Vol. 72: 283–287, 1991 | MARINE ECOLOGY PROGRESS SERIES
Mar. Ecol. Prog. Ser. | Published June 4

Yet another method of species-sequencing

W. T. Williams[1], J. S. Bunt[2], H. J. Clay[3,*]

[1] 10 Surrey Street, Townsville, Queensland 4812, Australia
[2] 4/6 McDonald Street, Potts Point, N.S.W. 2011, Australia
[3] Australian Institute of Marine Science, PMB No. 3, Townsville, Queensland 4810, Australia

ABSTRACT: Following a paper by Bunt et al. (1985) describing a numerical method for revealing mangrove species sequences along intertidal gradients, Sinclair (1986) drew attention to a source of bias in the procedure and offered a means to remove it. In exploring the matter further, it has now been recognized that neither method takes effective account of species which are either ubiquitous or distributed bimodally along particular transects. A new, and simpler, method is now described which pays attention to such problems and, at the same time, appears to avoid bias. The question of whether sequential measures of this kind may be subjected to tests of statistical significance is also considered.

HISTORICAL INTRODUCTION

In the first of what now becomes a series of 3 papers the present authors (Bunt et al. 1985) outlined a numerical model for objectively establishing the sequence of mangrove species along a transect from the water's edge to inland. This uses a transition-matrix approach. An empty matrix A, with general term a_{ij}, is first set up; we assume the transect consists of n sites, approximately equally-spaced. Beginning at the first site, nearest the water's edge, consider one of its species, i. Scan through Sites 2 to n and note whether a species other than i, say j, occurs in any of the higher-numbered sites. If j occurs anywhere in the higher-numbered sites, a_{ij} is incremented by 1; the *number* of times j occurs is disregarded. The process is repeated for all other species in Site 1. Now pass to Site 2 and repeat the process; continue in this way until Site $(n-1)$, the last site for which the procedure is possible. A is now a transition-matrix using the convention 'from rows to columns', in that every element represents the number of times the species defined by the column occurs higher in the sequence than that defined by the row.

A will normally be asymmetric; it is transformed into a skew-symmetric matrix, C, by the conventional transformation $c_{ij} = 1/2(a_{ij} - a_{ji})$. The column sums of C represent the sequence required, the highest negative sum being that of the species nearest the water's edge, the highest positive sum the species furthest inland. The method was used on a small set of 5 species from 4 transects, and appeared at first sight to give plausible results.

In the second paper Sinclair (1986) points out that the test-statistic of Bunt et al. (1985) is biassed against frequently-occurring species, in the sense that such a species will generate an unrealistically high negative column sum in the C matrix, and so will appear to be closer to the water's edge than is justified. He suggests, as an alternative, incrementing the A matrix for the *number* of occurrences of Species j above each occurrence of Species i. He then forms the C matrix as before, but standardizes each c_{ij} by dividing by the total number of occurrences of species i and j. We shall show that this procedure undoubtedly alleviates, and perhaps removes, the bias; but that there remains an underlying logical problem that both sets of authors have overlooked.

Finally in this section we note that in both the previous papers the possibility of estimating statistical significance is addressed; but we defer consideration of this aspect until the final section of this paper.

TWO FORGOTTEN PROBLEMS: UBIQUITY AND BIMODALITY

Consider a species which is in every, or almost every, site in the system, a situation we describe as *ubiquity*. Inspection of the incidence matrix for the Norman

* With a mathematical appendix by J. F. Hunter, Department of Mathematics, James Cook University, Townsville, Queensland 4811, Australia

0171-8630/91/0072/0283/$ 03.00

transects given by Sinclair (1986) shows that this situation is realized by *Avicennia* sp.: it occurs in every site of Transects 1, 3 and 4, and in 7 of the 9 sites of Transect 2. As Sinclair pointed out, the result of using the original Bunt et al. measure is that *Avicennia* has a very high negative column sum in the *C* matrix, and consequently appears to be the species closest to the water's edge – a clearly unacceptable result. Using Sinclair's improved statistic it appears to be approximately central along the transect – a result perhaps less unacceptable, but still obscuring the true situation. We maintain that, if only a single transect is being considered, a ubiquitous species cannot be part of a sequence; it is so much continuous 'background noise', and can be ignored. This is no longer true if an entire river-system is under examination. It is possible for a species such as *Avicennia* to be ubiquitous in the transects near the river-mouth, but to be no longer so in the higher reaches. In such a case, the distribution of ubiquity is itself an important feature of the species' distribution, and must be recognized and recorded.

A less common situation is that of a species which occurs in a few sites near the water's edge, is not encountered along the main sweep of the transect, but occurs again in a few of the extreme inland sites. We have even encountered a case of a species which occurred only in the first and last sites. This situation we describe as *bimodality*. Our examination of the data at our disposal is still in progress. Nevertheless, we have the impression that several species, notably *Aegialitis annulata, Bruguiera exaristata, Aegiceras corniculatum* and, as noted by Macnae (1966), *Avicennia* sp., all tend to be distributed along transects in this fashion. However, we again believe that a species exhibiting this behaviour cannot, in a single transect, be considered as part of a simple linear sequence; but again, within a complete river-system, it is a potentially important feature of a species' distribution. As with ubiquity, it needs to be recognized and recorded.

We now define a new procedure for determining species sequences, in the course of which ubiquitous and bimodal species can be easily recognized.

THE NEW METHOD

In this method we dispense completely with the transition-matrix concept. Instead, on any one transect, we work species-by-species and estimate the position along the transect at which each species is most likely to be found. The method can most easily be described by means of a simple worked example. We have chosen a single transect near the mouth of the Norman river, in North Queensland; the transect comprised 16 sites, numbered serially from 1, at the water's edge, to 16, the furthest inland. Seven species occurred in this transect; we shall refer to them by the numbers used in the data-base already prepared by Bunt (1982), since this will facilitate later comparison with other rivers. With code numbers, they are as follows:

3: *Aegialitis annulata* R. Br.
4: *Aegiceras corniculatum* (L.) Blanco
5: *Avicennia* sp.
7: *Bruguiera exaristata* Ding Hou
11: *Ceriops tagal* var. *australis* C. T. White
14: *Excoecaria agallocha* L.
24: *Rhizophora stylosa* Griff.

Only presence was recorded; Table 1(a) shows the species present in each site, set out in the format used in the Bunt (1982) data-base. For each species, every occurrence is given a score, which is simply the serial number of the site in which it occurs, less one. The reason for the 'less one' is to provide the system with a meaningful origin, so that '0' represents the water's edge. The resulting table of scores is given in Table 1(b).

It is immediately obvious that Species 5 is ubiquitous and Species 3 is bimodal. Species 11 (*Ceriops tagal* var. *australis*) and 14 (*Excoecaria agallocha*) both have poorly defined distributions, a matter of relative rarity along the transect, but are not bimodal. For routine data analysis, we accept as bimodal situations those in which the species in question is missing from at least 4 sites consecutively along a transect.

If there are n sites in the transect, the maximum possible value of the range is $(n-1)$; any species whose range equals, or even closely approaches, this value is likely to be ubiquitous or bimodal, and requires further

Table 1 Mangrove occurrences along a transect in the Norman River, North Queensland with scoring as described in text

Site no.	(a) Raw data Species occurring	(b) Species scores Species code 3	4	5	7	11	14	24
1	3/5	0	–	0	–	–	–	–
2	3/5	1	–	1	–	–	–	–
3	3/4/5/24/	2	2	2	–	–	–	2
4	3/4/5/24	3	3	3	–	–	–	3
5	3/4/5/6/24	4	4	4	4	–	–	4
6	3/4/5/7/24	5	5	5	5	–	–	5
7	7/24	–	–	–	6	–	–	6
8	5/7/11/24	–	–	7	7	7	–	7
9	5/7/24	–	–	8	8	–	–	8
10	5/7/11/24	–	–	9	9	9	–	9
11	3/5/7/14	10	–	10	10	–	10	–
12	3/5/7/14	11	–	11	11	–	11	–
13	3/5/7/11	12	–	12	12	12	–	–
14	3/5/11	13	–	13	–	13	–	–
15	5/14	–	–	14	–	–	14	–
16	5	–	–	15	–	–	–	–

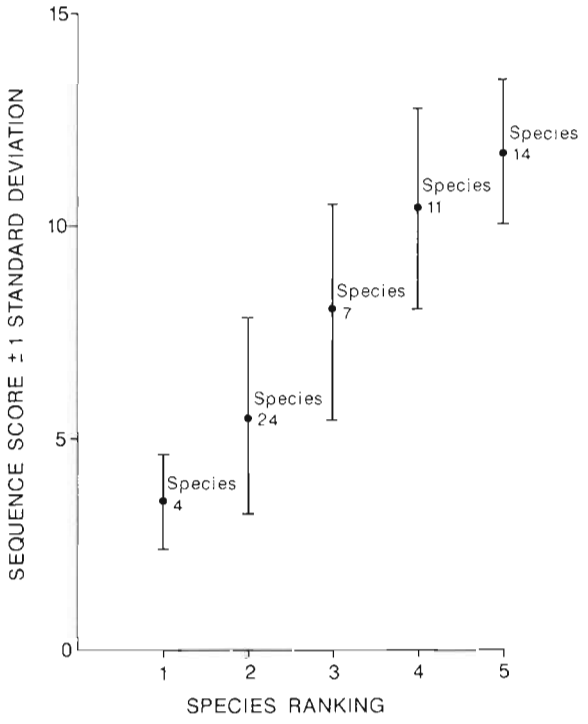

Fig. 1. Diagrammatic representation of the sequence of species along a transect in the Norman River, North Queensland. Scoring values ± 1 standard deviation procedure described in text. Ubiquitous Species 5 and bimodal Species 3 not shown

examination. Since in the present exercise we are considering only a single transect we can, for the reasons given above, set aside both Species 3 and 5.

For each species remaining in the system we calculate the mean and standard deviation of the scores in Table 1(b). The results can form the basis of the elegant graph of Fig. 1, in which the fiducial limits represent ± one standard deviation. The nature of the sequence is now manifestly *Aegiceras corniculation* – *Rhizophora stylosa* – *Bruguiera exaristata* – *Ceriops tagal* – *Excoecaria agallocha*, with *Avicennia* sp. ubiquitous and *Aegialitis annulata* bimodal, as already noted. In such a simple situation it may be seen readily that the derived pattern is compatible with the raw data. In more complicated cases, however, determination of a sequence from an examination of the primary data would not be feasible and, were it attempted, would be highly subjective.

THE CONSOLIDATION PROBLEM

We assume a computer program which works transect by transect (see next section); the problem is to consolidate the results to give a generalized sequence for an entire estuary. The species-score means are not additive unless every transect in the estuary has the same number of sites, which is unusual. However, if there are n sites in a transect, the highest possible mean score is $(n-1)$, attained by a species with a single record in the highest-numbered site. Consequently, if the species-means in any one transect are divided by $(n-1)$, all are constrained between 0 and 1; the means are then additive, and can themselves be meaned if desired. The same is true of the ranges. It is obviously desirable that any computer program should be able to provide this form of standardization.

There remains the problem of the overall analysis of a complete estuary. If every species anywhere represented in the estuary was somewhere present in every transect, there would be no difficulty. The sums of squares of the species-scores could then be partitioned into the components due to species, transects and their interaction; the interaction would provide some estimate of the extent to which sequences changed as one moved further up the estuary. However, in practice we should expect an appreciable number of missing values. The method of handling missing values in an analysis of variance situation is still to some extent controversial, and some degree of approximation will be inevitable. We shall return to this problem in a later paper dealing with a complete river-system.

COMPUTATIONAL CONSIDERATIONS

A computer program has been written by one of us (H.J.C.) in Turbo PASCAL which implements the procedure. We here give only a brief outline; any reader desiring a detailed specification, or a copy of the program, should write direct to Mr H. John Clay at the address given in the article heading. For each transect the program accepts a set of data in the format of Table 1(a). It detects, and prints out, n, the number of sites in the transect; it then prints out the complete list of species-numbers involved, and a table of scores in the format of Table 1(b). For each species it calculates the range, mean, and standard deviation of the scores; these are printed out both as calculated from the raw data, and as standardized by division by $(n-1)$. It preserves the raw-data means and, at the end of the transect, prints them out, together with their species reference-numbers, in ascending order.

At the end of a complete river the program also prints out a 'co-occurrence matrix'; this summarizes, for all possible species-pair, the number of times each of the members of a pair occurs together in the same site. As a model of community-structure we admit that this is exiguous, but we believe it provides useful information. For a single transect, such as the Norman, such a table is of little use and we do not present it.

CONSIDERATIONS OF STATISTICAL SIGNIFICANCE

The question as to whether sequential measures such as those under discussion should be subjected to tests of statistical significance is itself somewhat controversial. We return to this aspect at the end of the present section; meanwhile, we are obliged to point out that no less than 3 separate significance tests have been proposed. These are as follows:

Method 1 (Bunt et al. 1985). This method is admittedly crude. It regarded the *C* matrix column sums as equivalent to the treatment sums for a single variable, and undertook a between/within single-factor analysis of variance; the degrees of freedom were adjusted for the fact that the principal diagonal of a skew-symmetric matrix is by definition everywhere zero. Sinclair (1986) pointed out that the distribution of the elements of a skew-symmetric matrix is quite unknown, and almost certainly non-normal. Consequently there would be a real danger that the *F*-ratios would be grossly overestimated, producing a spurious degree of significance. On re-examining the same data by his own improved Method 2 (below) he claimed that this was indeed the case.

A Method 1 test applied to the score matrix of the new method would at least be more plausible, since the distribution of the elements of the score matrix is known. If we ignore the origin-adjustment the r individual score values for a single species represent a set of r integers selected without replacement from the set of integers 1 to n; but though the distribution is known, it is not normal, and the Sinclair criticism still applies.

Method 2 (Sinclair 1986). This involves a Monte Carlo simulation; the null hypothesis is that the species are distributed randomly along transects. A set of $(m-1)$ simulations is run in which the observed number of each species in each transect is retained, but the occurrences are randomly allocated along the transect. For each simulation a variance ratio is calculated as in Method 1. For the method of converting the results to an exact test the reader is referred to Sinclair's paper· he states that a value of $m = 100$ is adequate for a 5% test. It must be admitted that such a method is computationally far more time-consuming than is the basic analysis itself.

Method 3 (novel). As pointed out above, if we ignore the origin-adjustment, the set of r scores for a single species in a single transect represents a sample of r integers selected without replacement from the integers 1 to n. The random expectation for the mean and standard error of such a sample could presumably be calculated; and at first sight it might appear that these could provide the basis for an attractive significance test. The necessary calculation is by no means trivial, and the solution is given in Mr Hunter's Appendix.

The result may be somewhat surprising to a non-statistician: the random expectation of the mean is *always* $1/2(n + 1)$, irrespective both of the value of r and of the part, or parts, of the sequence from which its items are drawn. Although the standard error – which is not independent of r – is also now known, it could only be used to test the deviation of a given sample mean from the mid-value $1/2(n + 1)$. This is of no use in the present context; interesting though the concept may have been, biologically it represents a rigorous statistical answer to the wrong question.

It is evident that only Sinclair's Method 2 provides a rigorous significance test for the entire sequence. Nevertheless, we cannot now evade the question of whether any form of significance test is necessary or even desirable. The situation is not unlike that for intrinsic classification, in which in only a very few highly specialized cases is a significance test even possible. Ultimately, we are seeking a pattern in a complex set of data; as Mackay (1969) has pointed out, such a pattern is never unique – there is only 'pattern-for-an-agent', a pattern which the agent can interpret, and finds helpful for further speculation or possible more rigorous test. We incline to the view that the search for pattern in mangrove sequences is similar to this, both in concept and intent.

GENERAL CONCLUSIONS

Mangrove species zonation is northern Australia is often complex and frequently differs markedly from the simple pattern considered by Macnae (1966) to be characteristic for the region. Critical or detailed examination of the vegetational variability within and between estuaries demands an objective means of identifying species sequences along intertidal transects. The procedure now described appears to overcome earlier difficulties experienced in satisfying the essential requirements and we believe, is especially useful in handling floristically complicated data sets. A paper using this approach to consider underlying patterns of mangrove zonation at a number of locations in northern Australia is in preparation in elaboration of a more general account by Bunt & Williams (1981).

APPENDIX

J. F. Hunter

Problem: r integers are selected randomly from the integers 1, 2 n without replacement. We require the expected value of the mean $\bar{y}$ of the sample and its standard error. It will be convenient first to consider the sum of the r integers, denoted by S_r.

Solution: For each integer i ($i = 1, 2 \ldots n$) define an indicator variable Θ_i such that $\Theta_i = 1$ if the integer i is in the sample, $= 0$ if it is not; denoting probability by p, and assuming that each integer has the same chance of being selected, we have:

$p(\Theta_i = 1) = r/n$; $p(\Theta_i = 0) = 1 - r/n$; and $p(\Theta_i^2 = 1)$ also $= r/n$

The Θ_i values are not independent, but the joint probability of $(\Theta_i = 1, \Theta_j = 1)$ can be calculated from

$$p(\Theta_i = 1, \Theta_j = 1) = \frac{\text{no. of ways of selecting } r-2 \text{ integers from } n-2}{\text{no. of ways of selecting } r \text{ integers from } n}$$

because, in the numerator, 2 integers (i and j) are already selected.

Therefore
$p(\Theta_i = 1, \Theta_j = 1) = {}^{n-2}C_{r-2}/{}^{n}C_r = r(r-1)/n(n-1)$

The quantity of interest, S_r, can be written as

$$S_r = \sum_{i=1}^{n} \Theta_i . i$$

$$\text{Now, } ES_r = \sum_{i=1}^{n} E\Theta_i . i = \sum_{i=1}^{n} \frac{r}{n} . i$$

$$= \frac{1}{2} r (n+1)$$

$$\text{Also } ES_r^2 = E\left[\left(\sum_{i=1}^{n} \Theta_i . i \right) \left(\sum_{j=1}^{n} \Theta_j . j \right) \right] =$$

$$\sum_{j \neq i}^{n} ijE\Theta_i \Theta_j + \sum_{i=1}^{n} i^2 E\Theta_i^2$$

$$= \sum_{j \neq i}^{n} i . j \frac{r(r-1)}{n(n-1)} + \sum_{i=1}^{n} i^2 \frac{r}{n}$$

$$= \frac{1}{12} \frac{r(n+1)}{n(n-1)} \{3rn^2 + n^2 - rn - 2r - n\}$$

But var $S_r = ES_r^2 - (ES_r)^2$

$$= \frac{1}{12} r (n-r) (n+1)$$

Hence $E\bar{y} = ES_r/r = \frac{1}{2}(n+1)$

and its standard error is $\sqrt{\frac{(n-r)(n+1)}{12r}}$

Acknowledgements. The authors are grateful to the Australian Institute of Marine Science for typing the manuscript.

LITERATURE CITED

Bunt, J. S. (1982). Mangrove transect data from northern Queensland. Australian Institute of Marine Science Data Report, Townsville. AIMS-CS-82-1: 1–41

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

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

Mackay, D. (1969). Recognition and action. In: Watanabe, S. (ed.) Methodologies of pattern recognition. Academic Press, London, p. 409–416.

Macnae, W. (1966). Mangroves in eastern and southern Australia. Aust. J. Bot. 14: 67–104

Sinclair, D. F. (1986). A new test of species sequencing. Mar. Ecol. Prog. Ser 30: 283–286

This article was submitted to the editor

Manuscript first received: August 21, 1990
Revised version accepted: February 5, 1991