Vol. 30: 283–286, 1985 | MARINE ECOLOGY – PROGRESS SERIES Mar. Ecol. Prog. Ser. | Published May 22

A new test for species sequencing

D. F. Sinclair

CSIRO, Division of Mathematics and Statistics, Private Mail Bag, PO Aitkenvale, Queensland 4814, Australia

ABSTRACT: A test is proposed to determine whether a presumed environmental gradient is accompanied by a change in species composition. A measure of asymmetry of the relation between each pair of species leads to the data being summarised as a skew-symmetric matrix. The column sums of this matrix reflect species sequencing. A Monte Carlo test procedure is described. The method is shown to be an improvement over a technique proposed recently in this journal (Bunt et al. 1985). Two sets of mangrove distribution data from northern Australian rivers are used for illustration.

INTRODUCTION

Zonal patterning of vegetation has received a good deal of attention in the ecological literature (see Pielou 1977, for a general discussion). A claim frequently made is that a presumed environmental gradient is accompanied by a progressive change in species composition. The reality of such species sequences was questioned by Bunt & Williams (1981); subsequently Bunt et al. (1985) suggested a test for the detection of species sequences, with particular application to mangrove species in northern Australia.

In this paper we critically examine the Bunt et al. (1985) test procedure and propose an alternative approach. Two sets of data on the occurrence of mangrove species from estuaries in northern Australia are used for illustration.

CURRENT TEST PROCEDURE

The Bunt et al. (1985) test for species sequencing is developed along the following lines. Suppose an estuary contains a total of n mangrove species whose sequential tendency it is desired to study. A series of t transects are laid out perpendicular to the estuary, each beginning at the water's edge and continuing inland to the mangrove limit (see Fig. 1). At a variable number of sites (roughly evenly spaced) along each transect the presence of any of the species of interest is recorded. Thus an incidence matrix recording presence and absence of each species at each site is formed for each transect. From these incidence matrices, n × n matrices A_k are formed, with entries a_{ij}^k = number of sites in Transect k at which Species i occurs *and* Species j occurs somewhere higher in the transect. This information is then combined in a matrix

$$A = \sum_{i=1}^{t} A_i.$$

Bunt et al. (1985) then form the skew-symmetric matrix $C = \frac{1}{2}(A - A')$, where A' is the transpose of A, with elements

$$c_{ij} = \frac{1}{2}(a_{ij} - a_{ji}).$$

C summarizes the magnitude and sign of the asymmetry in A. The entries c_{ij} are meant to reflect the relative positions of Species i and j along transects, a positive

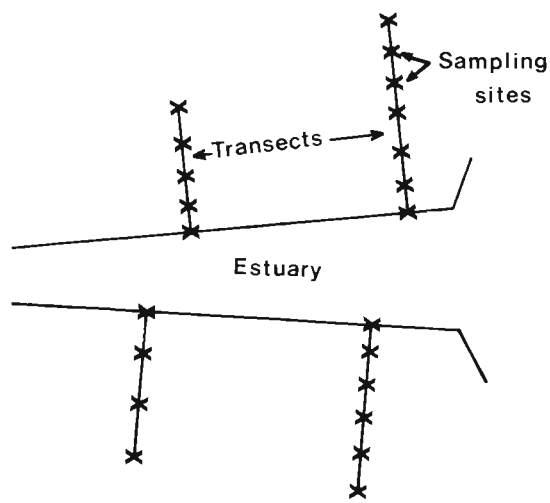

Fig. 1. Schematic representation of sampling procedure

0171-8630/86/0030/0283/$ 05.00

value indicating that Species i tends to follow Species j and a negative value indicating that Species j tends to follow Species i. While c_{ij} has intuitive appeal we demonstrate later that it is, in fact, biased.

Bunt et al. base their test for sequencing on the column sums of C, the ordering of which is felt to provide the underlying sequence. The proposed test statistic is the variance ratio of between-column mean square to within-column mean square, ignoring the n diagonal terms which are, by construction, zero. The authors acknowledge that the substantial structure in C strictly invalidates the use of standard significance levels for this 'pseudo' F-statistic, and can only draw on their limited experience with relatively few data sets to gauge the statistical significance of the test.

A VALID TEST PROCEDURE

We propose a Monte Carlo testing procedure (Barnard 1963) to determine the significance of the observed variance ratio test statistic, F_1, under the null hypothesis H_0 that the species are distributed randomly along transects. The test procedure is quite straightforward. We run $(m-1)$ simulations in which the observed number of each species in each transect is retained, but the occurrences are randomly allocated to the sites in the transect. Most large statistical computing packages have reliable random number generators which can be adapted for this purpose. For each simulation the variance ratio is calculated, giving us values $F_2, \ldots, F_m$. The corresponding order statistics are

$$F_{(1)} < F_{(2)} < \ldots < F_{(m)}.$$

Since under H_0,

$$P(F_1 = F_{(i)}) = 1/m,$$

where P denotes probability, the rank of F_1 is used to construct an exact test of H_0. If $F_1 = F_{(m-2)}$, say, then the attained significance level for the test is 100(3/m) %. For a test at the conventional 5 % level an adequate choice of m is 100 (Diggle 1983).

Monte Carlo tests provide a convenient method for testing hypotheses when the underlying null distribution of the test statistic is intractable. They have been used widely in the analysis of spatial patterns (e.g. Diggle 1983). The computer programming required is usually quite straightforward, although extensive simulation can consume considerable machine time.

The method is quite general, of course, and any other test statistic could be used. In the next section we propose an improvement to the Bunt et al. (1985) test statistic.

AN ALTERNATIVE TEST STATISTIC

The test statistic proposed by Bunt et al. (1985) is the variance ratio from an analysis of variance on the column entries of Matrix C, ignoring the zero diagonal entries. The heuristic argument behind this test statistic is that a positive column sum indicates a species which generally occurs higher up transects, and conversely for a negative column sum. Consider, however, the following simple situation of 2 species and 1 transect, with incidence matrix:

Site	Species	
	1	2
1	1	0
2	1	0
3	1	1
4	1	0
5	1	0

(1)

Clearly this is a situation in which one would want a measure to reflect a stalemate between the 2 species. However, the Bunt et al. measure indicates that Species 1 tends to follow Species 2 in the transect ($a_{12} = 2$, $a_{21} = 1$ and $c_{12} = \frac{1}{2}$, $c_{21} = -\frac{1}{2}$). This simple example illustrates the general bias of the method against frequently occurring species and in favour of rare species.

It would seem, then, that C does not provide an ideal reflection of species sequencing. If the column sums of C are to accurately reflect the sequencing of the species, then an alternative, unbiased measure is needed. The following variation seems an attractive alternative. Instead of incrementing a_{ij}^k by one each time an occurrence of Species i is followed somewhere in the transect by an occurrence of Species j, increment by the *number* of occurrence of Species j above each occurrence of Species i. Thus we define new matrices A_k with entries

a_{ij}^k = total number of occurrences of Species j following occurrences of Species i in Transect k,

$$\text{and } A = \sum_{i=1}^{t} A_i \text{ as before.}$$

Now consider a new C with entries $c_{ij} = (a_{ij} - a_{ji})/$ (total number of occurrences of Species i and Species j in all transects).

This measure avoids the bias problem of the Bunt et al. version. In our simple example (1), $a_{12} = (1 + 1)$ and $a_{21} = 2$, whence $c_{12} = 0$. The 'normalizing' denominator of this new c_{ij} maintains a sensible relativity in the deviations from symmetry in the relation between

Table 1. Incidence matrices for Norman River example (1 = presence; 0 = absence)

	Transect 1					Transect 2					Transect 3					Transect 4				
	Species					Species					Species					Species				
Site	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0
2	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	1	0	0	0	0
3	1	0	0	0	1	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0
4	1	0	0	0	1	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0
5	1	0	0	0	1	1	0	1	1	0	1	0	1	0	0	1	0	0	0	0
6	1	0	0	0	1	1	0	1	1	0	1	0	1	0	0					
7	1	0	0	0	1	1	0	1	0	0	1	0	0	0	0					
8	1	1	0	0	1	0	0	0	1	0	1	0	0	0	0					
9	1	0	0	0	1	1	0	0	1	0	1	0	0	0	0					
10	1	1	0	0	1						1	0	0	0	0					
11	1	0	1	0	0						1	0	0	0	0					
12	1	0	1	0	0															
13	1	1	0	0	0															
14	1	1	0	0	0															
15	1	0	1	0	0															
16	1	0	0	0	0															

Species i and j. For example, consider the 2 incident matrices:

Site	Species 1	2
1	1	1
2	1	0
3	1	0

Site	Species 1	2
1	1	1
2	1	1
3	1	0
4	1	0
5	1	0
6	1	0

(2)

These give $c_{12} = -\frac{2}{4}$ and $-\frac{6}{8}$ respectively. A ratio of 2 seems a reasonable measure of the relative degree of asymmetry in the 2 cases. Without the normalizing factor the ratio would be 4.

The proposed test procedure, then, is to use the variance ratio for the columns of this new Matrix C as the test statistic, and use Monte Carlo simulations to test for significance.

EXAMPLES

Bunt et al. (1985) illustrate their approach on the estuary of the Norman River in northern Australia. There were 5 species of interest (1 = *Avicennia* sp.; 2 = *Ceriops tagal*; 3 = *Excoecaria agallocha*; 4 = *Lumnitzera racemosa*; 5 = *Rhizophora stylosa*), and 4 transects. The incidence matrices are given in Table 1.

Bunt et al. calculate their resulting C matrix and obtain an observed variance ratio of $F_1 = 3.658$. They suggest that this provides evidence for 'a fairly strong sequential pattern'. However, when compared with 99 Monte Carlo simulations, the attained significance level was found to be 0.49. Thus there is no statistical evidence of sequencing in this data set using their test statistic. The Monte Carlo simulations give a 5 % critical value of 4.17, whereas the corresponding $F_{4,16}$ value is 3.01. If the F-distribution had been erroneously assumed, the observed test statistic would have been declared significant at the 5 % level, when in fact it is far from being so.

When our proposed unbiased procedure is applied to the Norman River data, we obtained the new Matrix C given in Table 2. Note that the bias in the Bunt et al.

Table 2. The new C matrix for the Norman River using the unbiased measure of sequencing

	Species				
Species	1	2	3	4	5
1	0.0	0.84	−0.38	0.0	−1.35
2	−0.84	0.0	0.57	0.0	−2.17
3	0.38	−0.57	0.0	0.46	−2.18
4	0.0	0.0	−0.46	0.0	0.0
5	1.35	2.17	2.18	0.0	0.0
Column sums:	.89	2.44	1.91	0.46	−5.70

Table 3. The *C* matrix for the Normanby River using the unbiased measure of sequencing

Species	Species 1	2	3	4	5	6	7	8	9
1	0.0	−0.32	−1.08	0.95	−0.17	0.23	0.16	−1.23	−0.01
2	0.32	0.0	−0.66	0.70	1.22	0.50	0.92	−0.32	0.05
3	1.08	0.66	0.0	1.29	1.21	0.81	1.20	0.02	−0.02
4	−0.95	−0.70	−1.29	0.0	−1.07	−0.04	−0.49	−1.89	0.0
5	0.17	−1.22	−1.21	1.07	0.0	−0.46	0.59	−0.90	0.02
6	−0.23	−0.50	−0.81	0.04	0.46	0.0	0.64	−0.36	−0.03
7	−0.16	−0.92	−1.20	0.49	−0.59	−0.64	0.0	−0.65	0.0
8	1.23	0.32	−0.02	1.89	0.90	0.36	0.65	0.0	0.02
9	0.01	−0.05	0.02	0.0	−0.02	0.03	0.0	−0.02	0.0
Column sums:	1.48	−2.74	−6.25	6.45	1.93	0.79	3.67	−5.34	0.03

measure which produced, for example, the pronounced negative column sum for the commonly occurring Species 1 (Table 1) is no longer present.

The accompanying F-value is 2.955. The attained significance level from 99 Monte Carlo simulations was found to be 0.53. Thus the conclusion using this more appealing test statistic is, again, that there is no evidence of sequencing.

The unbiased method has also been applied to a much larger data set, from the Normanby River in northern Australia. Here there were 9 species of interest (1 = *Avicennia* sp.; 2 = *Bruguiera gymnorhiza;* 3 = *Bruguiera parviflora;* 4 = *Ceriops tagal;* 5 = *Excoecaria agallocha;* 6 = *Heritiera littoralis;* 7 = *Lumnitzera racemosa;* 8 = *Rhizophora stylosa;* 9 = *Xylocarpus granatum*) and 16 transects, with an average of 14 sites per transect.

The *C* matrix for this example is given in Table 3. The resulting F-statistic is 5.99, which, when compared with 99 Monte Carlo simulations, attains a significance level of 0.03. Thus there is quite strong evidence of sequencing in this data. In particular, *Ceriops tagal* is exhibiting a preference to be higher up the transects, while *Bruguiera parviflora* and *Rhizophora stylosa* tend to occur closer to the water's edge.

Interestingly, despite the lack of statistical significance for the Norman River data, there is a good degree of consistency between the mangrove distributions of the 2 rivers. The 5 Norman River species were included in the Normanby River survey, and in both cases *Ceriops tagal* and *Rhizophora stylosa* occupied the 2 extremes.

CONCLUSION

This modification of the Bunt et al. (1985) measure of species sequencing along an environmental gradient removes the inherent bias of the original measure and seems to provide an accurate reflection of the relative positions of species. The Monte Carlo testing procedure allows a valid means of determining the significance of the pattern, and obviates the need to know the exact distribution of the test statistic.

The method would seem to be most useful in situations in which the environmental gradient is roughly constant. Caution would need to be heeded in interpreting the analysis when marked irregularities in the gradient are suspected.

Acknowledgements. I thank Bill Williams for bringing this problem to my attention, John Bunt for providing the data and Richard Sinclair for generous computing assistance.

LITERATURE CITED

Barnard, G. A. (1963). Contribution to the discussion of Professor Bartlett's paper. J. R. statist. Soc. B. 25: 294

Bunt, J. S., Williams, W T (1981). Vegetation relationships in the mangroves of tropical Australia. Mar. Ecol. Prog. Ser. 4: 349–359

Bunt, J. S., Williams, W T., Clay, J. (1985). The detection of species sequences across environmental gradients. Mar Ecol. Prog. Ser 24: 197–199

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

Pielou, E. C. (1977). Mathematical ecology. John Wiley & Sons, New York

This article was submitted to the editor; it was accepted for printing on February 27, 1986