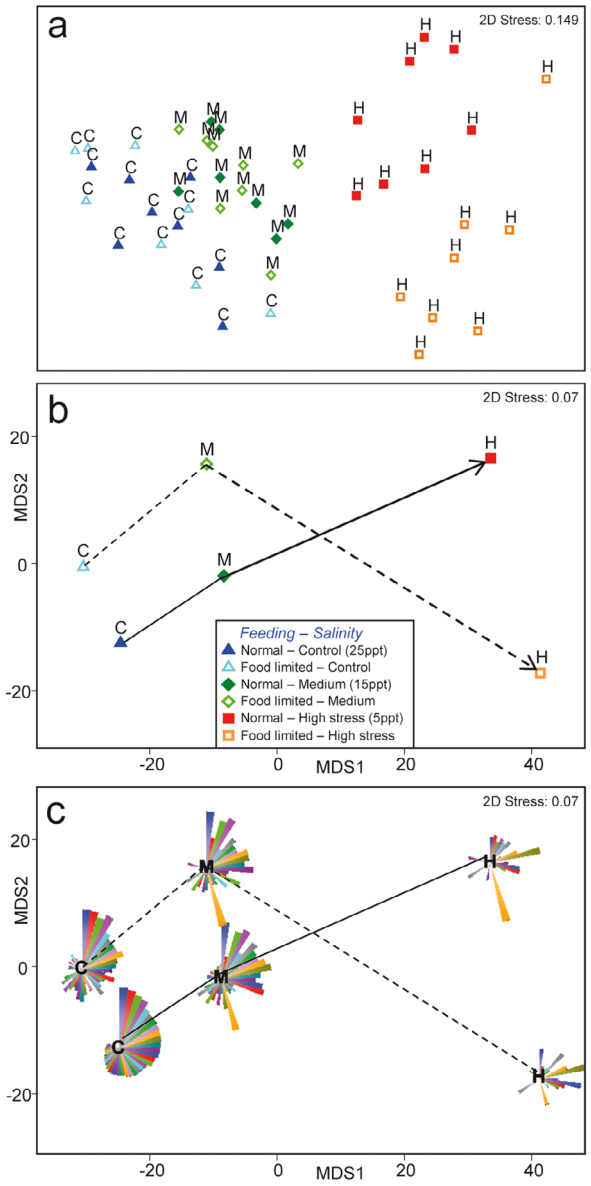

Fig. 4. (a) nMDS ordination of Bray–Curtis similarities from fourth-root transformed abundances of 45 estuarine meiofaunal taxa after 48 weeks in a microcosm experimental design with two crossed factors of Feeding regime (closed symbols: Normal, open symbols: Feeding limited) and Salinity stress (Control, 25ppt; Medium, 15ppt; High stress, 5ppt), with 8 replicates in each combination (one replicate lost for N-M condition); (b) 'means plot', a metric (mMDS) ordination of averages over replicates of transformed abundances, with split trajectories indicating increasing salinity stress for the two feeding regimes; (c) 'Nautilus plot', a segmented bubble plot showing averaged transformed abundances of all 45 species superimposed on the means plot, the segments using a common size scale and the order of species round the circle given by decreasing abundance for the control condition for both factors (N-C).

doi:10.1111/aec.13059

an issue for the ANOSIM, or PERMANOVA, methods discussed in this paper.)

Similarly significant is $\overline{R} = 0.686$ ($p < 0.01\%$) for the unordered ANOSIM test of salinity effects (removing any feeding effects), which compares with the ordered test statistic $\overline{R}^{Oc} = 0.749$. The larger value for the latter thus indicates that a serial change in communities over the three salinity levels reflects the data better than unordered change, and this is clearly seen in the means plot (Fig. 4b), of averages for the transformed replicate data in each group, input to Bray–Curtis similarity and ordinated by metric MDS (mMDS), for example Clarke *et al.* (2014). The latter is preferable to non-metric MDS (nMDS) when there are few points to display, and consequently few rank dissimilarities to constrain the placements in nMDS. With only 6 points, the higher stress of an mMDS is acceptably low here (at 0.07).

Turning to the converse test for effects of the feeding regime (ordered or unordered is immaterial, since there are only two levels), removing any effects of salinity change, there is a clear demonstration of the effect of food limitation since $\overline{R} = 0.229$ is strongly significant at $p = 0.08\%$, on 9999 random permutations (from the possible 2.8×10^{11}). Interestingly, however, this is an average of $R = -0.011$ ($p = 45.0\%$), $R = 0.239$ ($p = 3.0\%$) and $R = 0.460$ ($p = 0.10\%$) for the individual 1-way tests of feeding regimes within each of the salinity strata of Control, Medium and High stress, respectively. This is a clear indication of one type of interaction between the two experimental factors – feeding limitation has no effect at control salinity levels, but a large one at high salinity reduction. The issue of types of interaction and their consequences for the outcomes of ANOSIM and PERMANOVA tests is further elaborated, for this design, in the Discussion.

DISCUSSION

Within the unified ANOSIM framework, tests for 1-way or 2-way designs, having unordered or ordered factors, either with or without replication (where a test is possible), may be constructed. The tests require no distributional assumptions and are non-parametric.

With any statistical test, the investigator should understand what the test is doing, what the specific hypothesis being tested is, and what a rejection of the null hypothesis actually implies. This discussion attempts to unpack this statement with a few simple illustrations of a 2-way crossed ANOSIM test run in parallel with the semi-parametric PERMANOVA method (Anderson 2001a, 2017; Anderson *et al.* 2008), the comparison being perfectly valid when starting from the same chosen dissimilarity matrix. In brief, the dependence of ANOSIM only on the rank order of the dissimilarities confers robustness and, in addition to its use as a test statistic, ANOSIM R has meaning as a universal measure of effect size (much as a correlation coefficient does). ANOSIM cannot, however, deliver the richer sophistication of PERMANOVA modelling, for example the direct partitioning into main effects and interactions for multi-way designs.

The null hypothesis for ANOSIM is that between- and within-group dissimilarities, for the factor under consideration, are indistinguishable (characterised as H_0: $R = 0$). In the case of 2-way crossed ANOSIM, this applies within each level of the second factor; dissimilarities crossing those levels have no constraints under H_0, since they are never considered. The one-sided alternative hypothesis for ANOSIM is quite widely specified (characterised as H_1: $R > 0$), making ANOSIM a portmanteau test, namely having the property of being at least moderately powerful against a range of departures from the null hypothesis, including an ability to reject H_0 in response to heterogeneous 'dispersion' (the spread of within-group dissimilarities) across the groups. We leave aside for this discussion, however, the different ways in which ANOSIM and PERMANOVA respond to such heterogeneity since this has been extensively covered in a simulation study by Anderson and Walsh (2013), and concentrate on some key differences for interpretation using examples in which within-group dispersions are relatively homogeneous.

In PERMANOVA, the response is parameterised as an additive combination of main effects, interaction term(s) and an additive 'error', modelled in a high-dimensional (and complex, in mathematical language) 'dissimilarity space' created from the chosen resemblance measure. No assumptions are made, however, for the probability distribution of the errors in that space – the method is 'distribution-free'. To allow this construction, the measurement scale of the dissimilarities (values, not ranks) must be taken seriously. In contrast, ANOSIM does not create such a space and is fully non-parametric (also distribution-free). Its tests do not differ under monotonic transformations of the resemblance matrix; only the rank order relationships among dissimilarity values matter. For example, a PERMANOVA table will differ if Euclidean or Manhattan distances are used, but ANOSIM must give the same outcome for the two cases (Euclidean being just the square of Manhattan distance). The maxim, familiar from univariate statistics, is that the more structured the modelling the deeper the inference (PERMANOVA), the fewer the assumptions the more robust the conclusions (ANOSIM).

Another important distinction concerns the null hypotheses being tested, which in PERMANOVA make statements about the multivariate group

doi:10.1111/aec.13059

centroids in the space of the chosen resemblance matrix. For the main effect of the first factor in a 2-way crossed design, the null hypothesis is H_0: there are no differences in centroid location for the levels of that factor, when averaged across all levels of the second factor (and vice-versa). This is very different than the ANOSIM test for the first factor, whose computation – as noted above – involves no dissimilarities which cross the levels of the second factor. This should be an immediate flag that the two test results cannot be interpreted in the same way.

It should be stressed that the usual ANOVA mantra 'beware interpreting the test results for (either of the) main effects if there are interactions' applies only to PERMANOVA. In 2-way ANOSIM, the tests for each factor are straightforwardly interpretable whether or not there are interactions. The 1-way ANOSIM R for the first factor is calculated separately within each of the k strata of the second factor and averaged; if there are no effects whatsoever of the first factor then all k of the R values are close to zero, and their average (denoted $\overline{R}_1$) is close to zero, and ANOSIM will not reject the null hypothesis. Where there *are* effects of the first factor, some or all of these R values become positive; if enough of them do so – or one or two of them do so enough – $\overline{R}_1$ will be significantly greater than zero, and an effect detected. Two-way crossed ANOSIM is thus fundamentally measuring the average *magnitude* of the effect of the first factor, excising *any* information from the second factor. Exactly the same logic applies to the converse test, of effects in the second factor removing any contribution whatsoever from the first factor, using a test statistic denoted $\overline{R}_2$. Furthermore, the size of the test statistics $\overline{R}_1$ and $\overline{R}_2$ can be directly compared with each other, indicating perhaps that one is the more dominant of the two effects overall; this follows from the universality of interpretation of the scaled R statistic, across factors in a dataset, across different pre-treatment or dissimilarity choices, and even across different data sets. Such overall comparisons cannot necessarily be made so straightforwardly between mean squares, (pseudo-)F statistics or components of variation for main effects in (PERM)ANOVA tests in equivalent 2-way (or multi-way) designs, especially in the presence of interactions.

This can be simply illustrated for multivariate data by setting some artificially constructed data into the context of the two-way crossed design presented earlier (Fig. 4), namely estuarine nematode communities in experimental microcosms, subjected to differing food regimes (two levels, Normal and Food Limited) and stress from salinity reduction (two levels, Control 25ppt and High stress 5ppt, the Medium stress level being ignored for the moment). Figure 5a shows an ordination (a scatter plot in effect) of simulated 2-d data, with 8 replicates (as in Fig. 4) from each of the 2×2 treatment levels. The PERMANOVA table in this case, Table 2a, gives a large Salinity main effect (C to H, indicated in the figure by the arrows) but a completely non-significant Feeding main effect, with a (pseudo-)F value close to 1 and a negligible component of variation. Yet there clearly are quite strong differences between the closed (Normal feeding) and open (Food limited) symbols in the MDS plot, for both the C and H levels of the salinity factor, and the ANOSIM test shows this unequivocally. $\overline{R}_1$ for the Feeding test is 0.557, strongly significant at $p < 0.02\%$, and this $\overline{R}_1$ is an average of 1-way ANOSIM values $R = 0.573$ within the C stratum and $R = 0.541$ within the H stratum ($p < 0.1\%$ in both cases). The apparent inconsistency between ANOSIM and PERMANOVA results is, of course, simply explained by the standard mantra at the start of the previous paragraph, because here the PERMANOVA table shows a strong interaction effect, so interpretation of main effects in a (PERM)ANOVA needs to be suitably circumspect, and follow-up analyses of subsets of the data are very necessary (e.g. tests of the feeding levels separately for the two salinity strata are both significant at $p < 0.1\%$). Two-way ANOSIM also correctly and immediately gauges the relative strength of the two factors, evident from the MDS plot, with the Salinity test (C to H) giving an $\overline{R}_2$ of 0.997 ($p < 0.01\%$), much larger than the $\overline{R}_1 = 0.557$ value for the Feeding effect. It is also interesting to note here that the individual R values within each stratum of the second factor, namely $R = 0.997$ c.f. $R = 0.998$ for the Salinity effect (or, as we have seen, $R = 0.573$ c.f. $R = 0.541$ for the Feeding effect) are essentially identical, giving no suggestion of interaction caused by differing *magnitude* of effects of factor 1 for the different levels of factor 2 (or vice-versa). The interaction indicated by PERMANOVA is thus purely one of *direction* of those effects in the high- (here 2-) dimensional space (evident mainly from the ordination rather than the tests), and this interaction masks the importance of one of the factors in the PERMANOVA, at least if the main effect term (or its component of variation) is naïvely seen as measuring overall magnitude of change due to that factor.

A second artificially constructed example provides a contrast with the above, in demonstrating how significant interactions also (and often) arise from varying magnitudes for the effects of one factor in the different strata of the second factor, and this time it is the main effect of Salinity, not Feeding, which disappears. The ordination of Fig. 5b shows an apparently strong Salinity stress effect (C to H, indicated by arrows), similar to that seen in Fig. 5a, but in a more markedly different direction for the Normal (closed symbol) and Food Limited (open symbol) groups. However, the

doi:10.1111/aec.13059

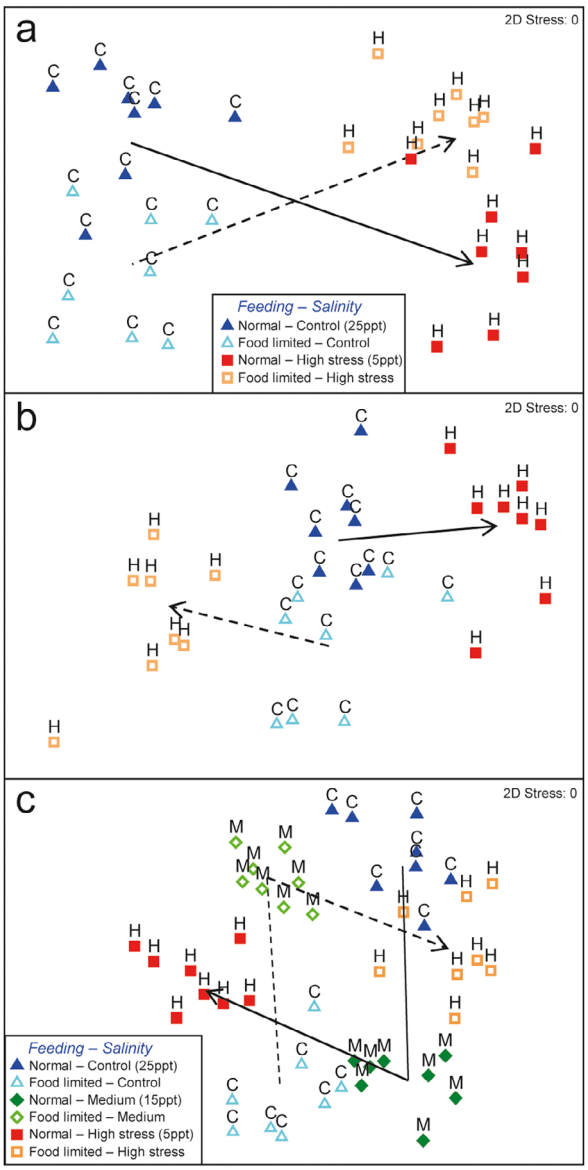

Fig. 5. 2-d scatter plots representing ordinations from three (a-c) illustrative data sets (simulated), set in the context of the nematode microcosm experiment of Fig. 4, namely 8 replicates from each combination of levels for the crossed factors Salinity stress (C: control, M: medium, H: high level stress) and Feeding (closed symbols: normal, open symbols: feeding limitation). (a) and (b) have no Medium level in the simulation, for clarity of example.

PERMANOVA (Table 2b), in contrast with Table 2a, now shows no main effect whatsoever of Salinity, with (pseudo-)$F < 1$, whereas there is now a strong Feeding main effect and, of course, a strong interaction. The Salinity main effect has disappeared in the PERMANOVA but the 2-way ANOSIM shows that the overall magnitude of this effect is as large as for the Feeding factor (actually, numerically slightly larger), since $\overline{R}_2$ for the Salinity effect is 0.818 ($p < 0.01\%$) and for the Feeding effect $\overline{R}_1 = 0.722$ ($p < 0.01\%$). Breaking these averages down, the Salinity effect is again seen – as in the first example – to be fairly consistent in magnitude for the two Feeding regimes (Salinity effect: $R = 0.898$ within Normal and $R = 0.738$ within Food limited strata), but there is now a starker contrast in magnitude of the Feeding effect for the two Salinity regimes ($R = 0.444$ within Control salinity and $R = 1$ within High salinity stress). This example therefore illustrates how a stronger type of interaction in a PERMANOVA test can be suggested also by the ANOSIM statistics, because the interaction reflects differences of both direction and *magnitude* of change in one factor over the levels of the other. But again, in this case, the PERMANOVA main effect terms on their own do not provide a reliable assessment of the relative importance of the two factors. Naturally, in situations where the main effects components of variation greatly dominate that of the interaction, the possibilities for a misleading interpretation of this sort are much reduced. But these two examples make it clear that comparison of the components of variation for the two main effects cannot be as reliable in gauging magnitudes of overall change as that obtained by contrasting the ANOSIM statistics $\overline{R}_1$ and $\overline{R}_2$.

Re-introducing the Medium salinity stress level (15ppt) so that the 2-way crossed design is now 2×3 combinations of levels – again with 8 replicates in each combination, as in the design for the estuarine meiofauna of Fig. 4 – it is even possible to contrive a PERMANOVA test which shows no main effects and component of variation contribution for *either* factor (Table 2c), in the presence of clear differences between all levels of both factors. These differences are very evident from the corresponding ordination, Fig. 5c. In complete contrast, ANOSIM gives near-maximal $\overline{R}_1$ values of 0.998 for the Feeding effect (an average of $R = 0.997$, 1.0 and 0.998 within the C, M and H Salinity stress regimes), and similar $\overline{R}_2 = 0.987$ for the Salinity effect (an average of 0.998 and 0.976 within the two Feeding regimes). Naturally, PERMANOVA shows a highly significant interaction effect, and here the (arrowed) progression from Control to Medium to High levels of the Salinity stress factor is in exactly contrary directions for the two Feeding levels. As noted earlier, what the main effects of a factor in a PERMANOVA table represent are the differences in position of centroids (in the high-d 'dissimilarity space') for its different levels, when averaged across all levels of the other factor(s). It is clear therefore why all the main effects disappear in the PERMANOVA partitioning for this (2-d) example: all the centroids corresponding to the main-effect levels of either Feeding (all filled *vs* all empty symbols) or Salinity (different symbol types, ignoring filling) are coincident with the overall centroid in the middle of the space.

doi:10.1111/aec.13059

Table 2. PERMANOVA tables and associated components of variation of 2-way crossed tests for illustrative (simulated) 2-d data in Fig. 5(a-c), respectively. *P*, from 9999 random permutations, is given as a probability, conventional for ANOVA ($P = 0.0001$ implies significance at the $p = 0.01\%$ level)

(a) PERMANOVA table

Source	df	SS	MS	(Pseudo-)*F*	*P*
Feeding	1	0.34	0.34	1.19	0.3168
Salinity	1	42.861	42.861	150.81	0.0001
Feeding × Salinity	1	7.86	7.86	27.76	0.0001
Residual	28	7.958	0.284		
Total	31	59.019			

Components of variation

Source	Estimate	Sq. Root
Feeding	0.0035	0.0588
Salinity	2.6611	1.6313
Feeding × Salinity	0.947	0.9731
Residual	0.2842	0.5331

(b) PERMANOVA table

Source	df	SS	MS	(Pseudo-)*F*	*P*
Feeding	1	12.272	12.272	61.06	0.0001
Salinity	1	0.055	0.055	0.27	0.7508
Feeding × Salinity	1	8.474	8.474	42.16	0.0001
Residual	28	5.628	0.201		
Total	31	26.428			

Components of variation

Source	Estimate	Sq. Root
Feeding	0.7545	0.8686
Salinity	−0.0091	−0.0956
Feeding × Salinity	1.0341	1.0169
Residual	0.201	0.4483

(c) PERMANOVA table

Source	df	SS	MS	(Pseudo-)*F*	*P*
Feeding	1	0.02	0.02	0.12	0.8968
Salinity	2	0.302	0.151	0.89	0.4735
Feeding × Salinity	2	46.551	23.275	136.93	0.0001
Residual	42	7.14	0.17		
Total	47	54.01			

Components of variation

Source	Estimate	Sq. Root
Feeding	−0.0063	−0.0791
Salinity	−0.0012	−0.0344
Feeding × Salinity	2.8882	1.6995
Residual	0.1699	0.4123

As is clear from such simple examples, the significance of an interaction term in PERMANOVA requires the investigator to unpack its meaning by appropriate pairwise or other subset tests, and (importantly) by examining directional changes in low-dimensional approximations to the dissimilarity

doi:10.1111/aec.13059

space, where such approximations are sufficiently reliable – pairwise tests alone cannot 'see' the directional nature of an interaction, such as occurs in Fig. 5a for example, just as the equivalent 1-way ANOSIM tests could not. The third example above is certainly contrived, but it does starkly make the important point that ANOSIM is testing something quite distinct from PERMANOVA: its null hypothesis is not making a statement about the positions of main-effect centroids in the dissimilarity space of the PERMANOVA model. Instead, as has been described, it tests for the presence of *any* effects of a factor, by non-parametrically calculating the magnitude of differences between its levels (using a standardised measure of distinctiveness among groups, R), separately for each stratum of the other factor(s), and averaging those magnitudes. The null hypothesis, that none of these differences in levels exist, is tested solely by permuting within the strata of the other factor(s), ensuring incidentally that this is always an 'exact' test, in terms of its p-values, under all possible permutations. In contrast, PERMANOVA permutes estimated residuals from the fitted model (Anderson 2001b) and the resulting P-values are only *asymptotically* exact, that is for 'sufficiently large' degrees of freedom in the denominator of the specific pseudo-F statistics. The example of Fig. 5a illustrates that ANOSIM statistics can never 'see' the PERMANOVA-based interaction effects arising from differing directions of change in the dissimilarity space, which do not also generate differing magnitudes of change (differing distinctiveness of the groups for the tested factor) at each level of the removed factor. But, equally of course, any interaction effects cannot interfere with the assessment of size or significance of the overall changes for each factor which the 2-way ANOSIM detects. Importantly, the ANOSIM R statistics scale effects in a way that can be fully compared, across both the overall assessment ($\overline{R}$ or $\overline{R}^{O}$) and within each stratum of the removed factor (R).

Returning to the real data set of the microcosm experiment on estuarine nematode communities for this same 2-way crossed design of 2×3 levels, the PERMANOVA table and components of variation are given in Table 3. Again an unwisely naïve interpretation of the main effects would conclude that, though there is a large effect of salinity – clearly seen in the placement of C, M and H symbols in the mMDS of Fig. 4a – there is not a significant effect of the feeding regime ($p \approx 7\%$), indicated by closed *vs.* open symbols. In contrast, whilst the 2-way ANOSIM test for an overall salinity effect, eliminating any contribution from feeding effects, also has a large and significant ordered statistic of $\overline{R}^{Oc} = 0.75$ (or unordered statistic $\overline{R} = 0.69$), its test for an overall feeding effect removing any salinity contributions is also strongly significant ($p < 0.1\%$), though with a smaller average value of $\overline{R} = 0.23$. There is, of course, a significant interaction in Table 3 ($p < 0.01\%$), illustrating for the fixed effects crossed design once again – for a real data set this time – that a main effect in a PERMANOVA table is not a direct measure of *magnitude* of community change associated with that factor, when this is averaged over the levels of the other factor(s), as was defined above for the ANOSIM statistic. This is a case where ANOSIM can suggest the presence of an interaction since a 1-way test shows no feeding effect within the Control salinity level ($R = -0.01$) but clearly significant effects for Medium and High salinity stress ($R = 0.24$ and 0.46); there is a progressive change in the magnitude of feeding effects with increasing salinity stress.

Interestingly, whilst not clear from the nMDS of replicates in Fig. 4a, the means plot of Fig. 4b shows that the interaction detected by PERMANOVA involves a contrary direction of change, the latter having the effect of more or less eliminating the feeding main effect, much as in the 2-d illustration of Fig. 5a (Table 2a). That the interaction is not likely, however, to be solely one of directional change is suggested by the comparison of magnitudes of feeding effects, seen above for the individual ANOSIM statistics at each salinity stress level. It is important for the graphical side of this interpretation to note that the PERMANOVA-based centroids plot (Anderson *et al.* 2008) is literally indistinguishable here from the means plot of Fig. 4b – as seems almost always to be the case when the averages are *based on similar numbers of replicates at each level*. To be clear, the means of transformed replicate data, then subjected to dissimilarity computation, are *not* the same as centroids calculated in the high-d 'dissimilarity space', but the internal *relationships* among the former are often indistinguishable from those among the latter, and thus low-dimensional ordination plots of means or centroids often appear effectively identical.

Averaging in the transformed species space (means plot), rather than in the dissimilarity space (centroid plot), has one substantial advantage: it retains a species × samples matrix, allowing average transformed abundances of particular species to be superimposed on the means ordination, as in bubble (or segmented bubble) plots (e.g. Clarke & Gorley 2015). Figure 4c shows an extreme form of segmented bubble plot in which the means for fourth-root transformed counts of all 45 species, for the 6 experimental combinations, are displayed in clockwise segments (technically sectors) at each of the ordination positions in Fig. 4b. A constant scale for segment length is used throughout and, naturally, species ordering is the same, with the sequence determined by descending

doi:10.1111/aec.13059

Table 3. PERMANOVA for 2-way crossed test (9999 random permutations) of the real estuarine nematode data of Fig. 4. *P*, from 9999 permutations is given as a probability, that is $P = 0.0690$ implies non-significance at the 5% level. The residual df is smaller (by 1) than for Table 2c, due to loss of a replicate from the design

PERMANOVA table

Source	df	SS	MS	Pseudo-*F*	*P*
Feeding	1	1602.4	1602.4	1.92	0.0690
Salinity	2	30376	15188	18.16	0.0001
Feeding × Salinity	2	5969.9	2984.9	3.57	0.0001
Residual	41	34294	836.4		
Total	46	72243			

Components of variation

Source	Estimate	Sq. Root
Feeding	32.61	5.711
Salinity	916.92	30.281
Feeding × Salinity	274.81	16.577
Residual	836.44	28.921

values for (average transformed) abundance in the control condition of Normal feeding and Control salinity. (The almost inevitable shape made by the segments at this control condition suggests the name 'Nautilus plot' for this construction which, as with all plots in this paper, was created using a combination of options in PRIMER v7). The visual effect is immediate, the main feature is the major reduction in diversity and abundance of the community evident with increasing salinity stress, particularly at the highest level (lowest salinity), but also clear is one possible explanation for the interaction in the PERMANOVA analysis. A single opportunist species (*Diplolaimella stagnosa*), seen (at about 05.30 on the clock face) in modest numbers in the control salinity conditions, increases strongly with salinity stress, under both normal and limited feeding, except at the highest combination of both stressors – where it returns to a level comparable with that for lower or absent stress levels, thus potentially explaining the reversal of trajectories seen in Fig. 4b.

In summary, PERMANOVA can play a crucial role in two-way (and, more generally, multi-way) multivariate analysis of resemblance matrices by providing direct tests for interactions between factors. It can also handle more complex experimental designs, including continuous quantitative covariates. However, PERMANOVA performs an additive partitioning that relies on a strict definition of what is meant by an 'effect' – specifically, an additive shift in the position of the centroid, measured in the units of a chosen dissimilarity measure. Main-effect shifts (differences in the positions of centroids) for individual factors, just as in univariate ANOVA, are calculated in a way that does not take into account any potential influences on these effects that may be introduced by a second factor. The presence of these influences is seen, however, in the test for interaction, though it follows that such an interaction can disrupt simple comparative interpretation of main-effect sizes for the factors.

In contrast, ANOSIM calculates group differences using a standardised *R* statistic calculated separately *within* each level of the second factor, which effectively *removes* any influences of that second factor on the measurement of group differences. The statistic in a 2-way ANOSIM is the mean of these individual *R* values, and permutations are similarly suitably restricted so that each test reflects genuinely independent information about that factor's influence overall, so the magnitude of these statistics may validly be compared, whether or not the two factors interact. The standardisation of these *R* statistics also gives them an absolute interpretation in a wider context, for example for different pre-treatments or even different datasets (from commensurate studies), and they are able to incorporate, in a fully comparative way, a robust concept of serially ordered change for the factor under test.

It should be noted that PERMANOVA and ANOSIM also differ – though less markedly – even for the one-way case, in terms of their null hypotheses, interpretation and response of the statistics to differences in variation among groups (e.g. Anderson & Walsh 2013). Here, we have articulated the generalised 2-way ANOSIM approach and its variants in detail and have given several examples clarifying how and why 2-way ANOSIM and PERMANOVA results may differ dramatically for a given set of data. Typically, the

doi:10.1111/aec.13059

nature of the study design and hypotheses of interest will direct a researcher towards one or other of these approaches in any given situation. ANOSIM usefully yields a more general and broader non-parametric inference, whereas PERMANOVA gives a semi-parametric partitioning with a focus on quantifying positions of centroids in dissimilarity space. We have also demonstrated, however, that ANOSIM and PERMANOVA may be used constructively *in tandem*, along with judiciously chosen ordinations of averages or centroids, to yield even deeper insights into the salient factors or combinations of factors that create structure in multivariate data.

ACKNOWLEDGEMENTS

All those involved with field and experimental work and the subsequent generation of the data used in this work are gratefully acknowledged, including B. Brown, J.M. Gee, R.M. Warwick, M. Austen, A. Rowden, S. Widdicombe, J. Davey, M. Frost, T. Dorrington, A. Douglas, M. Kendall, N. Villano, A. McEvoy, S. Dashfield, M. Attrill, A. Rogers and A. Hood. The authors are immensely grateful to the Reviewers and especially the Guest Editor, Prof. M. J. Anderson, for the time expended on highly detailed and perceptive reviews, unprecedented in our experience, which have led to a greatly improved manuscript.

AUTHOR CONTRIBUTIONS

Paul J. Somerfield: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Validation (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **K. R. Clarke:** Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **R. N. Gorley:** Conceptualization (equal); Formal analysis (equal); Methodology (equal); Software (equal); Validation (equal).

FUNDING

P.J.S. acknowledges funding support from the UK Natural Environment Research Council (NERC) through its National Capability Long-term Single Centre Science Programme, Climate Linked Atlantic Sector Science (Grant no. NE/R015953/1) and from the NERC and Department for Environment, Food and Rural Affairs, Marine Ecosystems Research Programme (Grant no. NE/L00299X/1).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

All data are freely available from the corresponding author.

REFERENCES

Anderson M. J. (2001a) A new method for non-parametric multivariate analysis of variance. *Austral Ecol.* **26,** 32–46.

Anderson M. J. (2001b) Permutation tests for univariate or multivariate analysis and regression. *Can. J. Fish. Aquat. Sci.* **58,** 626–39.

Anderson M. J. (2017) Permutational Multivariate Analysis of Variance (PERMANOVA). Wiley StatsRef: Statistics Reference Online https://doi.org/10.1002/9781118445112.stat07841.

Anderson M. J., Gorley R. N. & Clarke K. R. (2008) *PERMANOVA+ for PRIMER: Guide to Software and Statistical Methods*. PRIMER-E, Plymouth, UK.

Anderson M. J. & Walsh D. C. I. (2013) PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecol. Monogr.* **83,** 557–74.

Austen M. C. (1989) Factors affecting estuarine meiobenthic assemblage structure: a multifactorial microcosm experiment. *J. Exp. Mar. Biol. Ecol.* **130,** 167–87.

Brown B. E., Dunne R. P., Somerfield P. J. *et al.* (2019) Long-term impacts of rising sea temperature and sea level on shallow water coral communities over a ~40 year period. *Sci. Rep.* **9,** 8826.

Clarke K. R. (1988) Detecting change in benthic community structure. Proc XIVth Int Biometric Conf, July 1988, Namur: Invited Papers I-6, p13-24. Société Adolphe Quetelet, Gembloux, Belgium.

Clarke K. R. (1993) Non-parametric multivariate analyses of changes in community structure. *Aus. J. Ecol.* **18,** 117–43.

Clarke K. R. & Gorley R. N. (2015) *PRIMER v7: User Manual/tutorial*. PRIMER-E, Plymouth, UK.

Clarke K. R., Gorley R. N., Somerfield P. J. & Warwick R. M. (2014) *Change in Marine Communities: An Approach to Statistical Analysis and Interpretation*, 3rd edn. Plymouth, PRIMER-E.

Clarke K. R. & Green R. H. (1988) Statistical design and analysis for a 'biological effects' study. *Mar. Ecol. Progr. Ser.* **46,** 213–26.

Clarke K. R., Somerfield P. J., Airoldi L. & Warwick R. M. (2006) Exploring interactions by second-stage community analyses. *J. Exp. Mar. Biol. Ecol.* **338,** 179–92.

Clarke K. R. & Warwick R. M. (1994) Similarity-based testing for community pattern: the 2-way layout with no replication. *Mar. Biol.* **118,** 167–76.

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

Freedman D. & Lane D. (1983) A nonstochastic interpretation of reported significance levels. *J. Bus. Econ. Stat.* **1,** 292–8.

doi:10.1111/aec.13059

Gee J. M. & Warwick R. M. (1994a) Metazoan community structure in relation to the fractal dimensions of marine macroalgae. *Mar. Ecol. Progr. Ser.* **103,** 141–50.

Gee J. M. & Warwick R. M. (1994b) Body-size distribution in a marine metazoan community and the fractal dimensions of macroalgae. *J. Exp. Mar. Biol. Ecol.* **178,** 247–59.

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

Law R. J. & Kelly C. (2004) The impact of the "Sea Empress" oil spill. *Aquat. Living Resour.* **17,** 389–94.

Mantel N. (1967) The detection of disease clustering and a generalised regression approach. *Cancer Res.* **27,** 209–20.

Somerfield P. J., Clarke K. R. & Gorley R. N. (2021) A generalised Analysis of Similarities (ANOSIM) statistic for designs with ordered factors. *Austral Ecol.* **46,** 901–910.

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,** 581–93.

Somerfield P. J. & Warwick R. M. (1999) *Sea Empress contract No. FC 73-02-68. Appraisal of environmental impact and recovery using Laminaria holdfast faunas*. Final report. Plymouth Marine Laboratory, Plymouth, UK.

SUPPORTING INFORMATION

Additional supporting information may/can be found online in the supporting information tab for this article.

Appendix S1. Numbers of potentially distinct permutations for 1- and 2-way ANOSIM tests.

doi:10.1111/aec.13059