Journal of Vegetation Science 4: 341-348, 1993

Comparing three classification strategies for use in ecology

Belbin, Lee[1]* & McDonald, Cam[2]

[1]*CSIRO Division of Wildlife & Ecology, P.O. Box 84, Lyneham, A.C.T., Australia, 2602;*
[2]*CSIRO Division of Tropical Crops & Pastures, 306 Carmody Rd, St. Lucia, Qld., Australia, 4067;*
**Fax +61 6 2413343*

Abstract. We compare three common types of clustering algorithms for use with community data. TWINSPAN is divisive hierarchical, flexible-UPGMA is agglomerative and hierarchical, and ALOC is non-hierarchical. A balanced design six-factor model was used to generate 480 data sets of known characteristics. Recovery of the embedded clusters suggests that both flexible UPGMA and ALOC are significantly better than TWINSPAN. No significant difference existed between flexible UPGMA and ALOC.

Keywords: Agglomerative strategy; ALOC; Clustering algorithm; Divisive strategy; TWINSPAN; UPGMA.

Introduction

Classification and ordination provide fundamental tools for the exploratory analysis of ecological data. Such pattern finding techniques are usefully employed at an early stage to detect errors, suggest trends, identify outliers and generally provide a succinct summary of multivariate data. The problem with such methods is however the selection of appropriate strategies from an increasing range of choices. Ordination could be considered to be a more powerful class of pattern analysis techniques than classification. Ordination can better detect gradients and the nature of clusters, but is computationally more demanding and more likely to mislead. For example, the inclusion of an outlier will have a significant effect on all ordination methods but no side-effects on clustering.

The need for efficient classification algorithms for ecological data is undeniable. Systematic evaluations of alternatives in ecology are however extremely scarce. Milligan (1980, 1989), Milligan & Cooper (1986, 1987) provide a number of useful comparisons of clustering algorithms using 'non-ecological' data. These studies provided evidence of the utility of techniques such as UPGMA and non-hierarchical methods such as k-means (McQueen 1967). The possible restriction in applying these results generally in ecology is that the 'data-generator' used assumed groups that were multivariate normal (see Milligan 1985). This assumption is unlikely in ecology. Curtis & McIntosh (1951), Goodall (1954), Bray & Curtis (1957), Dagnelie (1960), van Groenewoud (1976), Whittaker (1978), Austin (1985) and Austin & Smith (1989) among others, suggest that the distribution of species form a continuum along environmental gradients. If this assumption is reasonable, well-defined clusters ('natural-groups') may not exist (Belbin 1992). One implication is that between any pair of samples, there exists in theory, a sample of intermediate species composition. Clusters could therefore be considered the result of inadequate sampling.

The problem in evaluating different clustering strategies for use in ecology is to provide data that is 'ecological' in its characteristics and where the 'structure' of the data is known. It is not difficult to find ecological data sets. Neither is it difficult to construct artificial data where all characteristics are perfectly known. The problem is in the marriage; to construct data sets that have an ecological character and where the structure is known. In classification, cluster membership defines the structure. Jain & Dubes (1991) provide a variety of definitions of what can constitute a cluster. Our definition in this study is reasonably broad-minded as the section on data generation should infer.

This paper uses the COMPAS algorithm of Minchin (1987) to simulate 'ecological data'. We evaluated the extent of the recovery of the known clusters by three types of classification algorithms. The methods evaluated were ALOC (Belbin 1987), TWINSPAN (Hill 1979a), and a flexible variant of UPGMA (see Sneath & Sokal 1973) by Belbin, Faith & Milligan (in press). These methods were selected because they are supported examples of the major classification alternatives (non-hierarchical, hierarchical divisive and hierarchical agglomerative).

Test data

The simulated data contained a predetermined number of clusters in a two-dimensional (Euclidean)

'environmental-space'. Gauch & Whittaker (1981) used a similar method to evaluate classification methods. Our first step used the RAND algorithm (Belbin 1991) to generate 10 sample configurations (named 0-9) of 80 sample sites in a two-dimensional environmental space. The first column contains uniform random values (0 - 100) representing co-ordinates on the first dimension. Similarly, values in the second column corresponded to co-ordinates on the second dimension and ranged from 0 - 80. Each dataset used a different random number seed.

Each of the 10 data sets contained either 2, 3, 4, 5 or 8 clusters. Flexible-UPGMA and ALOC can extract a given number of clusters but TWINSPAN tends to produce 2, 4, 8, 16, 32, 64 … 2^n clusters where n is the number of dichotomies. TWINSPAN can produce an odd number of clusters if a group is too small to split or outliers exist. For this reason, we assisted TWINSPAN by generating two clusters plus an outlier for the 3-cluster cases and four clusters plus an outlier for the 5-cluster cases.

Eliminating 30 of the 80 sample points in the environmental space for all 10 sample configurations produced the 'true-clusters'. Such clusters were readily identifiable by eye, so they were well-defined, yet they had few other regular features. For example, there was no regularity in cluster outline, shape, size or density. Classifying the 50 sites using Euclidean distance on the environmental co-ordinates and UPGMA ($\beta = 0$) verified the identity of the clusters. Note that the conclusions drawn from this study would be as equivocal as the cluster definitions. If the clusters were too 'fuzzy', conclusions would be equivocal. On the other hand, extreme cluster separation and regular boundaries are unrealistic.

The second step in the data simulation involved the use of COMPAS (Minchin 1987) to generate the species response surfaces on the two gradients. Beta functions (Austin 1976; Minchin 1987) provided a method of generating the abundance of the species in the environmental space. With these functions, the alpha and gamma values determine the shape of the surface. In this study, alpha and gamma varied (independently) from 0.5 to 4.0. This permitted a considerable variation in curve shape from symmetric to skewed with the latter predominating. We constructed six sets of species definitions. This procedure generates a multidimensional 'species-space' with each species abundance (0 to a potential maximum of 100) represented as one axis. The mean species widths and standard deviations for the long gradient models were 66/33 (primary axis) and 100/50 (secondary axis) respectively. For the short gradient, the associated parameters were 100/50 and 120/60 respectively. It is important to realise that the clustering algorithms operate directly within this species-space and only indirectly with the underlying two-dimensional environmental space.

Van Groenewoud (1992) concluded that correspondence analysis (Hill 1974), detrended correspondence analysis (Hill 1979b) and TWINSPAN (Hill 1979a) failed badly when secondary gradients approached primary gradients in length. In this study the ratio of the second to the first gradients for the long and the short models was 0.66 and 0.83 respectively. We believed that this strategy would enable us to assess van Groenewoud's (1992) conclusions.

The third data generation step used COMPAS to amalgamate the site and species definitions into 480 data sets. To enhance the ecological character of the data, we added a number of factors that were either 'off' or 'on'. The first factor modified the abundance of the species to introduce the notion of a varying carrying capacity across the environmental space. This is equivalent to a site standardisation, for example, dividing abundance values in one site by the total site abundance. The second factor also modified the abundance of the species, this time to introduce an element of 'competition'. The application of this factor permitted 'shoulders' and multi-modalities in the species response surfaces. The last factor introduced 'noise'; modifying the abundance of a species by a factor proportional to the square root (Minchin 1987). These amendments to the site and species specifications are additional to the procedures used to generate test data by van Groenewoud (1992).

The COMPAS program of Minchin (1987) represents a comprehensive and sophisticated tool in simulating 'ecological data'. We feel that any shortfalls in the production of the data sets used here reside with the authors and not COMPAS. Limitations in 'quantifying nature' will remain for some time to come. There is no doubt that more realistic generators of 'ecological-data' will be forthcoming as statistical models of species distributions based on environmental data are established, for example Whittaker (1956), Whittaker & Niering (1965), Westerman (1975, 1991), Austin (1976), Austin et al. (1983, 1989).

Clustering methods

The agglomerative algorithms have been by far the most popular in ecology as well as most other disciplines. Flexible-UPGMA (Belbin, Faith & Milligan in press) is based on UPGMA (see Sneath & Sokal 1973). The divisive methods are attractive from a theoretical point of view but have not, until TWINSPAN (Hill 1979a) had a large following. ALOC provides an example of a

non-hierarchical algorithm that can accommodate tens of thousands of samples with good cluster recovery (Belbin 1987). We provide a basic outline of the three clustering algorithms; the primary literature provides more comprehensive details. On the recommendation of Faith, Minchin & Belbin (1987), the Bray & Curtis (1957) association measure was applied to the data standardised by maximum species abundance (UPGMA and ALOC). TWINSPAN's default recoding of abundance to a 1 - 5 scale was deemed appropriate and its measure of association (χ^2) in-built.

Flexible-UPGMA

Flexible-UPGMA (Belbin, Faith & Milligan in press) provides the same extension to UPGMA (Unweighted Pair Group using ArithMetic Averaging; see Sneath & Sokal 1973) as the flexible procedure of Lance & Williams (1967) did to WPGMA (Weighted Pair group using ArithMetic Averaging). Flexibility means the ability to contract or dilate the multivariate space (see Belbin 1975, 1991) by altering a parameter β in the Lance & Williams flexible formula. As dilation increases (β is moved from 0 toward – 1) clusters appear increasingly well-defined, regardless of true data structure. With contraction (β is moved from 0 toward +1), the opposite occurs and clusters tend to show a chaining effect (see Williams, Clifford & Lance 1971). A recommended β value of – 0.1 (Belbin, Faith & Milligan in press) was used.

ALOC

The ALOC algorithm (Belbin 1987) produces a partitioning of the data into clusters, not a hierarchy. This aspect is appealing for there is little reason to consider survey sites as being hierarchically related. The algorithm has four distinct phases. The first phase makes a single pass over all sites, sequentially comparing each site to one or more 'seed-sites'. The first site is nominally chosen as the only seed. If any site exceeds a user-defined threshold association (radius) to all seeds, it becomes an additional seed. This procedure samples the volume of the multivariate space. Defining a small radius results in more clusters than a large radius. The second phase allocates all sites to their closest seeds. The third phase replaces the seeds with cluster-centroids based solely on group membership. The last phase is iterative. Each site is extracted from its assigned cluster (centroid re-calculated) and the distance to all cluster-centroids determined. The site is then allocated to the closest centroid. This process of extraction, testing and re-allocation continues until cluster membership stabilises (no sites change cluster membership).

TWINSPAN

Two-Way INdictor SPecies ANalysis (TWINSPAN, Hill; 1979a, p. 3) is based on a 'hand' method of constructing two-way tables of sites and species from Mueller-Dombois & Ellenberg (1974, Chapter 9). TWINSPAN dichotomises sites based on the first reciprocal averaging axis. This axis is divided roughly in the middle. Differential species are defined by a preference for sites on one or other side of the dichotomy. This is the primary ordination. A refined ordination of sites is then constructed by assigning weights to those species which are preferential to either side of the dichotomy. Preferential species weights are summed for each site and the location of the dichotomy re-positioned. The last step is an indicator ordination; the weights of the most preferential species are summed across each site. The indicator ordination is an attempt to reproduce the refined ordination using minimal few species. It is a diagnostic tool, providing a simplified key to the dichotomies and not used here for group definition. The default options for TWINSPAN suited the simulated test data sets.

TWINSPAN applies a similar strategy to classify the species. This aspect was not examined because it was considered that sites and species are not as interchangeable as the mathematics may indicate. This applies particularly in this context to reciprocal averaging where average site scores are calibrated by species and vice-versa. While sites may generally be represented by points in environmental-space, species will occupy a finite volume with a varying density distribution.

Evaluation criteria

Each of the three clustering methods was constrained, as far as possible, to produce the known number of clusters. While the true number of clusters is rarely if ever known in real data, a viable alternative procedure could not be established without introducing mitigating evaluation factors. The three clustering algorithms were applied to all 480 data sets and results compared with the known clusters of the 50 sites using the Hubert & Arabie (1985) version of the Rand statistic (1971). Hubert & Arabie modified the original statistic to take a value of zero as the number of agreements reached those expected by chance (the null model). A value of one implies a perfect match between the clusters. The measure also accommodates comparison of different numbers of clusters.

Results

An analysis of variance was applied to the recovery levels (Rm) across all factor-combinations. Results indicate three significant ($P < 0.001$) single factors and two two-factor interactions. Flexible-UPGMA provided the best mean recovery overall (0.786). ALOC was next best with a mean recovery of 0.767. The difference between ALOC and flexible-UPGMA was not significant ($p > 0.05$). TWINSPAN with a recovery of 0.635 was significantly less than both flexible UPGMA or ALOC ($p < 0.001$).

Recovery was significantly better for the 'long' gradient in five out of the ten configurations (see Table 1). For configuration 8 this was reversed. In the other four cases no significant difference was observed. One explanation could be that the longer gradient simply made it easier for the clustering algorithms to partition the data in this direction. One would expect that this would be relevant to TWINSPAN where dichotomisation is always along the primary gradient. Gradient length × algorithm was not however significant at $p < 0.05$.

The most interesting of the significant interactions was algorithm × sample configuration. This showed that the average TWINSPAN recovery for each of the ten configurations was significantly less than the next best algorithm (either flexible-UPGMA or ALOC) except for sample 7. In this case ALOC was significantly better than flexible-UPGMA and TWINSPAN ($p < 0.001$). Over all ten sample configurations, flexible-UPGMA won six times, ALOC four and TWINSPAN nil (Table 1).

A typical dataset was selected to enable a more detailed examination of the differences in recovery by algorithm. A dataset identified as 'd-120106' provided a

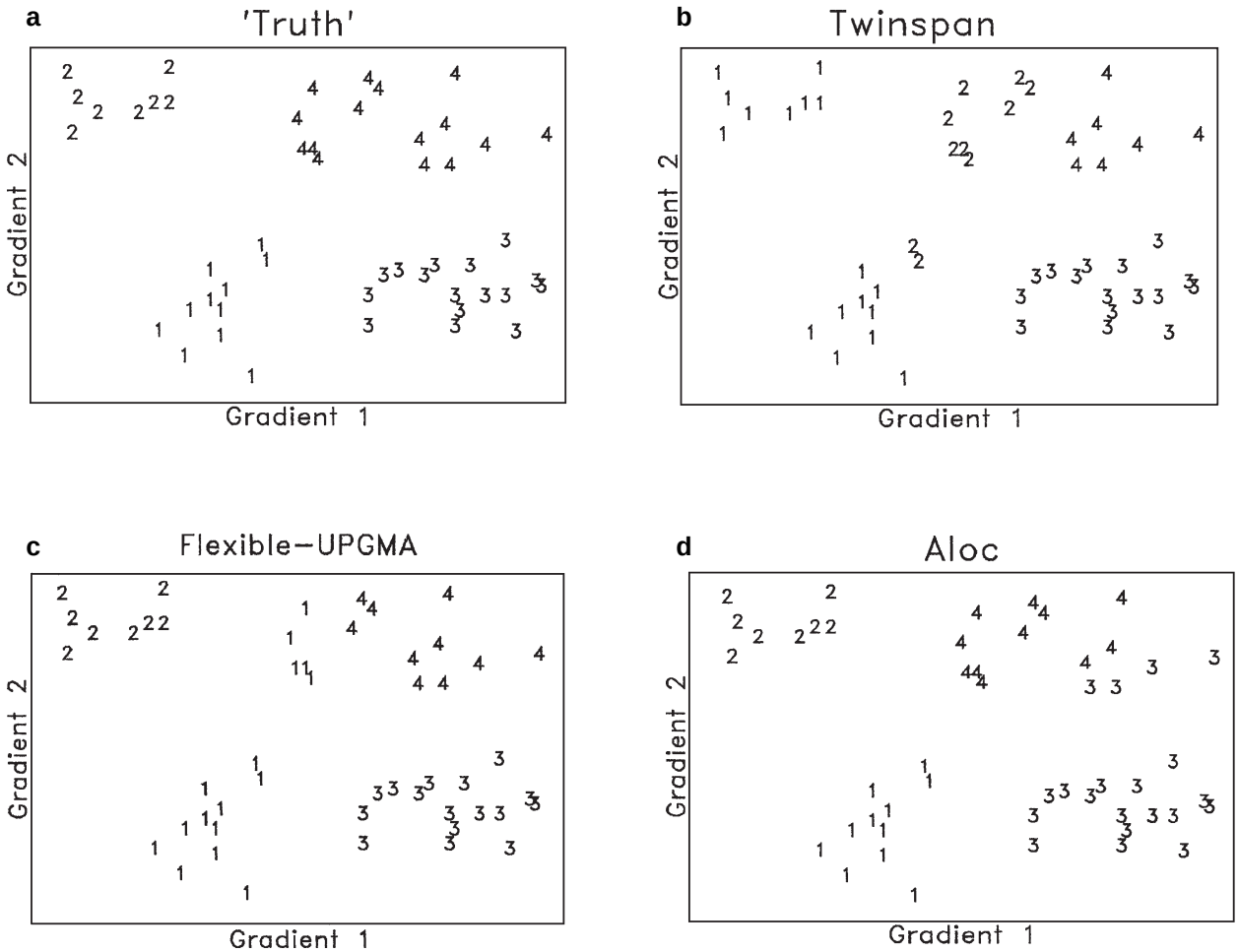

Fig. 1. (a) True configuration of the 50 sites and four groups in sample six, (b) TWINSPAN configuration, (c) flexible UPGMA configuration and, (d) ALOC configuration.

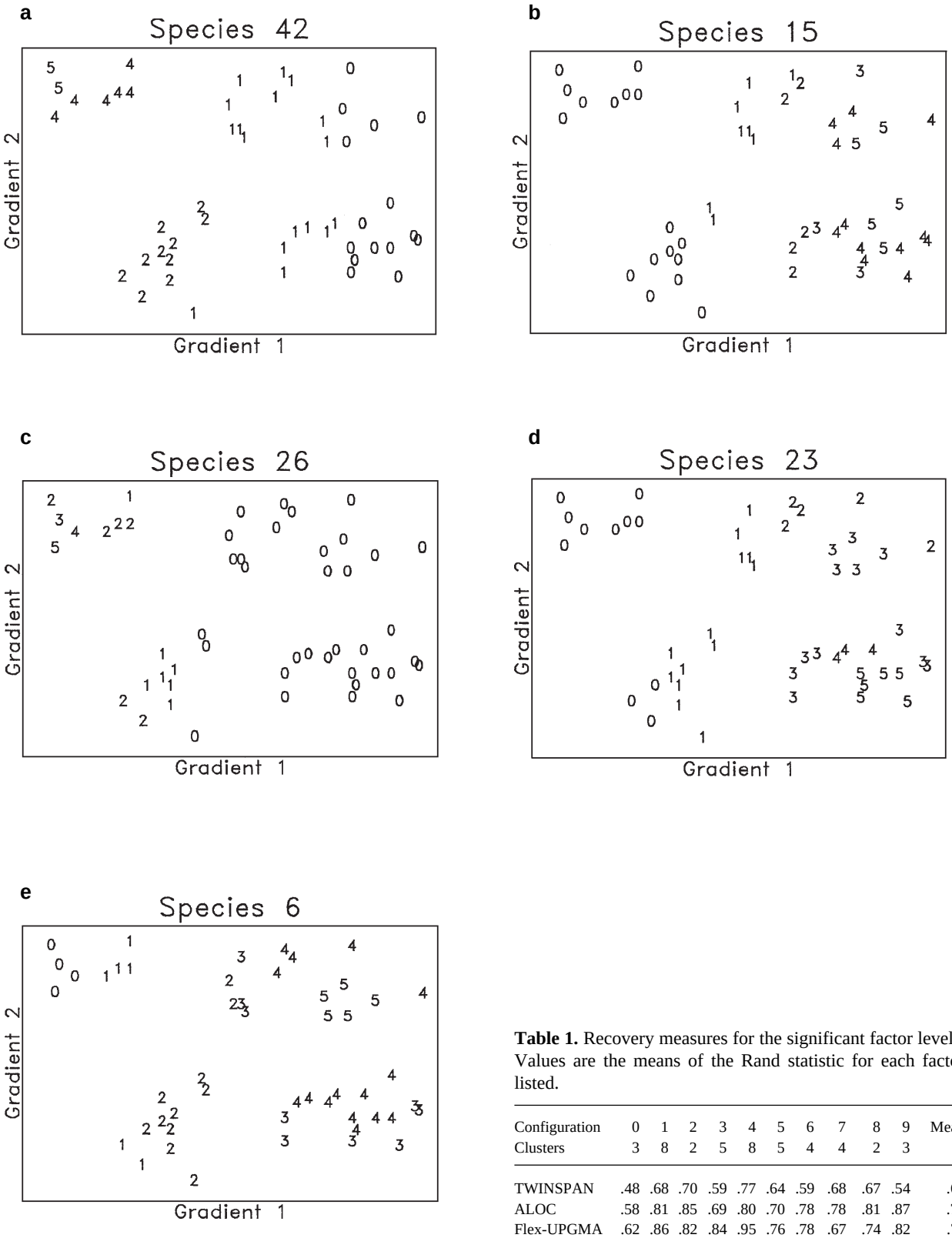

Fig. 2. Abundance of the significant species at the 50 true site configuration of sample 6. Values are scaled 0 (absent) to 5.

Table 1. Recovery measures for the significant factor levels. Values are the means of the Rand statistic for each factor listed.

Configuration	0	1	2	3	4	5	6	7	8	9	Mean
Clusters	3	8	2	5	8	5	4	4	2	3	
TWINSPAN	.48	.68	.70	.59	.77	.64	.59	.68	.67	.54	.63
ALOC	.58	.81	.85	.69	.80	.70	.78	.78	.81	.87	.77
Flex-UPGMA	.62	.86	.82	.84	.95	.76	.78	.67	.74	.82	.79
Short gradient	.51	.79	.76	.71	.84	.67	.65	.64	.80	.74	.71
Long gradient	.61	.78	.83	.71	.84	.73	.79	.78	.67	.74	.75
Mean	.56	.78	.79	.71	.84	.70	.72	.71	.74	.74	

close fit to the mean recovery values for each algorithm; 0.65, 0.77 and 0.77 for TWINSPAN, ALOC and flexible UPGMA respectively. This set used the short gradient model, no species interactions, a 'linear ramp' carrying capacity, no noise and sample configuration 6. Fig. 1a shows the sample configuration of 'd-120106' in the environmental-space with the true clusters imposed. Figs. 1b-d display respectively, the TWINSPAN, flexible UPGMA and ALOC clusters.

Discussion

Why have TWINSPAN, flexible UPGMA and ALOC produced their respective partitions from 'd-120106'? Was there a pattern in how particular species contributed to the partitioning? There are potentially two ways to examine the contribution of each species. One could examine the role of species at each stage in the clustering procedure. However, TWINSPAN is the only procedure that directly identifies preferential species. An alternative is to examine the ability of species to discriminate between the resulting clusters. This procedure was used to evaluate flexible-UPGMA, ALOC and the 'true' cluster composition. In the discussion below, 'cluster' is used to refer to the 'true-groups' while 'groups' refer to the results produced by the algorithms.

Cluster 3 was the only one fully identified by TWINSPAN (Fig. 1b). Out of the 50 sites, 18 or 36 % were mis-classified. The notable error was lumping of clusters 1 and 2 into TWINSPAN group 1 (compare Figs. 1a and 1b). To compensate for this, cluster 4 was split into TWINSPAN's group 2 and 4. Once cluster 4 splits, the divisive nature of TWINSPAN precludes re-connection. This highlights probably the most severe drawback of divisive clustering methods; an early error has significant implications. The distance between the closest neighbours from clusters 1 and 2 was 0.44 (Bray & Curtis). This compares with 0.34 between the closest neighbours across the gap in cluster 4. In other words, a good partitioning of the sites should have preferred to split across the obvious gap between clusters 1 and 2 rather than split cluster 4. This corroborates van Groenewoud's conclusion (1992, p. 245) that "groups of sample points that straddled division points were split at that point".

TWINSPAN's best negative preferential on the first dichotomy was species 42 (Fig. 2a). This species was localised in sites to TWINSPAN groups 1 and 2. The best positive preferential was species 15 (Fig. 2b), in TWINSPAN groups 3 and 4. It appeared that TWINSPAN did not split clusters 1 and 2 on the basis that both shared 18 of the 50 species. The problem with this is that 13 of these 18 species showed virtually no overlap in abundance values. TWINSPAN (by default) looks at the merit of each species on the basis of a single abundance code (1-5). Examining the differences between cluster 1 and 2 revealed that 14 species showed a significant preference for one or other. This is weighty evidence for these clusters to be separated.

Given that the Reciprocal Averaging axis is from the southeast to the northwest corners of the environmental space, it is understandable why TWINSPAN split where it did. The slight gap in cluster 4 and the gap between clusters 1 and 3 line-up about half way along the first RA-axis. The largest gap in sites along this line between cluster 2 and the rest occurs in the northwest third of the gradient. Failing to recognise this gap is most evident when TWINSPAN groups 1 and 2 were created. These groups are separated along a line running from southwest to northeast. This highlights a significant feature of TWINSPAN; groups separated by angles that are near perpendicular to the primary reciprocal averaging axis will not be identified at that stage of division.

TWINSPAN's second dichotomy placed the gap between cluster 1 and 4 too far 'south', incorporating two sites of cluster 1 into cluster 4. The indicator species here (26, Fig. 2c) was localised to cluster 1. The two mis-classified sites, in common with cluster 4, had no species 26. There are however another ten species that would have correctly discriminated all members of these two clusters. TWINSPAN correctly identified clusters 3 and 4. Species 23 (Fig. 2d) and 6 (Fig. 2e) were the best discriminators of all four TWINSPAN groups. All the discriminating species were relatively frequent, ranging from 28 sites for species 22, 34 sites for species 15, 39 sites for species 23, 40 sites for species 42 and 46 sites for species 6.

Fig. 1c shows that flexible UPGMA incorrectly placed five sites from cluster 4 into group 1. The dendrogram from the classification showed that cluster 4 would have been sub-divided into east and west subgroups (as in TWINSPAN) at an association value just slightly below the four cluster level. This was to be expected. Why did the 5 sites from cluster 4 get classified into group 1? The 2 group-1 sites just off-centre to the southwest on Fig. 1c are the catalyst. These two sites are slightly closer (using the Bray & Curtis dissimilarity), to the other group 1 sites to the northeast than to the group-1 sites to the southwest. The fusion of these two sites with the northeast group moves the group centre southwest to facilitate a subsequent fusion into the complete group 1. The distance between the closest pair of sites between cluster 4 and group 1 was greater than the gaps within group 1. Considering the dissimilarity values, the logic here is justifiable and highlights the imperfect relationship between the species-space and

the environmental-space; all clustering methods are similarly disadvantaged.

14 species showed excellent discrimination among the four flexible-UPGMA groups. These species had little or no overlap in abundance values (highly significant *f*-values). Species 42 (Fig. 2a) and 22 (Fig. 2f) were the most discriminating species. These same species were identified as the best discriminators for both the 'true' and ALOC configurations.

In the case of ALOC (Fig. 1d), four sites from cluster 4 were mis-classified as group 3. The seed samples were within clusters 3 and 4, but the distance of the mis-classified sites to the centroid of cluster 3 was 10 % closer than to cluster 4. ALOC is not sensitive to the fact that there were very small dissimilarity values between the mis-classified sites and adjacent sites in cluster 4.

Why did TWINSPAN achieve a better average recovery than UPGMA for sample configuration 7 (see Table 1), even if this difference was not significant? Sample configuration 7 was the only sample distribution where groups could be readily split on the primary RA axes. All four clusters are approximately equal in size and show good separation at a point half way along the primary ordination axes. This situation would be uncommon.

Conclusions

From the test data, flexible-UPGMA and ALOC provide a better recovery of true cluster structure than TWINSPAN. The insignificant difference between flexible-UPGMA and ALOC was unanticipated (we expected ALOC not to perform as well as flexible UPGMA).

TWINSPAN appears to have two identifiable problems; dependence on a predominant primary gradient (noted by van Groenewoud 1992) and dichotomising at an inappropriate point on this axis. There is no doubt about the use of ordination axes to identify predominant gradients. Problems arise when second or subsequent gradients exist in the data. The placement of the first RA-axis is crucial; groups cannot be re-formed once split. The difficulty in identifying the primary gradient with the first RA-axis depends on the relative significance of subsequent gradients. If the primary gradient is not accurately detected, separation between groups on this gradient is lost. Any offset of the RA-axis from the primary gradient results in groups being incorrectly split (Fig. 1b). Once this occurs, subsequent dichotomisations compound the problem.

Even if TWINSPAN detects the primary gradient, groups arranged along this axis may not be accurately recovered (Fig. 1b). The primary RA-axis is ca. 25 ° off. This error could have been countered if the first dichotomisation separated the northwest group (cluster 2, Fig. 1a). A secondary axis could then be placed trending northeast-southwest with the possibility of splitting cluster 1 from clusters 3 and 4. This did not occur because TWINSPAN opted in the secondary species scoring phase to dichotomise at a point approximately 50% along the primary axis. The failure to recognise the large gap around cluster 2 is significant.

Acknowledgements. We wish to thank CSIRO, Division of Tropical Crops and Pastures for funding Cam McDonald's travel to Canberra for collaboration on this paper. Dan Faith, Mike Austin and Fiona Gell provided constructive comments on drafts of this paper.

References

Austin, M. P. 1976. On non-linear species response models in ordination. *Vegetatio* 33: 33-41.

Austin, M. P. 1985. Continuum concept, ordination methods, and niche theory. *Annu. Rev. Ecol. Syst.* 16: 29-61.

Austin, M. P. 1989. Vegetation: data collection and analysis. In: Margules, C. R. & Austin, M. P. (eds.) *Nature Conservation: cost effective biological surveys and data analysis*, pp. 37-41. CSIRO, Melbourne.

Austin, M. P. & Smith, T. M. 1989. A new model for the continuum concept. *Vegetatio* 83: 35-47.

Austin, M. P., Cunningham, R. B. & Good, R. B. 1983. Altitudinal distribution of several Eucalypt species in relation to other environmental factors in southern New South Wales. *Aust. J. Ecol.* 8: 169-180.

Belbin, L. 1975. A FORTRAN V Program for Agglomerative Fusion for Minicomputers. *Comput. Geosci.* 10: 361-384.

Belbin, L. 1987. The use of non-hierarchical allocation methods for clustering large sets of data. *Aust. J. Comp.* 19: 32-41.

Belbin, L. 1991. *PATN Technical Reference Manual.* CSIRO Division of Wildlife & Ecology, Canberra.

Belbin, L. 1992. Comparing two sets of community data: a method for testing reserve adequacy. *Aust. J. Ecol.* 17: 255-262.

Belbin, L., Faith, D. P. & Milligan, G. W. (in press). *A comparison of two approaches to beta-flexible clustering. Multivariate Behavioural Research.*

Bray, J. R. & Curtis, J. T. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecol. Monogr.* 27: 325-349.

Curtis, J. T. & McIntosh, R. P. 1951. An upland continuum in the prairie-forest border region of Wisconsin. *Ecology* 32: 476-496.

Dagnelie, P. 1960. Contribution à l'étude de communautés végétales par l'analyse factorielle. *Bull. Serv. Carte Phytogéogr. B.V.* 1: 7-71.

Faith, D. P., Minchin, P. R. & Belbin, L. 1987. Compositional Dissimilarity as a Robust Measure of Ecological Distance: A Theoretical Model and Computer Simulations. *Vegetatio* 69: 57-68.

Gauch, H. G. & Whittaker, R. H. 1981. Hierarchical classification of community data. *J. Ecol.* 69: 537-557.

Goodall, D. W. 1954. Objective methods for the classification of vegetation. III. An essay on the use of factor analysis. *Aust. J. Bot.* 2: 304-324.

Hill, M. O. 1974. Correspondence analysis: a neglected multivariate method. *J. Roy. Stat. Soc. Ser. C.* 23: 240-354.

Hill, M. O. 1979a. *TWINSPAN: A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes.* Section of Ecology and Systematics. Cornell University, Ithaca, NY.

Hill, M. O. 1979b. *DECORANA: A FORTRAN program for detrended correspondence analysis and reciprocal averaging.* Section of Ecology and Systematics. Cornell University, Ithaca, NY.

Hubert, L. & Arabie, P. 1985. Comparing Partitions. *J. Classif.* 2: 193-218.

Jain, A. K. & Dubes, R. C. 1991. *Algorithms for Clustering Data*. Prentice Hall, New Jersey.

Lance, G. N. & Williams, W. T. 1967. A General Theory of Classificatory Sorting Strategies: I. Hierarchical Systems. *Comput. J.* 9: 373-380.

McQueen, J. B. 1967. *Some methods for classification and analysis of multivariate observations.* Proc. 5th Symp. Math. Statist. and Probability, Berkeley, 1: 281-297. AD669871, Univ. of California Press, Berkeley, CA.

Milligan, G. W. 1980. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*. 45: 325-342.

Milligan, G. W. 1985. An algorithm for generating artificial test clusters. *Psychometrika* 50: 123-127.

Milligan, G. W. 1989. A Study of the beta-flexible clustering method. *Multivar. Behav. Res.* 24: 163-176.

Milligan, G. W. & Cooper, M. C. 1986. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivar. Behav. Res.* 21: 441-458.

Milligan, G. W. & Cooper, M. C. 1987. Methodological review: clustering methods. *Appl. Psychol. Meas.* 11: 329-354.

Minchin, P. R. 1987. Simulation of Multidimensional Community Patterns: Towards a Comprehensive Model. *Vegetatio* 71: 145-156.

Mueller-Dombois, D. & Ellenberg, J. 1974. *Aims and Methods of Vegetation Ecology*. Wiley, New York.

Rand, W. M. 1971. Objective criteria for the evaluation of clustering methods. *J. Amer. Stat. Ass*. 66: 846-850.

Sneath, P. H. A. & Sokal, R. R. 1973. *Numerical Taxonomy*. Freeman, San Francisco.

van Groenewoud, H. 1976. Theoretical considerations on the covariation of plant species along ecological gradients with regard to multivariate analysis. *J. Ecol.* 64: 837-848.

van Groenewoud, H. 1992. The robustness of correspondence analysis, detrended correspondence analysis and TWINSPAN analysis. *J. Veg. Sci.* 3: 239-246.

Westerman, W. E. 1975. Edaphic climax pattern of pygmy forest regions of California. *Ecol. Monogr.* 45: 109-135.

Westerman, W. E. 1991. Measuring realized niche spaces: climatic response of chaparral and coastal sage scrub. *Ecology* 72: 1678-1684.

Whittaker, R. H. 1956 Vegetation of the Great Smoky Mountains. *Ecol. Monogr.* 26: 1-80.

Whittaker, R. H. 1978. *Ordination of plant communities, 2nd ed.* Junk, The Hague.

Whittaker, R. H. & Niering, W. A. 1965. Vegetation of the Santa Catalina Mountains, Arizona: a gradient analysis of the south slope. *Ecology* 46: 429-452.

Williams, W. T., Clifford, H. T. & Lance, G. N. 1971. Group-size dependence: a rationale for choice between numerical classifications. *Comp. J.* 14: 162-165.

Received 26 August 1992;
Revision received 3 November 1992;
Accepted 3 December 1992.