Austral Ecology (2021) ••, ••–••

Analysis of similarities (ANOSIM) for 3-way designs

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, PL1 3DH, UK*

Abstract Analysis of similarities (ANOSIM) is a robust non-parametric hypothesis-testing framework for differences in resemblances among groups of samples. To date, the generalisation and use of ANOSIM to analyse various 2-way nested and crossed designs with unordered or ordered factors has been described. This paper describes how the 2-way tests may be extended and modified for the analysis of 3-way designs, including the introduction of a different type of constrained permutation procedure for a design in which one factor is nested in another and crossed with a third. The construction of 3-way tests using the generalised statistic in various nested and crossed designs, with or without ordered factors, and with or without replication, is described. Applications of the new tests to ecological data are demonstrated using three marine examples. They are as follows: a study of changes in fish diet for fish of increasing size sampled in different locations at different times (a 3-way fully crossed design with ordered factors); a hierarchical spatial study of the fauna inhabiting kelp holdfasts (a 3-way fully nested design with unordered factors); and a study of infaunal macrobenthos in which sites within areas were resampled over a long time series (a design in which sites are nested in areas but crossed with years, both latter factors potentially being ordered). The magnitudes of the ANOSIM statistics provide information about relative effect sizes (accounting for other factors), which is often a focus for multifactorial designs. Though the described ANOSIM tests do not provide parallels for all the range of 3-way mixed-factor designs possible in ANOVA (and its multivariate semi-parametric counterpart PERMANOVA), it is seen that for nested factors these ANOSIM tests parallel the matching PERMANOVA random-effects models, and not their fixed-effects counterparts, thus allowing the same broader inference about the space from which these random factor levels are drawn.

Key words: hypothesis tests, multifactorial designs, multivariate data, non-parametric statistics, ordered factors.

INTRODUCTION

The assumptions underlying many univariate and multivariate statistical tests are often grossly invalid for multivariate ecological community data, such as abundances of taxa in samples, owing to the nature of the data (variables are generally right-skewed and heteroscedastic, the dominant entry in the matrices is zero, etc.). To address the many statistical difficulties, a robust non-parametric multivariate strategy for the analysis of community data was described by Field *et al.* (1982). The analytical strategy and methods were expanded and clarified by Clarke (1993) and continue to develop (Clarke *et al.* 2014; Somerfield *et al.* 2021a,b). A key formal hypothesis test within the framework is ANOSIM (Analysis of Similarities), originally described for one-way layouts by Clarke and Green (1988). Clarke (1988, 1993) showed how ANOSIM can be extended to two-way nested and crossed layouts with replication: 2-factor nested, B within A (denoted by B(A)) and 2-factor crossed (denoted A × B). Clarke and Warwick (1994) described how the special case of A × B in which there are no replicates may be analysed. Such designs sometimes arise either because only one sample was taken for each combination of A and B, or replicates were taken but considered to be 'pseudo-replicates' (*sensu* Hurlbert 1984) and pooled.

Somerfield *et al.* (2021a) redefined the ANOSIM *R* statistic of Clarke and Green (1988), demonstrating that a generalised ANOSIM statistic R^O is the slope of a linear regression of the ranks of observed resemblances on the ranks of model distances, where the model is a resemblance matrix characterising the alternative hypothesis. This formulation extends ANOSIM from a test for unordered differences among groups to a framework that can also be used to analyse ordered factors, for example testing for spatial or temporal trends. The statistic has a common form but the notation distinguishes the different hypotheses being tested: *R* is the classic ANOSIM statistic in a test for differences between unordered groups, R^{Oc} is the statistic for ordered groups when there are replicates within groups, and R^{Os} is the equivalent statistic when there are no replicates (each 'group' is a single sample). Somerfield *et al.* (2021b) showed how the treatment of ordered factors could

*Corresponding author.
Accepted for publication June 2021.

doi:10.1111/aec.13083

be incorporated into analyses of 2-way designs, and in this paper, the framework is extended to include 3-factor designs.

METHODS

ANOSIM for 3-way designs

The first step is briefly to recap the definition of 2-way hypothesis tests, as described in detail by Clarke (1993) and Somerfield *et al.* (2021b). In a 2-way crossed test (denoted A × B) the effect of factor B on a test of A may be removed altogether by calculating R_A (whichever of the definitions $R/R^{Oc}/R^{Os}$ is used) within each level of 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 to be within levels of B, corresponding to the null hypothesis that there is no effect of A at *any* level of B. As the design is crossed, the converse hypothesis may be tested, namely whether there is an effect of factor B having removed any effect of factor A.

An example of a fully crossed 3-way design (denoted A × B × C) could be replicate samples from a set of locations (A) each examined at the same set of times (B) and for the same set of depths (C). A fully symmetric design like this can be addressed by testing each factor in turn (A, say), by 'flattening' the other two into a single factor (B × C) whose levels are all the possible combinations of levels of B and C. The test for A from the relevant 2-way crossed design is then carried out (see Somerfield *et al.* 2021b, noting particularly the Discussion on the valid interpretation of the test statistic for A, irrespective of whether there are, or are not, interactions with the B × C levels). Similarly, the global test for time effects (B removing A × C) will only compare those different times at the same depth and location, and will then average those time-comparison statistics across all depth-by-location levels. Whichever of the definitions is used ($\overline{R}/\overline{R}^{Oc}/\overline{R}^{Os}$), the three global statistics (A removing B × C, B removing A × C, C removing A × B) can be directly compared to gauge overall relative importance of the A, B and C factors.

For a 2-way nested analysis with B nested in A (denoted B(A)), the initial test for effects of B (Clarke, 1993; Somerfield *et al.* 2021b) is performed in the same way as testing B in the crossed design B × A. The 2-way ANOSIM statistic ($\overline{R}_B/\overline{R}_B^{Oc}/\overline{R}_B^{Os}$ as appropriate) is computed and the permutations are carried out among levels of B within levels of A. For the (usually more important) second test for effects of A, in concept the averaged B levels over their replicates become the replicates for a 1-way ANOSIM test of A. (In reality, the non-parametric status can be maintained by averaging the ranks of the relevant dissimilarities and re-ranking the result).

This can be extended to the 3-way fully nested design C(B(A)), for example sub-areas (C) nested in sites (B), nested in locations (A), by repeated application of the 2-way case. This tests the lowest factor (C) inside the levels of the next highest (B), then averages at the replicate level so that levels of C are now replicates for a test of B, then averages at the levels of C so that B levels are the replicates for a test of A. The Discussion returns to the issue of the differing ways in which this averaging may be carried out.

Testing is also straightforward for designs having a structure of C(A × B), in which C is nested in all combinations of A and B. For example, multiple sites (C) are chosen from all combinations of location (A) and habitat type (B), in a case where all habitat types are found at each location, with replication (or not) at each site. The test for C uses the A × B 'flattened' factor at the top level of a 2-way nested design, and tests for A and B are exactly as for the 2-way crossed design but, if replicates exist, averaging over the appropriate ranks to obtain a reduced matrix, then re-ranked to utilise the levels of C as replicates for the crossed A and B tests.

The only other practical type of 3-way sampling design, and one which is quite frequently encountered, is B × C (A), in which only C is nested in A, and B is crossed with C. An example of such a design is when multiple sites (C) are identified in a number of areas (A), and the *same sites* are returned to at each of a number of times (B). (Note that here, and throughout the paper, the term 'Area' is used synonymously with 'Location', to represent the top level of a spatial design.) The building blocks of a test for A (Fig. 1 a) are the 1-way ANOSIM statistics R_A (or R_A^{Oc} if A is considered ordered) for a test of Areas (A), using as replicates the Sites (C) in each Area, computed separately for each Time (B). (If there are replicates within the sites these need to be pooled or averaged, perhaps by averaging the appropriate rank dissimilarities and re-ranking, as in the nested designs C(B(A)) and C(A × B). The key point is that the correct nested levels, that is the sites and not the replicates, must be used to test the areas.) The R_A (or R_A^{Oc}) statistics are then averaged over the levels of B, to obtain the overall test statistic for A of $\overline{R}_A$ (or $\overline{R}_A^{Oc}$) exactly as for the usual 2-way crossed case A × B. The crucial difference here is in generating the null hypothesis distribution for this test statistic (Fig. 1a). Permuting the sites across the areas separately for each year, as the standard A × B test would do, assumes that the sites are randomly drawn afresh each time from the defined area (a C(A × B) design), rather than determined only once and then revisited each time. Instead, the permutable units are the entire series of samples representing change through time at each site. Thus, the entire time series for each site is shuffled randomly among the areas, leaving intact the originally recorded ordering of observations through time for each site, and $\overline{R}_A$ (or $\overline{R}_A^{Oc}$) re-calculated for each permutation. There are consequently many fewer permutations for the test of A under this B × C (A) rather than C(A × B) design, but this may be compensated for by improved focus when examining the B time factor: subtle assemblage changes from year to year may be seen by returning to the same site(s), which might otherwise get swamped by large spatial variability from site to site if these are randomly reselected each year.

The test for times (B) is straightforward if there are genuine replicates taken from each site. This is now just a standard two-way crossed design ('B × C') where it is understood that C represents all the different sites, the area (A) which they come from being immaterial: site and area factors are excised as in all two-way crossed ANOSIM tests, by calculating R_B (or R_B^{Oc}) among times separately for

doi:10.1111/aec.13083

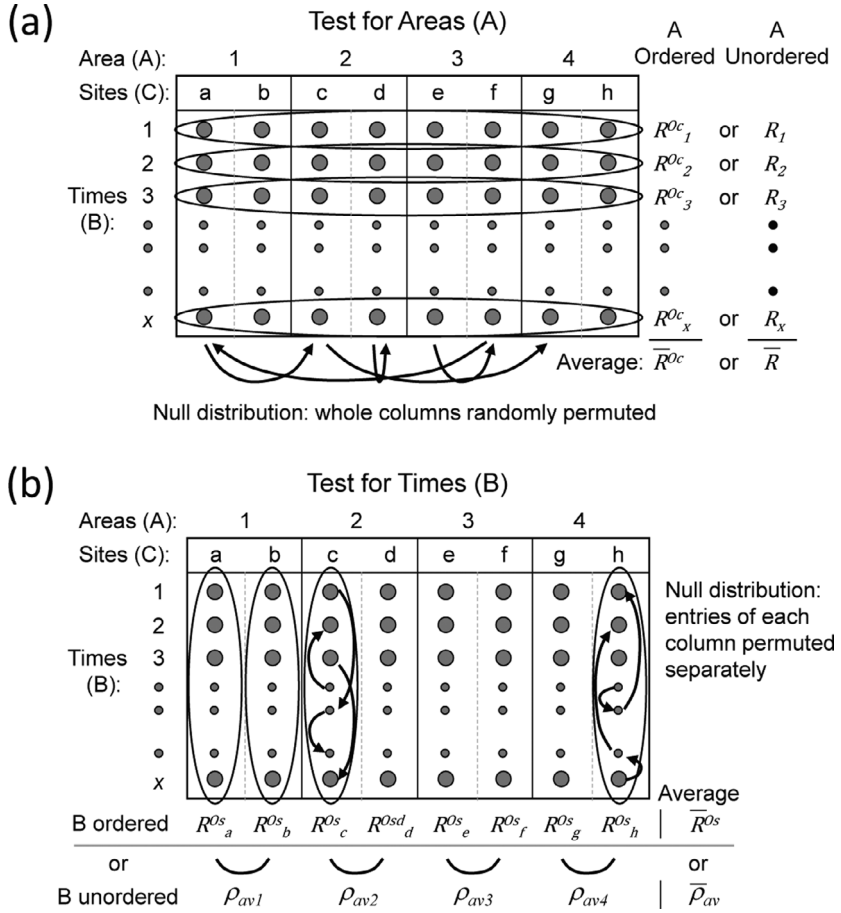

Fig. 1. Schematic diagrams of: (a) the test for factor A in a B × C(A) design with factors A: Areas (1–4); B: Times (1, 2, 3, . . . , *x*); and C: Sites (a–h), pairs of which are nested within Areas. Each circle represents a sample (or a set of samples, see text). The key point to note is that the permutable units are the whole columns representing time series at each Site; (b) the test for factor B in a B × C(A) design where replicates (if they existed) were pooled within combinations of Site and Time to give single samples. The permutable units are now single time observations within each site. See text for details.

each of these 'site within area' levels and averaging to give $\overline{R}_B^{Oc}$. Permutations are also carried out as usual, permuting the replicates across the times but constrained to stay within their own site. If only a single sample is available from each site at each time (Fig. 1b), perhaps because pseudo-replicates are pooled, a test statistic for times (B) must either exploit serial change (R^{Os} statistics) or the matching of time patterns (ρ_{av} statistics) among sites within locations. The latter is a weaker form of 2-way crossed test statistic available in the absence of replication, which establishes the significance of the time factor (B) by demonstrating commonality of time patterns across sites, here within areas (Clarke & Warwick 1994; Somerfield *et al.* 2021b). An average of either of these statistics across areas provides the test statistic for a time effect and permutation is again simply one of random shuffling of time labels independently for each site (Fig. 1b).

It is evident that there are a sizeable number of possible mixtures of design and test statistic, which are summarised in Table 1. This details all viable combinations of three factors, A, B, C, in crossed/nested form, with ordered or unordered factors, and with or without replication at the lowest level. For each, it gives the appropriate test statistic and its method of construction and indicates when pairwise tests are either not feasible (e.g. the test is based on a singly ordered R^{Os} or matching statistic ρ_{av}, which require more than two levels) or not logically desirable (e.g. pairwise tests of nested factors). In the final column, there are some examples of (marine) ecological studies in which the factors would have the right structure for such a test. The case designations in the first column (e.g. 3c) are cross-referenced in the Results.

Data analyses

All the analyses were undertaken with PRIMER v7 (Clarke & Gorley, 2015). Testing utilised the ANOSIM routine with 9999 random permutations, where the full set of possible permutations could not be enumerated. In order to visualise effects of factors, pre-treated data were averaged over replicates and inter-sample Bray–Curtis resemblances from these means ordinated using non-metric multidimensional scaling (nMDS).

doi:10.1111/aec.13083

Table 1. 3-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

No.	Type of design	Factors	Factor levels ordered?	Replicates?	Statistics used	Pairwise test?	Construction of test	Examples
3a	3-way crossed	A×B×C	A,B,C unordered	Yes	A,B,C: $\overline{R}$	Yes	As two-way crossed test, but combining pairs of factors in turn, for example calculating 1-way R for A within all B×C levels[†]	A: location, B: time, C: habitat
3b	3-way crossed	A×B×C	A,B,C unordered	No	A,B,C: ρ_{av}	No	As two-way crossed test with no replication, that is comparing resemblance matrices of A across combined B×C levels[†]	As 3a above but no reps (or pooled)
3c	3-way crossed	A×B×C	A,B unordered C ordered	Yes/no	A,B: $\overline{R}$ C: $\overline{R}^{Oc}/\overline{R}^{Os}$	Yes/no	A,B: as test 3a/3b C: as 2-way crossed test, collapsing A,B to single factor A×B[†]	A: location, B: time, C: depth range with/without reps in A×B×C cells
3d	3-way nested, C within B within A	C(B(A))	A,B,C unordered	Yes	A: R B,C: $\overline{R}$	A: Yes B,C: No	A,B: as 2-way nested test of B in A, using levels of C as replicates[‡] C: as 2-way nested test for C in all B levels (i.e. over all A levels)	A: region, B: location, C: site, with replicate samples at each site
3e	3-way nested, C within B within A	C(B(A))	A,B,C unordered	No	A: R B: $\overline{R}$ C: –	A: Yes B: No C: –	A,B: exactly as for test 3d (except no averaging of C level reps needed) C: no basis for a test	A: region, B: location, C: site, with one pooled sample at each site
3f	3-way nested, C within B within A	C(B(A))	A,B unordered C ordered	Yes/no	A: R B: $\overline{R}$ C: $\overline{R}^{Oc}/\overline{R}^{Os}$	A: Yes B,C: No	A,B: as 2-way test of B nested in A, using ordered C levels (/single C values) as reps[§] C: as 2-way ordered test of C nested in B(A), all B levels over A	A: location, B: shore, C: along shore transect, reps (or not) at transect points
3g	3-way nested, C within B within A	C(B(A))	A unordered B ordered, C either	Yes/no	A,C: as 3f B: $\overline{R}^{Oc}$	A: Yes B,C: No	A,C: as the relevant tests in 3d–3f B: as 2-way nested test for B (ordered) within A, using levels of C (/single C values) as reps[¶]	A: sea region, B: transect of sites, C: random days at each site (with/without rep trawls)
3h	3-way, C nested in A × B	C(A×B)	A,B,C ordered or unordered	Yes/no	Various	A,B: Yes C: No	A,B: as for 2-way crossed tests but using C levels as reps (averaged where needed)[††] C: as for 2-way nested, C in all combinations A×B	A: location, B: season, C: different site-day combinations in each A×B (with/without rep. cores)
3i	3-way, B crossed with C(A) (i.e. only C is nested in A)	B×C(A)	A,B,C unordered	Yes	A: $\overline{R}$ B: $\overline{R}$ C: $\overline{R}$	A: Yes B: Yes C: No	A: average the reps in C levels (on resemblances[‡]), then 2-way crossed statistic for A from A×B[‡‡] B: usual 2-way crossed test for B across all levels of C (over all A) C: usual 2-way nested test for C within all combined levels A×B	A: location, B: time, C: same random sites in location returned to each time, with replicate samples at sites
3j	3-way, B crossed with C(A)	B×C(A)	A,B,C unordered	No	A: $\overline{R}$ B: $\overline{\rho}_{av}$ C: $\overline{\rho}_{av}$	A: Yes B: No C: No	A: as 3i but with single C levels as reps (constrained perms again[‡‡]) B: ρ_{av} statistic for B patterns matched over C levels in each A, then averaged (normal perms)[§§] C: converse ρ_{av} of C patterns, for each A, matched across B levels, then ρ_{av} averaged over A[¶¶]	A: location, B: time, C: same random sites in location returned to each time for single sample (or pooled sample)

doi:10.1111/aec.13083

Table 1. *Continued*

No.	Type of design	Factors	Factor levels ordered?	Replicates?	Statistics used	Pairwise test?	Construction of test	Examples
3k	3-way, B crossed with C(A)	B×C(A)	A unordered B unordered C ordered	Yes/no	A: $\overline{R}$ B: $\overline{R}/\overline{\rho}_{av}$ C: $\overline{R}^{Oc}/\overline{R}^{Os}$	A: Yes B: Yes/no C: No	A: as test 3i/3j B: as test 3i/3j C: as 2-way test of C(A×B), that is C nested in all A×B combinations	A: location, B: time, C: same (representative) transect of sites in location returned to each time
3l	3-way, B crossed with C(A)	B×C(A)	B ordered A,C ordered or unordered	Yes/no	B: $\overline{R}^{Oc}/\overline{R}^{Os}$ A,C: as 3i–k or 3m	B: Yes/no A: Yes, C: No	B: 2-way crossed test of ordered B with all levels of C (in all A) A,C: as the relevant tests for A,C in 3i–k,m	A: location, B: yearly time trend, C: same random sites in each location each year
3m	3-way, B crossed with C(A)	B×C(A)	A ordered B,C ordered or unordered	Yes/no	A: $\overline{R}^{Oc}$ B,C: as 3i–l	A: Yes B: Yes/no C: No	A: 2-way crossed test of ordered A across B, using C levels as reps (C reps if present are averaged‡), B levels held as a block in perms‡‡ B,C: as the relevant tests in 3i–3l	A: latitudinal region, B: yearly trend, C: same transect of sites in region each year (with/ without reps). A, B,C ordered

Also given are the existence (or not) of pairwise tests, details of the test constructions and examples of contexts in which they might be employed.

†Test for A uses average of 1-way R (for A) across all levels of B and C in combination (B × C), then B v (A × C) and C v (A × B). Same idea is used for 3b (use the 2-way test for unreplicated designs), and if two of the factors are ordered still use 3a, b or c.

‡Starts from ranked resemblances of C replicates, which are then averaged and re-ranked (twice for the A test). Or (e.g. if unsure of quality of C replicates) test A & B by averaging C replicates in data matrix and using a 2-way test on A and B(A).

§C levels (averaged where needed, as in note ‡) are assumed representative replicates of B(A) condition.

¶If A ordered (whether B, C are or not), it changes nothing except the test of A, in which ordered levels of B are now assumed representative as replicates.

††Similar comments as for note ‡ apply, about whether it may be better sometimes to average replicates of C externally, on the data matrix, then calculate resemblances and submit to the 2-way crossed cases for A × B.

‡‡Note the necessity for a block-constrained permutation test here under the null, with values across B for each C level being permuted as a batch across C(A) and A levels. A common structure is A: locations, C: sites (nested in A), B: period, all sites visited in each period. Test for A uses sites as replicates but keeps the periods for each site together under permutation across locations.

§§This is a new doubly averaged statistic $\overline{\rho}_{av}$ matching patterns in B over the C levels for each A level (the usual ρ_{av}), then averaging ρ_{av} over A levels. Permutations are the usual random ordering of B for each C(A).

¶¶For example ρ_{av} calculated to match relationships among sites for different periods, separately for each location, then ρ_{av} averaged over locations. Standard permutation of sites within all levels of location × period.

Example data

Diets of Western King Wrasse Coris auricularis *(Valenciennes 1839)*

Lek *et al.* (2011) studied the diets of labrid fish in Western Australia. The data used here are the composition by volume of taxa found in the foreguts of Western King Wrasse from two regions of the western Australian coast. Taxonomic composition of the prey assemblage was recorded in 23 broad groups (gastropods, bivalves, annelids, ophiuroids, echinoids, small and large crustaceans, teleost fish, etc.). The fish are 'doing the sampling' of the assemblages so there is no control over the total volume of material in each gut. Individual fish may have little content in their foregut at any given time, so to make viable samples of ingested material, four fish guts at a time were randomly selected from a given size class, location and sampling time, and pooled to make a single sample. Total gut content of such a sample may still vary substantially over replicates, so these prey-category volumes were then sample-standardised, so the basic replicates input to all subsequent analyses are of percentage composition, that is all taxa add up to 100% for each sample. A mild (square root) transformation was applied and Bray–Curtis dissimilarity among samples was calculated. The sampling design for this study, carried out at Jurien Bay Marine Park (JMBP) and on the Perth coast and simplified for the current illustration, is here treated as a 3-way, fully crossed design (A × B × C), with factors: A = 3 regions/habitat levels (JBMP/exposed, JMBP/sheltered and Perth/exposed locations); B = 4 body sizes of the wrasse predator (ordered length classes, 1: <150 mm, 2: 150–199 mm, 3: 200–249 mm, 4: >250 mm); C = 2 seasons (summer/autumn and winter/spring), and with two replicate samples from each of the 24 combinations of these levels.

doi:10.1111/aec.13083

New Zealand kelp holdfast fauna

In north-eastern New Zealand, Anderson *et al.* (2005) examined assemblages of invertebrates colonising holdfasts of the kelp *Ecklonia radiata* (C. Agardh) J. Agardh. Holdfasts were collected according to a structured hierarchical experimental design. The largest spatial scale examined was locations (four levels, separated by hundreds of kilometres) spanning a large stretch (~290 km) of the north-eastern coast of New Zealand. Within each location, two sites were randomly chosen (separated by hundreds to thousands of metres), and within each site, two sub-areas (separated by 10s of metres) were chosen haphazardly. Within each sub-area (measuring approximately 10 × 10 m), five replicate holdfasts (separated by several metres) were taken haphazardly from the kelp forest by divers between 7 January and 5 February 2002. Bags containing holdfasts were brought to the surface, opened and the fauna was relaxed using a solution of magnesium chloride, then fixed and stored in formalin. In the laboratory, the fauna was rinsed onto a 0.5 mm sieve. All organisms retained on the sieve, or remaining attached to the holdfast, were identified to the finest level of taxonomic resolution possible. The abundance of each taxon was counted and recorded. For organisms that were encrusting or colonial (some sponges, ascidians, bryozoans, etc.), an ordinal semi-quantitative score from 0 to 3 was given according to the relative coverage of the organism on the holdfast: 0 = absent, 1 = present but rare, 2 = present and fairly frequently encountered and 3 = present and very common throughout the holdfast. Data consists of counts for each of 351 taxa from a total of 80 holdfasts. Abundances were sample-standardised (as holdfast volumes varied) and square-root-transformed prior to calculating Bray–Curtis similarities. This is a fully nested design, C(B(A)) with four locations (A), two sites (B) within each location, two sub-areas (C) within each site and five replicate holdfasts within each sub-area.

Tees Bay soft-sediment macrofauna

As part of a wider study of the Tees Bay and estuary off the northeast coast of England (Warwick *et al.* 2002), samples of soft-sediment macrofauna were collected annually in September from 1973 to 1996 from two sites within each of four areas in Tees Bay (Fig. 2a). Samples were collected using a 0.1 m^2 grab, sieved on a 1 mm mesh, preserved in formalin and subsequently identified and counted. The data are abundances of 282 taxa in 192 samples, representing a mixed nested and crossed design B × C(A) where A = sub-tidal Areas 1–4 (Fig. 3), with C = two Sites within each Area, the same sites being returned to each September over B = 24 years (1973–1996). Sites (C) are therefore nested in Areas (A) but crossed with Years (B). There was a further level of replication, with multiple grab samples collected at each site on each sampling occasion, but these have been averaged to give a more reliable picture of the assemblage on each occasion (the repeated grabs from a single stationing of the ship are considered 'pseudo-replicates' in time, and possibly space). Samples were fourth-root transformed before calculating Bray–Curtis similarities.

RESULTS

Diets of Western King Wrasse *Coris auricularis*

Using three-factor crossed ANOSIM (A × B × C, case 3c in Table 1, but for B ordered rather than C), testing for A (region/habitat) within all eight combinations of B (length class) and C (seasonal period) levels gives $\overline{R}_A = 0.26$ (p = 1.5%, on a random subset of 9999 from the 15^8 possible permutations, see Appendix to Somerfield *et al.* 2021b). The pairwise tests between the three regions/habitat levels (now on $3^8 = 6561$ permutations) give similar values of $\overline{R}_A$ between 0.20 and 0.29. The ordered ANOSIM test for length class B, across the six strata of A and C, has a larger $\overline{R}_B^{Oc}$ of 0.49 (p < 0.01%) with a clear pattern in the pairwise $\overline{R}_B$ values, which increase with increasing separation of the four wrasse size classes ($\overline{R}_{12}$, $\overline{R}_{23}$, $\overline{R}_{34}$ = 0, 0.21, 0.08; $\overline{R}_{13}$, $\overline{R}_{24}$ = 0.46, 0.5; $\overline{R}_{14} = 0.63$; p < 5% only for the last three tests). Unsurprisingly therefore, the ordered ANOSIM test outperforms the equivalent unordered test (case 3a), which has $\overline{R}_B = 0.33$ (p = 0.06%). The test for period C, removing A and B, gives no effect, with $\overline{R}_C = 0.003$ (p = 48.4%). The key point to note is that the three global statistics, $\overline{R}$ or $\overline{R}^{Oc}$ of A: 0.26, B: 0.49, C: 0.0 (and pairwise values), are directly comparable as measures of the effect size for each factor (and their pairs of levels). In general terms, the levels of the ANOSIM statistics are not affected by the differences in group sizes, in sharp contrast to the P (or p%) values in the associated hypothesis tests, which never escape strong dependence on the group sizes and thus on the number of potentially distinct permutations (see the Appendix to Somerfield *et al.* 2021b).

Although the non-parametric ANOSIM does not test directly for difference in 'centres' among groups of samples, a natural and appropriate follow-up to successful rejection of the null hypotheses is a means plot. This averages the (usually transformed) data for each combination of factor levels and plots an MDS ordination from Bray–Curtis resemblances on those means, thereby aiding the interpretation of effects found to be significant by ANOSIM. Since the period effect is absent ($\overline{R}_C = 0.003$), it is appropriate to average the square-root-transformed data over both the two replicates and two periods at each combination of wrasse size class and location. The resulting nMDS of the mean dietary assemblages for the four size classes at the three locations is shown in Fig. 3a. It has low stress (0.09) and shows the relationships seen in the tests with great clarity. The strongest effect (i.e. for ordered size classes, $\overline{R}_B^{Oc} = 0.49$) is seen as the primary structuring factor in the MDS, with changes in the mean dietary composition from small to large fish observed as broadly parallel trajectories from left to right (in this

doi:10.1111/aec.13083

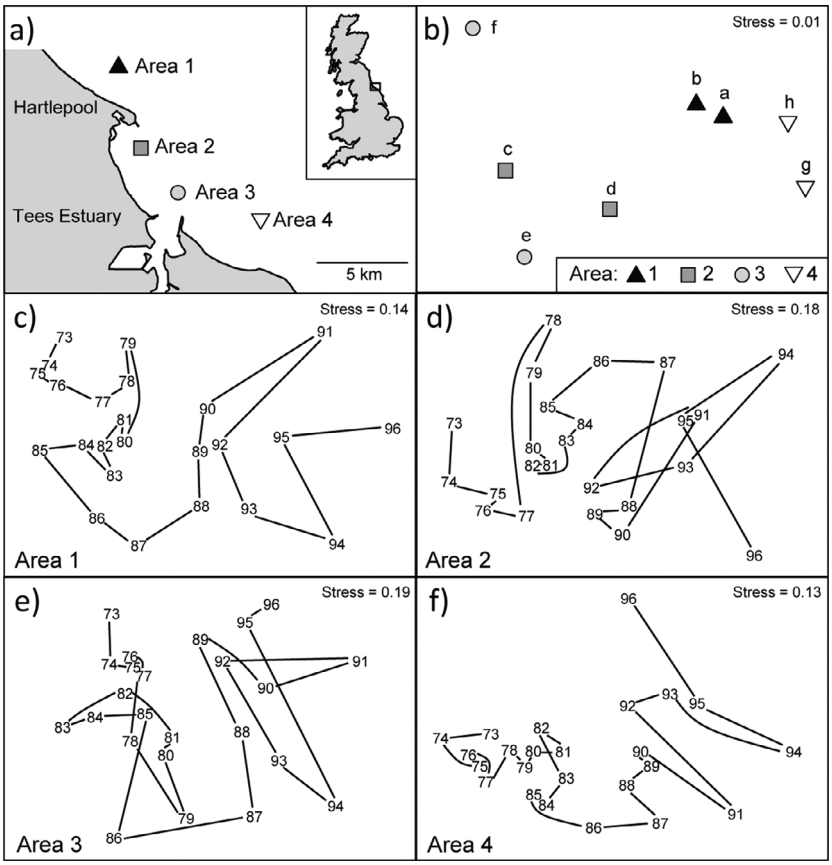

Fig. 2. Tees Bay macrofauna. (a) Map showing the locations of the four sampling areas (Areas 1–4) in Tees Bay, NE England. (b) nMDS of averages of transformed abundances over the 24 time points for the two sites (a–h) in each of the four areas. (c–f) Separate nMDS time-series plots for each area, over 24 years of September sampling. Abundances were fourth-root-transformed then averaged over the two sites in each area, prior to calculating Bray–Curtis similarity.

arbitrary orientation) for each of the three locations/habitat conditions. The latter are also separated (largely vertically) on the plot but to a lesser extent ($\overline{R}_A = 0.26$) than for the fish size classes.

The next step should be to relate the results back to the objective of the study, namely here to understand which taxa are mainly implicated in the steady change in the dietary assemblage through the size classes of King Wrasse in different places. One of the simplest and most effective tools is a bubble plot, superimposing on each ordination point a circle with size proportional to the (averaged) value for a specific taxon in that (averaged) sample. Fig. 3b shows a bubble plot for the 'large crustaceans', which are seen to become an increasing percentage of King Wrasse diet with size, in all three locations.

New Zealand kelp holdfast fauna

In this fully nested design, C(B(A)), factor A (location) is treated as unordered. As factors B (site) and C (sub-area) each have only two levels, there can be no distinction between treating them as ordered or unordered. In a 3-way fully nested ANOSIM for unordered factors (3d in Table 1) the test statistics are therefore $\overline{R}$ and R, and in the order in which they are constructed (small to large spatial scale) give: $\overline{R}_C = 0.28$, $\overline{R}_B = 0.44$ and $R_A = 0.71$ (Fig. 4). These three statistics are, again, directly comparable with one another. Though in this case they demonstrate an increase in magnitude with increasing spatial scale, it is important to appreciate that they are in no way constrained to do this. Their values reflect the degree of biological structuring at each spatial scale, having removed any such differentiation at other scales, and can be seen as a type of non-parametric analogue to variance (or variation) components for classical fully nested designs. They are not calculating variances as such, and there is no partitioning of total variation here. Rather, and arguably more usefully because of the universal scaling of R values, they measure the strength of distinctiveness of groups at each separate level within the hierarchy. Here, as assessed by average rank dissimilarities, the distinctions between sub-areas are small ($\overline{R}_C = 0.28$) in relation to assemblage variation from one holdfast

doi:10.1111/aec.13083

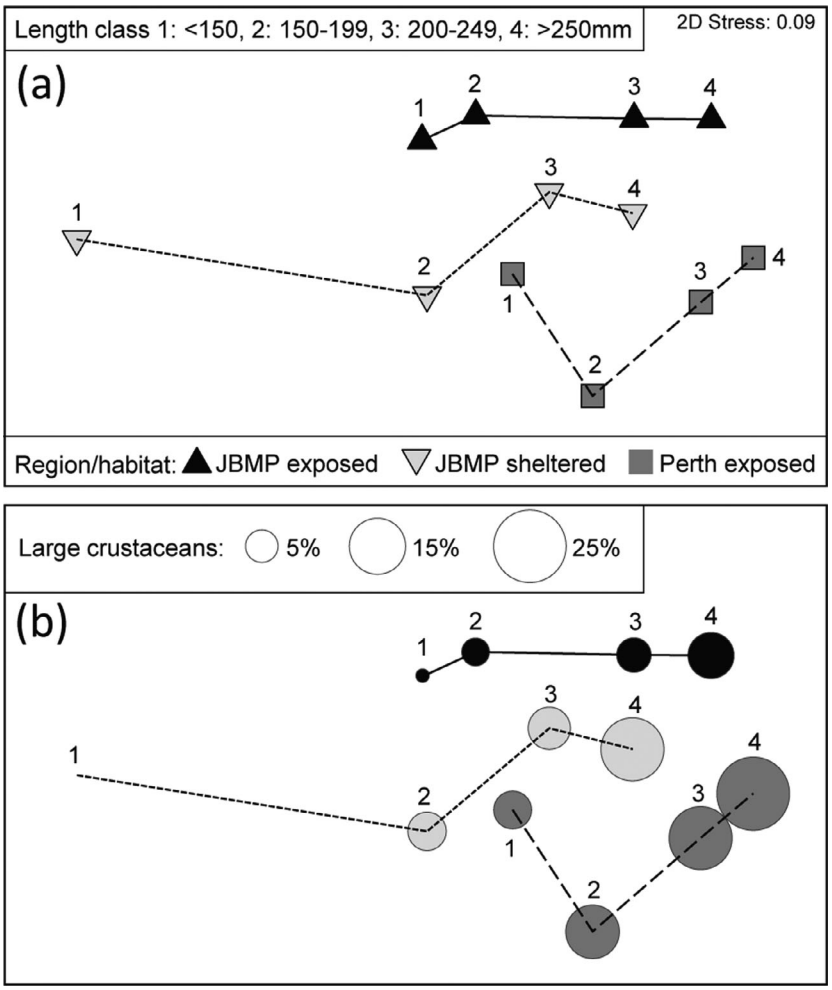

Fig. 3. (a) Ordination by nMDS of Bray–Curtis similarities among samples following standardisation of volumes, square-root taxon transformation and averaging over replicates and seasonal periods, showing clear dietary change with Western King Wrasse body size and between regions/habitats; (b) the same plot overlaid with bubbles of sizes proportional to the percentage of large crustaceans in samples, one component of the average diet.

to another, are somewhat larger between sites ($\overline{R}_B = 0.44$) compared with sub-area differences, and are very large among locations ($R_A = 0.71$) relative to change among sites within those locations. This is in stark contrast to the conclusions one might draw from looking only at the significance levels (as seen from the permutation distributions under the null hypotheses, Fig. 4), C: p << 0.01%, B: p = 1.2%, A: p = 1%, a result of the very different numbers of replicates at each level, and thus possible permutations ($126^8 \approx 6.4 \times 10^{16}$, 81 and 105 respectively). As always with ANOSIM, it is not the p% values but the *R* statistics which describe the magnitude of effects.

Pairwise tests are only meaningful at the top level of such a nested design and there are insufficient permutations here (3) for valid tests. As before, it is possible to follow up the global ANOSIM tests and visualise effect sizes in ordinations of appropriately averaged data, here averages of the (standardised and square-root-transformed) replicate counts for the 16 sub-areas (Fig. 5a). This 2-d non-metric MDS is only an approximation to the true underlying relationships (stress is 0.11) but clearly demonstrates the rather strong and significant separation ($R_A = 0.71$) of the four locations (different symbols) relative to the assemblage differences between the two sites at each location (the matching pairs of 'dumbbells'). Whilst pairwise *tests* at the location level are not viable, it is not irrelevant to note that the lowest observed pairwise *R* value (= 0.25) is that between Berghan and Home Points, reflected in the overlap of these locations in the top left of the ordination. At the next spatial scale down, what is being compared visually is the separation of pairs of sites (within locations) with pairs of sub-areas within those sites (the 'dumbbell lengths'), giving $\overline{R}_B = 0.44$. It is thus no surprise from the plot that the null hypothesis here (no site

doi:10.1111/aec.13083

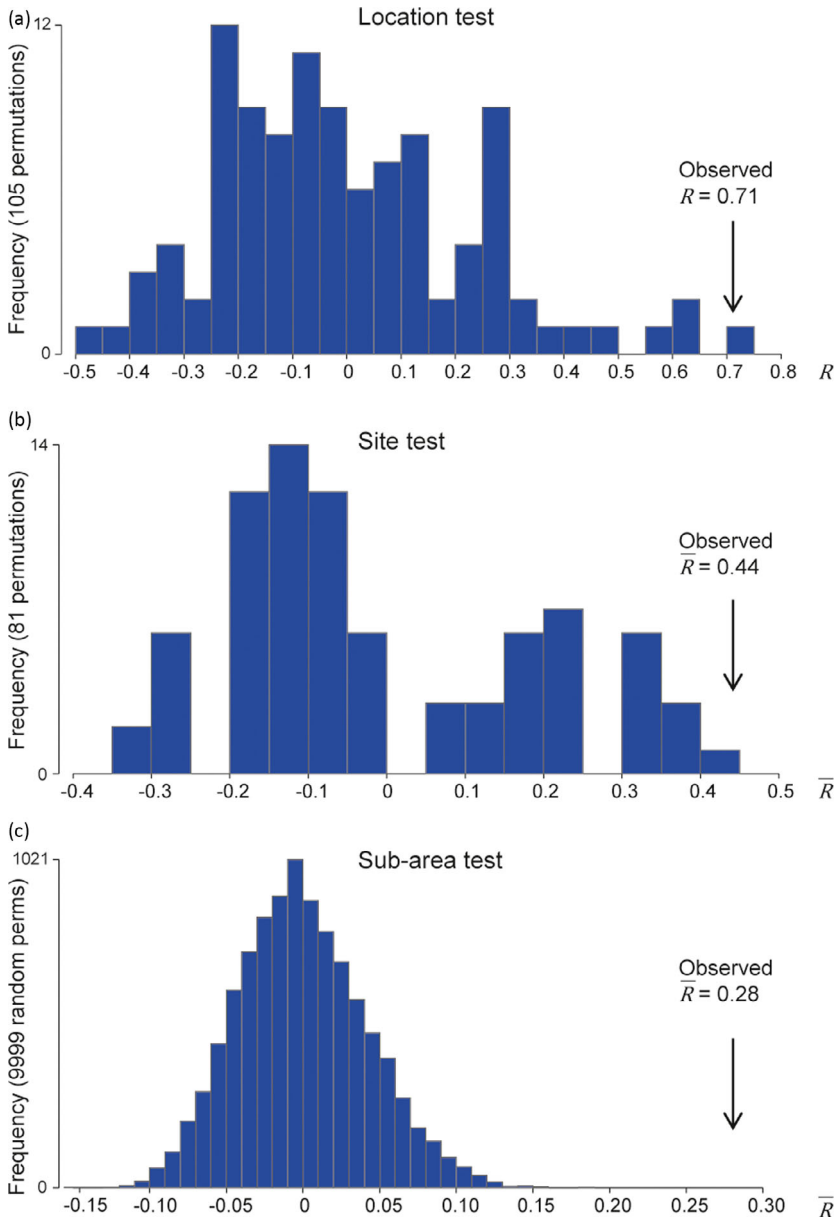

Fig. 4. Null distributions by permutation for 3-factor fully nested (unordered) ANOSIM tests for hierarchical differences in standardised square-root-transformed abundances of fauna inhabiting kelp holdfasts using Bray–Curtis similarity. The test is of C(B(A)), with five replicates from each of two sub-areas (C), nested in two sites (B) and nested in four locations (A). A very large number of permutations are possible for the lowest level test of sub-areas, so 9999 were selected at random; all permutations are computed for the site test (81) and the location test (105).

differences, within locations, over and above sub-area differences) has also been rejected by the ANOSIM test. However, visualising that R_A (= 0.71) appears larger than $\overline{R}_B$ (= 0.44) is more taxing ('symbol separation in relation to dumbbell separation, within symbols, is greater than dumbbell separation in relation to dumbbell length'). Visualisation of the next spatial scale down ($\overline{R}_C$ = 0.28) requires ordinations (not shown) for replicate holdfasts over sub-areas, separately by location in this case, otherwise the 2-d MDS plot has an unworkably high stress. It is clear therefore that the ANOSIM R values at the three scales, together with their tests, provide a very succinct and informative summary of the degree of spatial structuring at each level in this hierarchical design.

Tees Bay soft-sediment macrofauna

To recap, this is a B × C(A) design in which a total of eight Sites are nested in pairs (C levels: a,b; c,d; e, f; g,h) in four Areas (A levels: area 1–4) but crossed

doi:10.1111/aec.13083

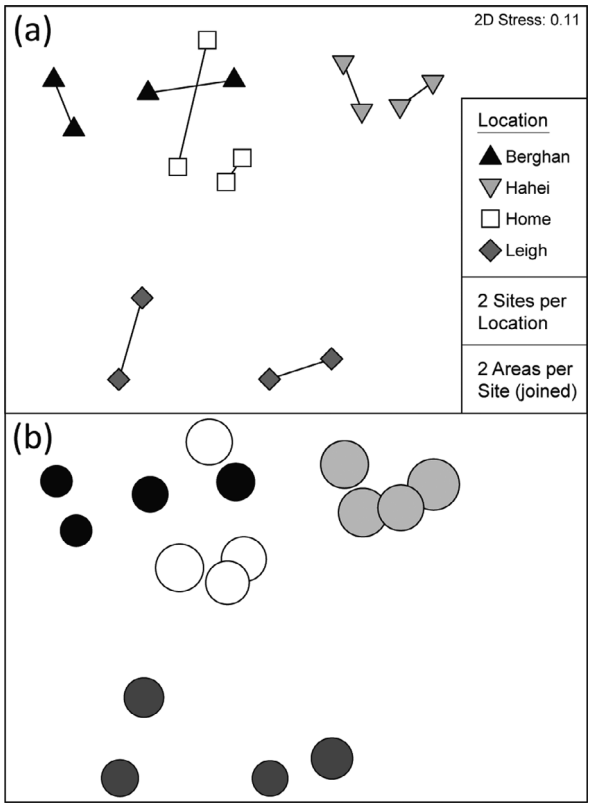

Fig. 5. (a) Ordination by nMDS of Bray–Curtis similarities calculated using square-rooted-transformed standardised abundances of 351 taxa, averaged over five replicate holdfasts in each sub-area (nested in site and location); (b) the same nMDS overlaid with bubbles scaled in diameter to reflect the average volume of holdfasts in each sub-area.

with 24 years (B levels: 73–96). The areas lie on a spatial transect (c. 5 km spacing, Fig. 2a) but are probably not ordered hydrodynamically, so it makes sense to consider both ordered and unordered tests for A (cases 3m/3j in Table 1). The years are also amenable to analysis under either alternative hypothesis. As it happens, there is a clear annual trend in assemblage structure over the period (seen in the nMDS plots of Fig. 2c–f, for the two sites in each area averaged), but the prior expectation might have been for a more complex time signal of cycles or short-term changes and reversions, so these data will serve to illustrate both alternatives for B (ordered or unordered, cases 3l/3j). There being only two sites in each area, it is then irrelevant whether C is considered to be ordered or not. With pseudo-replication at the sites, and thus only one genuine (pooled) replicate for each of only two sites in the four areas at each time, there can be no test for the site effect C (though with several sites in each area, returned to at each time, there would have been a test based on $\overline{R}^{Os}$ if the sites were considered ordered, or on an average of the matching statistic ρ_{av} if unordered, cases 3k/3j).

Figure 1a shows the construction of the ANOSIM permutation test for area (A), case 3m/3j (Table 1). The 1-way ANOSIM statistics R (or R^{Oc} if A is considered ordered) for a test of the four areas, using as replicates the two sites in each area, are computed separately for each year and then averaged over the 24 years, to obtain the overall test statistic for A of $\overline{R}$ (or $\overline{R}^{Oc}$). To generate the null hypothesis distribution for this test statistic the relevant permutation is to keep the columns of this schematic table intact and shuffle the eight whole columns randomly over the four areas, recalculating $\overline{R}$ (or $\overline{R}^{Oc}$) each time. There will be many fewer permutations for the A test under this B × C(A) design (8!/2!2!2!2!4! = 105 permutations for the unordered case, compared with 105^{24} for the standard A × B test which assumes that the sites are randomly drawn afresh each time from the defined area).

If area is considered an unordered factor $\overline{R} = 0.60$, a high value (and the most extreme of the 105 permutations, so p = 1%). This is clearly seen in the time-averaged nMDS plot for the eight sites (Fig. 2b). If treated as an ordered factor, the area test gives $\overline{R}^{Oc} =$ 0.13, now not even significant. These two values are directly comparable being the slopes of a linear regression of the types illustrated in Fig. 6, with the same y axis values but only two rather than four x-axis positions in the unordered case (within and among groups). The strong implication is that greater credence should be given to the unordered alternative hypothesis here, and the nMDS plot of sites in Fig. 2b makes clear the downside of an ordered test, based solely on geographical ordering of areas. The middle two areas are within the confines of Tees Bay (Fig. 2 a), with their assemblages potentially influenced by the hydrodynamics or even anthropogenic discharges from the Tees estuary. Thus areas 1 and 4 are rather similar to each other but differ from areas 2 and 3. Opting for what can be a more powerful test if there is a serial pattern risks failing to detect obvious differences when they are not serial, as illustrated (Fig. 6) for one of the 24 components of the average $\overline{R}$ and $\overline{R}^{Oc}$, namely the R and R^{Oc} constructions for 1978.

Turning to the test for the Year factor (B), case 3l/3j in Table 1, the schema for constructing the test statistic in both ordered and unordered cases was illustrated in Fig. 1b. Note that there are no replicates within levels of the time factor B at each site C. When years are considered ordered, the test reduces to the 2-way crossed layout B × C in which a 1-way ordered ANOSIM statistic without replicates (R^{Os}) is calculated over years, separately for each of the eight sites, and these values averaged to give $\overline{R}^{Os}$. The appropriate permutation is the usual one of samples in each site being randomly permuted across the

doi:10.1111/aec.13083

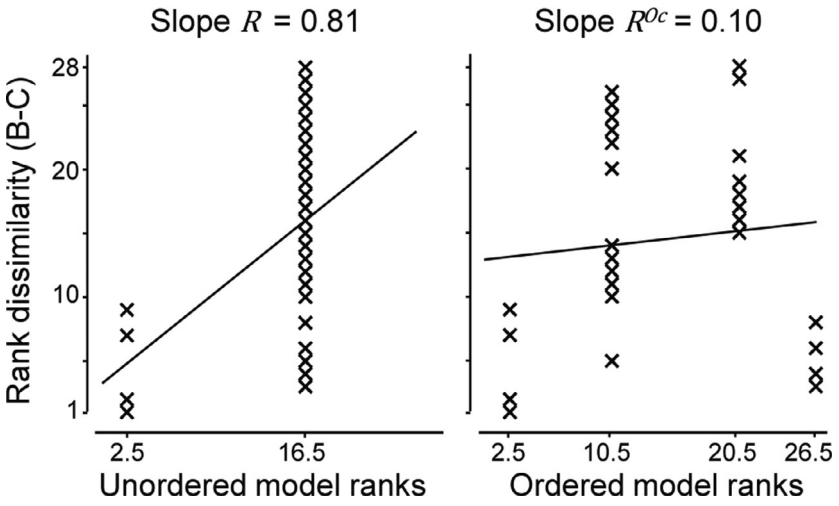

Fig. 6. Scatter plots of ranked Bray–Curtis similarities from observations of macrofauna from Tees Bay (at two sites within each of four areas, in 1978 only), against ranked distances in an unordered model (left, within group ranks and between group ranks only) and an ordered model (right, ranks within areas, between areas 1 step apart, 2 steps apart and 3 steps apart). The generalised ANOSIM statistic R^{Oc} is the slope of the linear regression.

years, since the null hypothesis specifies that there is no year effect, at any site. This will be roundly rejected, with global $\overline{R}^{Os} = 0.52$ in comparison with the permutation distribution, where none of 9999 (or even 99,999,999) randomly chosen permutations gave an $\overline{R}^{Os}$ value exceeding 0.1.

If it is considered unwise to test only for a serial trend through time, rather than a more general pattern of inter-annual changes, there is no replication which the test for B can exploit, so the design falls back on the indirect style of test (see Methods) in which evidence of differences among years is provided by observing significant similarities in temporal patterns across the spatially located sampling positions. A modified test statistic is needed here to cope with the structuring of the spatial factors into a 2-way nested design of sites within areas. As shown in Fig. 1b, a logical construction for the test statistic is to use the matching statistic ρ_{av} among the sites within each area (in this case there is only one ρ since there are only two sites) and then average these across the areas to give a doubly averaged $\overline{\rho}_{av}$ statistic. If site-to-site differences are temporally consistent within the areas, and there are genuinely no year-to-year changes within the sites, all the components of this statistic (and thus their mean $\overline{\rho}_{av}$) will be centred on zero, and the null hypothesis distribution to test $\overline{\rho}_{av}$ is created by the same permutations as for the ordered test. A significant effect of years may therefore be inferred from an observed consistency in the patterns through time across the (two) sites, in any or all of the areas. If (as might well be thought in this context) it is more appropriate to infer temporal change by noting commonality of inter-annual patterns at the wider spatial scale of areas, the sites should be averaged to leave a 2-way A × B design with both factors unordered, the B test then using the (singly averaged) ρ_{av} statistic. Here, $\overline{\rho}_{av} = 0.62$ (based on sites) and $\rho_{av} = 0.66$ (based on areas) are both highly significant, though note that unlike the variants of the ANOSIM R statistic their values cannot be compared with $\overline{R}^{Os}$ (= 0.52) for the ordered case, as the statistics are constructed very differently.

One other point should be noted for tests based on ρ_{av}. If there is a strong Year × Site interaction, that is there *are* inter-annual differences but these are entirely inconsistent across the sites within each area (or in the wider-scale test, inconsistent across the areas), then there will be no observed commonality of pattern and the test has no power to detect these annual changes. This parallels the situation in univariate ANOVA, or its multivariate PERMANOVA equivalent (Anderson *et al.* 2008), when there is no genuine replication within sites, as here. The (PERM)ANOVA table, for this unreplicated B × C (A) design, tests for Years by utilising the Year × Site interaction as its residual and it, likewise, cannot detect inter-annual change (i.e. an overall main effect of Years) if this is dominated by the inconsistency of temporal patterns over sites (i.e. a large Year × Site interaction mean square).

Returning to the ordered B test for temporal trend in Fig. 1b, doubly averaging R^{Os} by site then area could not actually change the previous $\overline{R}^{Os}$ value (0.52), though averaging sites first and performing the 2-way crossed test on Areas × Times (with no replication and ordered years) does increase the value slightly to $\overline{R}^{Os} = 0.60$. This statistic reflects the overall trend seen in the four time-series plots of Fig. 2c–f. It may also be of interest to ask whether the averaged $\overline{R}^{Os}$ hides a rather different trend for each area, and the individual trend values R^{Os} for each area (or site) can certainly be calculated and tested. The four areas here give reasonably consistent values of $R^{Os} = 0.67$, 0.54, 0.50, 0.67 respectively (all $p << 0.01\%$), though there is perhaps a suggestion here, and in the plots, that the

doi:10.1111/aec.13083

serial time trends are stronger in Areas 1 and 4 than they are in Areas 2 and 3, which lie within Tees Bay and are potentially influenced by outflows from the Tees River (Warwick *et al.* 2002).

DISCUSSION

Building on the work of Clarke (1993) and more recent developments such as the use of the generalised R statistic in 1-way (Somerfield *et al.* 2021a) and 2-way (Somerfield *et al.* 2021b) tests, the current paper demonstrates how, within the unified ANOSIM framework, fully non-parametric tests may be constructed for 3-way unordered and ordered designs, with or without replication (where a test is possible). The tests require no distributional assumptions. In this paper only an ordering consistent with a serial pattern of change is considered, but the definition and construction of the ordered R^{Oc} and R^{Os} statistics allow their calculation in any situation where a model describes the expected rank distances among sample groups under an alternative hypothesis (Somerfield *et al.* 2021a), so tests for cyclical (seasonal), spatial (based on distances among sample locations) and other patterns are entirely possible. If samples are genuinely ordered then using that ordering in constructing a test will give it more power (Somerfield *et al.* 2002, 2021a,b) but, as demonstrated with the Tees Bay 1978 analysis (Fig. 6), the converse is also true. If a genuinely unordered factor is analysed as though it were ordered, the test will lose power. On the subject of power, it is also worth noting that ordering may increase power in two ways. Firstly, using an ordered statistic R^{Oc} in a situation where inter-sample relationships genuinely are ordered will give a higher value than the alternative unordered R. Secondly, adding ordered categories also increases the number of possible permutations having the potential to give different values of the test statistic (Appendix to Somerfield *et al.* 2021b), which generally bears some relation to the power of the test. Primarily though, as both the 2-way designs of that paper, and the 3-way illustrations here demonstrate, it is the magnitudes of the ANOSIM statistics that provide information about relative effect sizes (accounting for other factors), which is often a main focus for a multifactorial design. In particular, the R statistics obtained by treating a factor as either ordered or unordered can be directly compared to ascertain the strength of evidence in the data to support these potential alternative hypotheses.

In the Western King Wrasse study (Lek *et al.* 2011) the original data had a 5-factor crossed design, treating region and habitat separately and with two further common labrid species studied, but such higher-way designs can always be analysed at a lower level, flattening pairs of factors, as described in this paper. In fact, Lek *et al.* (2011) analysed only three factors at a time to explore dietary change with region, habitat, species, size and season because there were no sheltered sites on the Perth coast, and not all labrid species and not all size classes were found in each location. Examining different hypotheses may often require separate analysis of different selections from a data set, especially in cases such as this, where several combinations of factors are not present. Received statistical wisdom is that all the data from a particular study should be analysed in a single procedure, to avoid biases from *a posteriori* selection, but this does not always reflect the reality of ecological fieldwork. Data are often collected for multiple purposes and designs are often asymmetric, as here, but an important step is to understand (*a priori*) how subsets of the data are to be extracted logically to tackle each major question.

In the nested New Zealand kelp fauna example, as Anderson *et al.* (2008) explain, the holdfasts had different volumes and, although here this is addressed by standardising all samples to relative composition, there may still be some artefactual dissimilarity arising from species-area relationships, that is higher species richness in larger holdfasts. One practical way to explore this is by superimposing, on an MDS ordination of the sub-areas (Fig. 5b), bubbles whose diameters are scaled to reflect the average volume of holdfasts in each sub-area. Indications of a substantial problem would be if the ordination pattern bore a clear relation to the differing volumes. For example, experience suggests that smaller samples with consequently lower species richness can (after standardisation) have higher Bray–Curtis inter-sample dissimilarities than among larger samples. This can result in smaller or sparser samples being seen as outliers on the MDS plot (Clarke *et al.* 2006; issues of the effects of unequal sample sizes on differing dissimilarity measures can be complex however, see Anderson *et al.* 2011). Here, Fig. 5b would not seem to indicate a significant issue, certainly after the averaging (post-transformation) over the five replicate holdfast assemblages represented by each of these ordination positions.

In the semi-parametric modelling context of PERMANOVA tests (Anderson 2001; Anderson *et al.* 2008), a more direct approach becomes possible, of attempting to remove the effects of covariates such as holdfast volume through linear regression in the high-dimensional 'dissimilarity space', before carrying out the tests on the main C(B(A)) factors (as Anderson *et al.* 2008 do for these data). Whether, in a specific case, the initial fitting of a linear covariate in the chosen dissimilarity space has, in fact, successfully removed its potential effect on the ensuing

doi:10.1111/aec.13083

analysis of the categorical factors can be uncertain. So-called 'linear models' are not constrained to be linear in the covariates, just in the parameterisation, so quadratic (and potentially higher-powered and/or interacting) covariate terms can be fitted. Unlike classical univariate ANOVA however, where simple model-checking plots are routine, it has been harder to visualise the success (or otherwise) of fitting such regression models, though the introduction (Anderson 2017) of the equivalent of univariate residual plots to this multivariate PERMANOVA context promises to be a useful step forward in improving model-checking capability for these more structured analyses.

Clearly the possibilities for anything similar in the non-parametric approach used here are very limited, though in simple 2-way designs there may sometimes be scope for coarse categorisation of a third, continuous, structuring variable into groups, then treated as another (ordered) crossed factor. Its effect would then be removed by the way the ANOSIM tests for the two primary factors are carried out (in parallel) solely within the strata of this third variable. It is also worth noting that one other significant feature of (PERM)ANOVA modelling, the distinction between fixed- and random-effects models, does have a parallel within the *nested* ANOSIM designs of this paper. Somerfield *et al.* (2021b) illustrate the way PERMANOVA and ANOSIM statistics are structured quite differently from each other in the 2-way A × B crossed design (and thus in the 3-way fully crossed A × B × C also). The 2-way (and thus the 3-way) nested designs, however, are much more congruent between ANOSIM and PERMANOVA. For the kelp holdfast communities, both test the changes at the location level, A, by taking the within-location replication level as that among sites, B(A), and testing at the site level uses within-site replication from sub-areas, C(B(A)). This is the classic random-effects model of (PERM)ANOVA and is not what the (almost universally erroneous) fixed-effects model would provide for this nested situation. Under fixed effect assumptions, the test of differences between sites, for example, would use the residual mean square as the denominator of the pseudo-*F* statistic, which measures variation among replicate *holdfasts*, and the conclusions could only infer site differences for the particular set of sub-areas from which those holdfasts are taken. That inference could *not* be extended to infer site differences for any conceived set of sub-areas from those sites which could have been sampled. For this reason, and in virtually every other situation in which nested factors are involved, such inference is achieved in PERMANOVA by specifying nested factors to be random, thus contrasting the mean square for sites to a residual which is the mean square for sub-areas. In ANOSIM, entirely equivalently, the contrast is of between-site average rank dissimilarities to within-site (between sub-areas) average ranks, and not to within sub-areas (between holdfast) ranks. Replicates, though not usually described as such, are simply random factors, permitting inference about the whole condition from which those replicates can be thought of as randomly drawn. Here both PERMANOVA and ANOSIM exploit the random-effects concept by using the same sequence of replicates: individual holdfasts to infer sub-area differences, sub-areas to infer site differences and sites to infer location differences.

Such equivalence of outlook for ANOSIM and PERMANOVA in nested models extends to some mixed models also, in respect of their nested factors, for example the Tees Bay design B × C(A), where years (B) are crossed with sites (C), which are nested in the four areas, A. Again, the inference for Area differences in the ANOSIM construction of Fig. 1a uses as 'replicates' the differences (in the whole time series) between the sites within each area. Site is conceived of as a random factor (nested in the fixed effect Area and crossed with the fixed effect Year) because the sites are randomly selected to represent their area – though then consistently returned to throughout the years, hence the retention of the integrity of each time series in the constrained permutations. The inference thus extends to establishing differences between those areas, not just differences between those *particular* sites in those areas. The concept of this test (though not the construction of the test statistic, naturally) is again entirely that of the PERMANOVA mixed model for these data, when Site is treated as a random factor nested in Area. This is also evident from the limited number of distinct permutations available for both the (correct) ANOSIM and PERMANOVA tests: $8!/[(2!)^4(4!)] = 105$, reflecting the comparison of four areas by two 'replicate' sites in each area (giving 3,4 d.f. for the PERMANOVA pseudo-*F*), and not the much greater number of permutations (and 3,92 d.f.) which would be obtained from an (inappropriate) PERMANOVA model in which the nested Site factor was treated as fixed.

Naturally, there are many multifactor models in PERMANOVA for which there are currently no robust ANOSIM equivalents. For example, as in classical univariate ANOVA, any factor in any given model (whether it be nested within or crossed with other factors) can be specified explicitly in PERMANOVA as being either fixed or random, in accordance with the appropriate design and the desired inference space for relevant hypotheses. As seen above, however, it would be wrong to assume that just because the robust non-parametric ANOSIM approach does not make use of the same explicit additive modelling structure, it cannot

doi:10.1111/aec.13083

produce tests which have equivalent generality in their inference space as the corresponding PERMANOVA test. For the suite of ANOSIM tests of Table 1 (and Table 1 of Somerfield *et al.* 2021b), it is important to appreciate that random effect assumptions are currently restricted to nested factors (indeed, are obligatory assumptions in those cases, in keeping with good design practice) but there is nothing which limits the scope of further potential tests involving random effects, and employing rank dissimilarities and ANOSIM-type statistics, to nested factors only.

Returning to the New Zealand kelp fauna, and the sequence of nested ANOSIM tests in the C(B(A)) design, there is a technical issue as to how best to combine the original replicates to provide 'sub-area replicates' for a test of site, and then how best to combine the sub-areas to provide 'site replicates' for a test of locations. There are many possibilities. PERMANOVA uses centroids calculated in its high-dimensional 'dissimilarity space' (see Anderson *et al.* 2008), whereas in the non-parametric approach described here the original resemblances are ranked, then averaged and re-ranked, at each level. Averaging the similarities rather than their ranks is another possibility, as is averaging the data (either transformed or untransformed). Only slight variations are likely to arise from the different choices, though experience suggests that averaging untransformed data is not optimal where a severe transformation is to be applied before computing dissimilarity. This is for much the same reasons as in univariate ANOVA where, if required, transformation is always needed first, to avoid distortion of computed means by outliers from strongly right-skewed distributions. One situation in which averaging on the untransformed scale may be considered appropriate is when the original replicates are sufficiently sparse and unreliable not to constitute a fair reflection of the assemblage structure at all (Clarke *et al.* 2014; Anderson & Santana-Garcon 2015). An example here is the use of multiple fish to make up single samples in the Western King Wrasse example. If in the nested case it is appropriate to pool samples (i.e. sum or average untransformed abundances) then a 3-way nested case could be analysed as 2-way nested for A and B(A) tests.

While use of pseudo-replicates (*sensu* Hurlbert 1984) in statistical testing is always to be avoided, repeated sampling, for example at one specific time and place, may still have an important role when those samples are pooled, in providing sufficient material for a sensible definition of a single replicate representing the community of that time and place. Though it is difficult to define power for any of the ANOSIM tests (Somerfield *et al.* 2002) it is important to ensure that sufficient replicates are taken at the right level of a multifactorial design to generate enough potentially distinct permutations (see Appendix to Somerfield *et al.* 2021b) for meaningful significance levels to be achievable. The balance of collection or analysis effort at different levels of a design is often context-dependent, and pilot experimentation will often reap dividends for the overall efficiency of a study. As a general rule, the aim should be to provide fully representative replication at the level immediately below the primary factor of greatest interest, and to use balanced crossed designs to eliminate non-negligible factors which are not the main focus of the study.

ACKNOWLEDGEMENTS

All those who contributed to collecting the data used in this study, or making it available, are gratefully acknowledged, including M.J. Anderson and colleagues, A.R. Brown and colleagues, R.M. Warwick and colleagues, and E. Lek, I.C. Potter and colleagues. In particular, we thank the Guest Editor and three anonymous reviewers, whose extraordinarily detailed and perceptive comments have greatly improved this paper.

CONFLICT OF INTEREST

The authors are unaware of any conflict of interest.

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).

DATA AVAILABILITY STATEMENT

All data are freely available from the corresponding author.

REFERENCES

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

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., Crist T. O., Chase J. M. *et al*. (2011) Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist. *Ecol. Lett.* **14,** 19–28.

Anderson M. J., Diebel C. E., Blom W. M. & Landers T. J. (2005) Consistency and variation in kelp holdfast assemblages: spatial patterns of biodiversity for the major phyla at different taxonomic resolutions. *J. Exp. Mar. Biol. Ecol.* **320,** 35–56.

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. & Santana-Garcon J. (2015) Measures of precision for dissimilarity-based multivariate analysis of ecological communities. *Ecol. Lett.* **18,** 66–73.

Clarke K. R. (1988) Detecting change in benthic community structure. *Proc XIVth Int Biometric Conf., July 1988, Namur: Invited Papers* I-6, pp. 13–24. 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. PRIMER-E, Plymouth.

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. & Chapman M. G. (2006) 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.

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.

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

Lek E., Fairclough D. V., Platell M. E., Clarke K. R., Tweedley J. R. & Potter I. C. (2011) To what extent are the dietary compositions of three abundant, co-occurring labrid species different and related to latitude, habitat, body size and season? *J. Fish Biol.* **78,** 1913–43.

Somerfield P. J., Clarke K. R. & Gorley R. N. (2021a) A generalised analysis of similarities (ANOSIM) statistic for designs with ordered factors. *Austral Ecol.* https://doi.org/10.1111/aec.13043.

Somerfield P. J., Clarke K. R. & Gorley R. N. (2021b) Analysis of similarities (ANOSIM) for 2-way layouts using a generalised ANOSIM statistic, with comparative notes on Permutational Multivariate Analysis of Variance (PERMANOVA). *Austral Ecol.* https://doi.org/10.1111/aec.13059.

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.

Warwick R. M., Ashman C. M., Brown A. R. *et al*. (2002) Inter-annual changes in the biodiversity and community structure of the macrobenthos in Tees Bay and the Tees estuary, UK, associated with local and regional environmental events. *Mar. Ecol. Progr. Ser.* **234,** 1–13.

doi:10.1111/aec.13083