PN-ABK-250
75235

WORKING PAPER SERIES

FISHERIES STOCK ASSESSMENT

TITLE XII
Collaborative Research Support Program

Fisheries Stock Assessment CRSP Management Office
International Programs, College of Agriculture
The University of Maryland, College Park, Maryland 20742

In cooperation with the United States Agency for International Development (Grant No. DAN-4146-G-SS-5071-00) the Fisheries Stock Assessment CRSP involves the following participating institutions:

The University of Maryland—Center for Environmental and Estuarine Studies
The University of Rhode Island—International Center for Marine Resource Development
The University of Washington—Center for Quantitative Sciences
The University of Costa Rica—Centro de Investigación en Ciencias del Mar y Limnología
The University of the Philippines—Marine Science Institute (Diliman)—College of Fisheries (Visayas)

In collaboration with The University of Delaware; The University of Maryland—College of Business and Management; The University of Miami; and The International Center for Living Aquatic Resources Management (ICLARM).

WORKING PAPER

Working Paper 70

TABLE ANALYSIS BY CLASSIFICATION OF ORDINATIONS (TACO): A Method for the Simultaneous Division

by

John W. McManus

Marine Science Institute, University of the Philippines
Diliman, Quezon city, Philippines

Int'l Center for Marine Resource Development
University of Rhode Island

December 1990

Fisheries Stock Assessment
Title XII
Collaborative Research Support Program

The Fisheries Stock Assessment CRSP (sponsored in part by USAID Grant No. DAN-4146-G-SS-5071-00) is intended to support collaborative research between U.S. and developing countries' universities and institutions on fisheries stock assessment and management strategies.

This Working Paper has been produced by The University of Rhode Island research component in collaboration with the University of the Philippines and in association with the International Center for Living Aquatic Resources Management. Additional copies are also available from:

International Center for Marine Resource Development
The University of Rhode Island
126 Woodward Hall
Kingston, Rhode Island 02881

ABSTRACT

A method is described by which objects and attributes may be divided simultaneously to form a hierarchical classification. TACO involves ordination by correspondence analysis and divisions based on minimizing variance. Significance testing or validation of both the axis and the division points is facilitated through iterative production of eigenvalues and reduced variance indices from the shuffled data matrix. Acceptance values for reduced variance can be used to produce a tree with several divisions at each level. Objects and attributes characteristic of a given class can be determined by use of an index and randomization procedure. TACO can be used with quantitative data without reduction to semi-quantitative form. Comparative tests with coral reef fish community data indicate that the method compares favorably with TWINSPAN and the unweighted pair groups agglomerative method.

INTRODUCTION

In classification studies, it is often desirable to classify objects and attributes simultaneously, so that the relationships between the two sets are made clear. For example, the traditional method of phytosociological table construction results in a sorted data table wherein "blocks" of relatively homogeneous relevés and their characteristic species are grouped together to facilitate further analysis (Braun-Blanquet 1932, Mueller-Dombois and Ellenberg 1974). Unfortunately, this approach is very tedious and a reasonable level of objectivity in the procedure can usually be achieved only after years of experience.

The most popular method for simultaneous classifications of species and sample units in ecology is probably the TWINSPAN program of Hill (1979)(Gauch 1982, Pielou 1984b). This program produces hierarchical divisive classifications of objects, and then conducts a similar analysis of the attributes (usually species). Finally, a table is produced based on row and column exchanges, which optimally displays the two sets of analyses in block form. Divisions are based on reciprocal averaging ordinations and designed to optimally divide species based on discriminate scores. In the process, indicator species are identified which are characteristic of each group of sample units. The complete algorithm is quite complex, and details may be found in Hill (1979).

The algorithm has a number of subjective decisions built in as default options, which may be altered. It is objective in the sense that two workers using the same options on the same data set will achieve the same results (prior to interpretation). This fact, and the impressive degree of speed, space efficiency and portability built into the program, make it a substantial improvement over previous methods. The program also has the property that it usually yields classifications which are intuitively understandable, probably because of the emphasis built into the algorithm on making choices the way an experienced ecologist would. However, the algorithm is not, strictly speaking, a simultaneous division of objects and attributes. The separation of object and attribute analyses into distinct analytical phases occasionally results in tables with poor block structure. The number of classes of objects may differ from the number of classes of species for a given minimal class size. The attribute groups are sometimes difficult or impossible to pair up with groups of species on a one-to-one basis. Other limitations include the difficulty with which objective validation methods and significance tests can be applied. Also, the method is based on conversions of semi-quantitative data to binary form, and does not lend itself to use with quantitative data without prior reduction. Hill

(1979) recognized a number of limitations in his program, particularly for use in a variety of specialized applications, and encouraged further work on method development. I present here a method intended to complement the TWINSPAN method, which has some advantages for specific applications.

The TACO Algorithm

Methods based on classifications of ordinations have been previously presented by Noy-Meir (1973), Lefkovitch (1976) and others. Noy-Meir (1973) proposed using a principal components analysis (PCA) as a basis for the ordination of objects. The optimal point for division along this ordination was that which resulted in the greatest reduction of variance of sample points projected along the ordination axis. The method had the advantage that successive hierarchical divisions need not be binary -- it is possible to develop a criteria for accepting two or more optimal divisions at each level. The fact that the method was based on a quantitative index (reduction of variance) meant that it could be validated or subjected to significance testing (depending on statistical approach and philosophy) by computer intensive methods such as those later outlined by Noreen (1989) and demonstrated below. A potential disadvantage which has become of recent concern involves the problem of nonlinearity in ecological data (Gauch 1982, Greig-Smith 1983, Pielou 1984b). A primary gradient reflected in sample units often appears as an arch with respect to the principal axis in PCA analysis (Fig. 1). The points near the ends of the arch are often inverted and compressed when projected onto the axis. If the optimal division was to be found among these points, a misleading result could be achieved.

One solution to the problem of end-point inversion would be to use a correspondence analysis (CA) (essentially equivalent to reciprocal averaging) axis for the ordination (Pielou 1984b). CA ordinations often result in an arch effect, but the end points generally do not invert. This is not an ideal solution, because the end points in CA are often compressed upon projection. However, this is much less of a problem than the end-point compression and inversion associated with projections along a PCA axis.

The ordination in CA analysis is usually strongly correlated with the strongest gradient affecting the species distributions, and both species and sites may be ordinated along the same axis. In this case, the species tend to be located based on their mean distribution among sites, and vice versa. It stands to reason that any substantial break in continuity of sites would be associated with a similar break in the distribution of species along the axis. Figure 2 shows a set of artificial data with a strong 3-block structure. Figure 3 shows a CA ordination of the

data based on the first two axes. There are two strong breaks in the distributions of both objects (sample units) and attributes (species) with respect to projections on the first axis. Figure 4 shows the data in the table rearranged into a roughly diagonal (block) form according to the CA coordinates. The two division points along the first axis can be used to divide both the sites and species into blocks.

In the current algorithm, the division points are located with respect to objects (sample units) first. The total variance for the points along the axis is calculated first, and then each possible division is tested. The variance of points to the left of each division is added to the variance of points to the right, and the total is subtracted from the overall point variance. The result is standardized by the squared range of values, as recommended by Noy-Meir (1973). The formula is:

$$SRV = \frac{V - (VL + VR)}{(High - Low)\ 2}$$

where
- SRV = Standardized reduction of variance
- V = total variance of points along the axis
- VL = variance of points on the left of a division
- V_R = variance of points on the right of a division
- High = Highest coordinate value
- Low = Lowest coordinate value

The division point (or points) which yield(s) the greatest reduction in variance is (are) selected. The division among attributes is based on similar trials among attributes (if any) located within the bounds of the coordinates of objects enclosing the best object break. If no attributes fall within the bounds, the division point for species is defined by the division site for sites.

When the method is applied to a smooth gradient, the best division tends to be in the center (Fig. 5). This is a desirable property, because the ordination already represents an ordering of extremes which would usually be best separated even in an arbitrary division. Trials using a simple reduction in information diversity in the unprojected data set produced a technique which was biased toward end-point divisions rather than the center of a smooth gradient, and this was deemed less desirable for most classification purposes.

Validation

Breaks in Gradients

To say that a classification of a smooth gradient is useless would be no more reasonable than saying that contour lines on a map are useless (Pielou 1984b). However, it is useful to know whether a smooth gradient or a substantial break has been encountered. This requires that some criteria separating the two conditions be applied. The fact that the reduction of variance can be summarized into a simple criteria index makes this a simple matter. Based on a number of experimental trials, Noy-Meir suggested a decision criteria in the range of 0.02 - 0.06 for his index of standardized data reduction. He suggested using the chosen value as a criteria for accepting multiple divisions at a given level in a hierarchical analysis (Noy-Meir, 1973).

The availability of high-speed microcomputers makes it feasible to use a randomization approach in determining whether a break is significantly non-random or not. More specifically, we can test whether the break is a consequence of the specific way the objects and attributes are ordered, or whether it is likely that a break of a given reduction level could have arisen merely as a consequence of the nature of the matrix in terms of dimensionality, patchiness, sparsity, and so forth. The TACO algorithm includes an option whereby the quantities in the matrix are shuffled with respect to both objects and attributes for a large number of times. The probability that the decision criteria could have been achieved randomly is (Hope 1968, Noreen 1989):

$$P = \frac{NGE + 1}{N + 1}$$

where NGE = the number of shuffled rearrangements which yielded a reduction value greater than or equal to the test value

N = the number of shuffled rearrangements

The number of rearrangements should generally be more than 20, and high levels of confidence can be achieved when N > 100. An arbitrary level of 5% can be used to distinguish between continuous and discontinuous gradients. It can be also used for determining whether or not to accept multiple divisions at a given level for circumstances wherein this is desirable. Future versions of the program are expected to incorporate a routine to

calculate the expost probability, i.e. the probability which would have been approached (assuming uniform priors) if N approached infinity (Noreen 1989).

In order to accomplish the above shufflings and still preserve dimensionality, it is necessary to build in a corrective routine. The routine tests each row and column to ensure that at least one non-zero value is present. When there is none, the routine searches randomly within the matrix to find a non-zero value in a row and column each with at least one other non-zero value present. This value is then moved to a random place in the aberrant row or column. The search continues until all rows and columns contain non-zero values. The procedure does not introduce any nonrandom bias, because ultimately, any non-zero value has an equal chance of being shuffled to any position in the matrix. Note, however, that 'randomness' as used above is limited to the pseudo-randomness provided in the computer algorithm using Quick Basic functions.

Trends Along Axes

The randomization approach can similarly be applied to determining the uniqueness of the gradient indicated by the ordination axis. Each CA axis is associated with a unique eigenvector, which is in turn related to a single eigenvalue describing the variability of attributes with respect to objects along the axis (ter Braak 1988). The hypothesis to test is whether the relationship thus indicated is uniquely related to the distribution of values among the objects and attributes as ordered. The null hypothesis would be that any matrix of similar dimensionality, patchiness, sparsity, etc., would have resulted in an eigenvalue of equal or greater value. The procedure follows that above, and is accomplished simultaneously in the current algorithm. Testing is done after converting the original and generated eigenvalues to percent of the total set of eigenvalues, so that the focus is on the proportion of variability explained. I suggest here that a similar validation procedure could be applied to virtually any ordination procedure based on eigenanalysis, including the oft-used PCA.

Indicator Objects and Attributes

It is often useful in interpreting the results of a classification to have a preset criteria for determining if an object is characteristic of a group of attributes, or if an attribute is characteristic of a group of objects. One possible index for whether or not a species can act as an indicator of conditions in a group of sites produced by a division is:

$$IR = \frac{NR / OR}{(NR/OR) + (NL/OL)}$$

where IR = Index value for a variable (species) being an indicator of the objects on the right side of a division (a high value would be a right indicator, a low value, a left indicator, and a moderate value an ubiquitous attribute).

NR = Tabular value (number of individuals) totalled among objects on the right side of a division for the attribute (species) under consideration.

NL = Same as NR, but for objects on the left side of a division.

OR = Number of objects on the right side of a division.

OL = Number of objects on the left side of a division.

Obviously, if the tabular value is very small, or if the distribution of values among objects is very patchy, then it will be difficult to confidently determine if the attribute is a right or left indicator. The confidence can be determined by shuffling the values randomly among objects (within a row) and recalculating the index many times. This results in a two-sided analysis, where a very low IR and high NG (number of randomized results greater than the test value) would be a left indicator, a high IR and low NG would be a right indicator, and everything else would be ubiquitous or indeterminate. This would be true assuming that the number of "equals" is very low and therefore that:

$$LG \approx N - NG$$

where LG = number of results after shuffling less than the test value.

Acceptance can be placed arbitrarily at 5% and 95% of trials, although this does not translate precisely into probabilities as in the prior cases. A similar procedure may be applied to objects by shuffling and calculating within columns. Table 1 shows the results for each attribute and object in the 3-block data set.

Field Trials

For comparisons in the performance of this method against other popular routines, it was desirable to use a data set known to have a considerable degree of continuity. If the test data represented a highly discrete set of objects, nearly any

classification method would tend to yield identical results. A suitably non-discrete data set is that of a fish community on the fore-reef slope of Santiago Island, Bolinao, Pangasinan. We have previously shown that the fish community is variable enough that no two sites classified together consistently across two-level TWINSPAN analyses of time-slice tables from nine samplings at alternate months (Nañola et al. 1990).

A set of 18 visual transect sites was sampled on alternate months for the year July 1987 - June 1988. Each transect was 200 m by 10 m, and paired observers counted fish to a height of 5 m from the bottom. Data from 6 periods were averaged into a single table of sites by species. This table was analyzed in two strategies. In one, a set of three CA axes was generated from the whole table, and each was divided once. These divisions were used to produce a hierarchical set of divisions, as recommended by Noy-Meir (1973). The second strategy involved extracting a primary axis, dividing it into two regions, and then subdividing the left and right sets by extracting new primary axes. This method of "diminishing area", as used in TWINSPAN, was expected to yield different results, because there is intuitively no reason to expect that the residual variance of the whole set of points should equate to the primary trend in a reduced subset of points. However, the results were the same for both strategies
(Fig. 6).

The site classifications were compared against those generated by TWINSPAN and the unweighted pair groups method (UWPGM) agglomerative method cut off at two levels of division (four classes). The results indicated that TACO and TWINSPAN each misclassified more sites relative to UWPGM than to each other (Table 2). There was no indication that TACO produced a less reliable or less interpretable result than did TWINSPAN. In fact, only TACO isolated four sites which were located in very shallow, wave-swept water which had been previously labelled as a unique set by fieldworkers during site selection (Fig. 7).

The significance of the axes determined iteratively (30 trials each) was compared to that determined by two existing methods for determining trend significance. Skewer analysis (Pielou 1984a) is a rank correlation method which seeks correlations between the presumed ordering and the ordering of the actual points projected on axes randomly thrust through the data cloud. Ordered Similarity Matrix Analysis (OSMA) (Pielou 1979, 1983 -- see also McManus 1985) involves a semi-quantitative analysis of a presumed trend in a set of data points based on expected orderings in a similarity matrix. The results indicate that if anything, the eigenvalue iteration method was more sensitive than the other methods (Table 3). The lack of significance in many of the divisions confirms the expectation that the fish community in question was fairly continuous in its distributions along multiple gradients.

DISCUSSION

The major shortcoming of the TACO method in its current state is the difficulty with which a table may be sorted to present the results of successive divisions. Results in TACO can currently be only presented either as tree diagrams for objects and attributes, or as sets of tables representing successive divisions. Combining these into a single table will require trade-offs, although an optimal solution is theoretically possible for a given set of constraints. Currently, if a simple sorted table is the objective, TWINSPAN is the preferable technique. However, in many cases, the trade-offs involved in the TWINSPAN tabular sorting steps can obscure the underlying structure of the data and lead to misleading interpretations. It is recommended that analytical emphasis be placed on the divisive steps in the order they were performed in either method, rather than only on the final tabulations.

TACO has the advantage that it can operate on quantitative data to a reasonable level of precision. TWINSPAN requires such data to be reduced to a scale of no more than 10 semi-quantitative categories.

The current TACO method is oriented specifically to the identification of major gradients and the division of objects and attributes with respect to them. This is a desirable feature in some studies, and makes interpretation of results fairly simple. For other purposes, however, the dependence on gradients may not be as helpful. An optimal division of the points as projected along the primary CA axis may not necessarily correspond to the optimal division of the same points in multivariate space. For some purposes, it may be desirable to change the SRV criteria into one which minimizes variance in multivariate space instead of depending on the univariate projections of the points. Preliminary trials with such a criterion have been promising. However, for many current uses in ecology and other fields where divisions reflecting data trends are desirable, the current implementation of TACO may be the method of choice.

ACKNOWLEDGMENTS

This research was sponsored under the U.S. AID Fish Stock Assessment Collaborative Research Support Program, University of Rhode Island / Philippines Component, "A Search for New Methods for Fish Stock Assessment" (USAID Grant No. Pan-4146-G-SS-5071-00). I would like to thank all members of the FSA-CRSP team in Bolinao, especially C.L. Nañola Jr., R.B.

Reyes Jr., and M.H.S. Jaranilla for computer and field work. This is publication number xxx of the Marine Science Institute, University of the Philippines.

REFERENCES

Braun-Blanquet, J. 1932. Plant sociology: the study of plant communities. (Transl. by G.D. Fuller and H.S. Conard.) Transl. of 1st ed. of Pflanzensoziologie (1928). McGraw-Hill, New York and London. 1983 reprint, Koeltz Scientific Books, Koenigstein, Germany. 438 p.

Gauch, H.G. Jr. 1982. Multivariate analysis in community ecology. Cambridge University Press 298 p.

Greig-Smith, 1983. Quantitative plant ecology. Studies in ecology, v. 9. University of California Press. Berkeley. (Blackwell Scientific Publications. U.K.) 359 p.

Hill, M.O. 1979. TWINSPAN -- A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Cornell University, Ithaca, NY. 90 p.

Hope, A.C.A. 1968. A simplified Monte Carlo significance test procedure. J. Royal Statistical Society. 30:582-598.

Lefkovitch, L.P. 1976. Hierarchical clustering from principal coordinates: an efficient method for small to very large numbers of objects. Math. Biosci. 31:157-174.

McManus, J.W. 1985. Descriptive community dynamics: Background and an application to tropical fisheries management. Ph.D. Dissertation, The University of Rhode Island. 217 p.

Mueller-Dombois, D. & Ellenberg, H. 1974. Aims and methods of vegetation ecology. Wiley, New York. 547 p.

Nañola, C.L. Jr., McManus, J.W., Campos, W.L., del Norte, A.G.C., Reyes, R.B.Jr., Cabansag, J.P.B., & Pasamonte, J.N.D. 1990. Spatio-temporal variations in community structure in a heavily fished forereef slope in Bolinao, Philippines. p. 377-380. In: Hirano, R. & Hanyu, I. eds., The Second Asian Fisheries Forum. Asian Fisheries Society, Manila, Philippines. 991 p.

Noreen, E.W. 1989. Computer intensive methods for testing hypotheses: an introduction. John Wiley & Sons, Inc. 229 p.

Noy-Meir, I. 1973. Divisive polythetic classification of vegetation data by optimal divisions on ordination components. J. Ecol. 61:753-760.

Pielou, E.C. 1979. Interpretation of paleoecological similarity matrices. Paleobiology 5(4):435-443.

Pielou, E.C. 1983. Spatial and temporal change in biogeography: gradual or abrupt? p. 29-56. In: Sims, R.W., Price, J.H., & Whalley, P.E.S., eds. Evolution, time and space: the emergence of the biosphere. Academic Press, London.

Pielou, E.C. 1984a. Probing multivariate data with random skewers: a preliminary to indirect gradient analysis. Oikos 42:161-165.

Pielou, E.C. 1984b. The interpretation of ecological data. John Wiley & Sons, Inc. N.Y. 263 p.

ter Braak, C.J.F. 1988. CANOCO -- a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal component analysis and redundancy analysis (Vers. 2.1). Agricultural Mathematics Group. Wageningen, The Netherlands. 95 p.

Table 1. Indicator analysis for species. A similar analysis can be performed on sites. (NG is the number of values from randomly shuffled rows greater than the test value).

Species	Indicator value	NG of 1000
1	0.0	0.0
2	0.0	0.0
3	0.0	0.0
4	47.8	30.9
5	75.0	57.8
6	100.0	67.1
7	100.0	66.5
8	100.0	85.3
10	100.0	65.8
9	100.0	66.6

High Indicator Value + High NG = Right Group Indicator
Low Indicator Value + Low NG = Left Group Indicator
Other species are unknown or ubiquitous.

Table 2. Classification comparisions. TACO and TWINSPAN results were relatively close.

TACO					
A:	5	8	9	10	11
	12	14	17	18	
B:	13	15	16		
C:	1	2	3	4	
D:	6	7			

TWINSPAN				
		9	10	11
12	14		18	
13	15	16	17	
		2	3	4
1	5	6	7	8

UWPGM				
5	8	9	10	11
13	14	15	16	17
	2	3	4	
1	6	7		

Missclassified relative to:

TACO 4

TACO 5
TWINSPAN 5

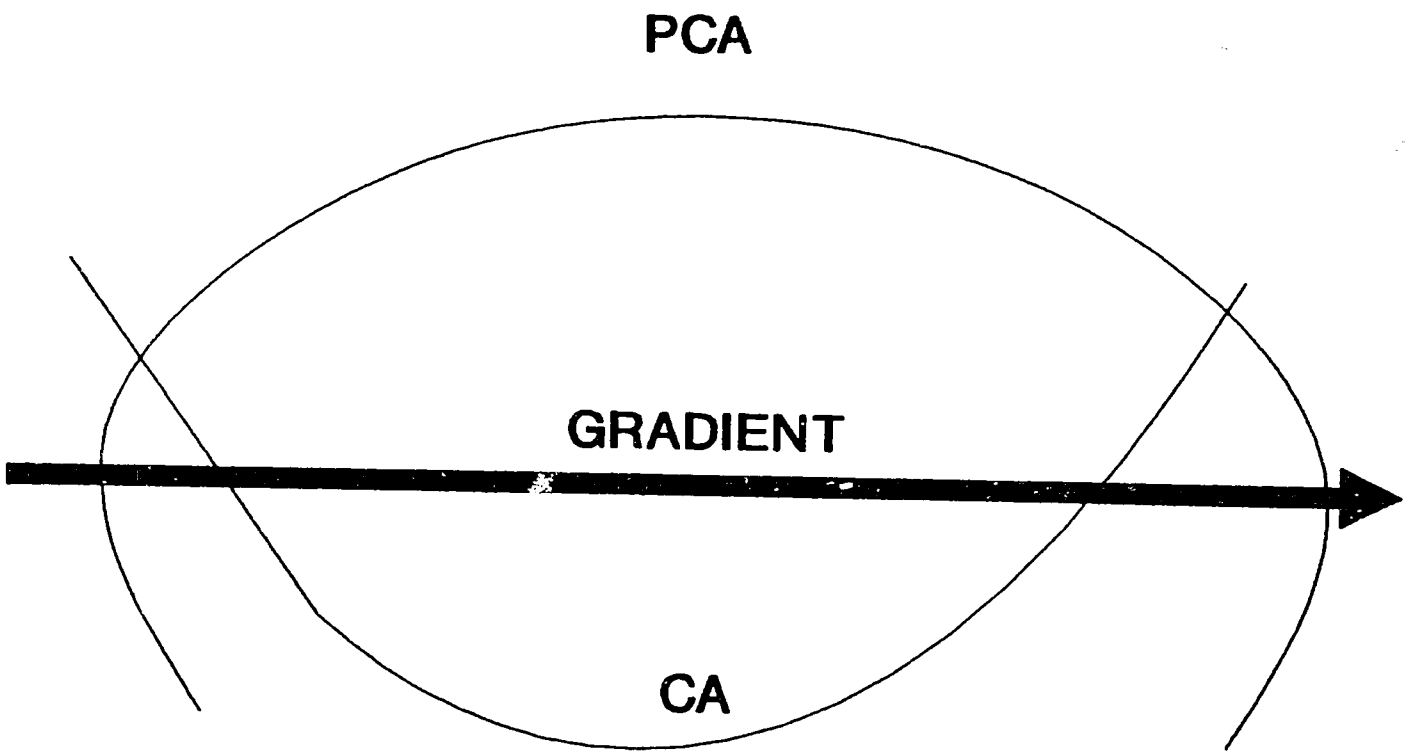
13.

Table 3. Significance tests. Axes 1-3 are from visual fish transect data. % Eig. refers to tests of the significance of the eigenvalue converted to percent of total values for all axes. The continuous nature of the data is obvious from the lack of significance in species divisions.

	Reduced Variance		Axis
Data set	Sites	Species	% Eig
Gradient	ns	ns	*
3-Block set	ns	ns	*
Axis 1	*	ns	*
Axis 2	*	ns	*
Axis 3	*	ns	ns
Axis 1 Left	*	ns	*
Axis 1 Right	ns	ns	ns

Station Order:		Species Order:	
OSMA	All Significant	OSMA	All Significant
Skewer	All Significant	Skewer	All Significant

Figure 1. The arch effect causes endpoint compression in correspondence analysis (CA) and principal components analysis (PCA), but PCA can also lead to inversion of points as they are projected onto a straight axis.



Objects (Sites)

Attributes (Species)										
	4	3	0	0	0	0	0	0	0	0
	3	4	0	0	0	0	0	0	0	0
	2	3	0	0	0	0	0	0	0	0
	1	2	1	4	3	2	1	0	0	0
	0	1	1	3	4	3	1	0	0	0
	0	0	0	2	3	4	1	0	0	0
	0	0	0	1	2	3	1	0	0	0
	0	0	0	0	1	2	1	4	3	2
	0	0	0	0	0	0	1	3	4	3
	0	0	0	0	0	0	1	2	3	4

Figure 2. A hypothetical data table with strong 3-block structure. Each number refers to the value of an attribute for an object, e.g. the number of individuals of a given species in a site. The data reflects the presence of 3 object groups characterized by 3 groups of attributes.

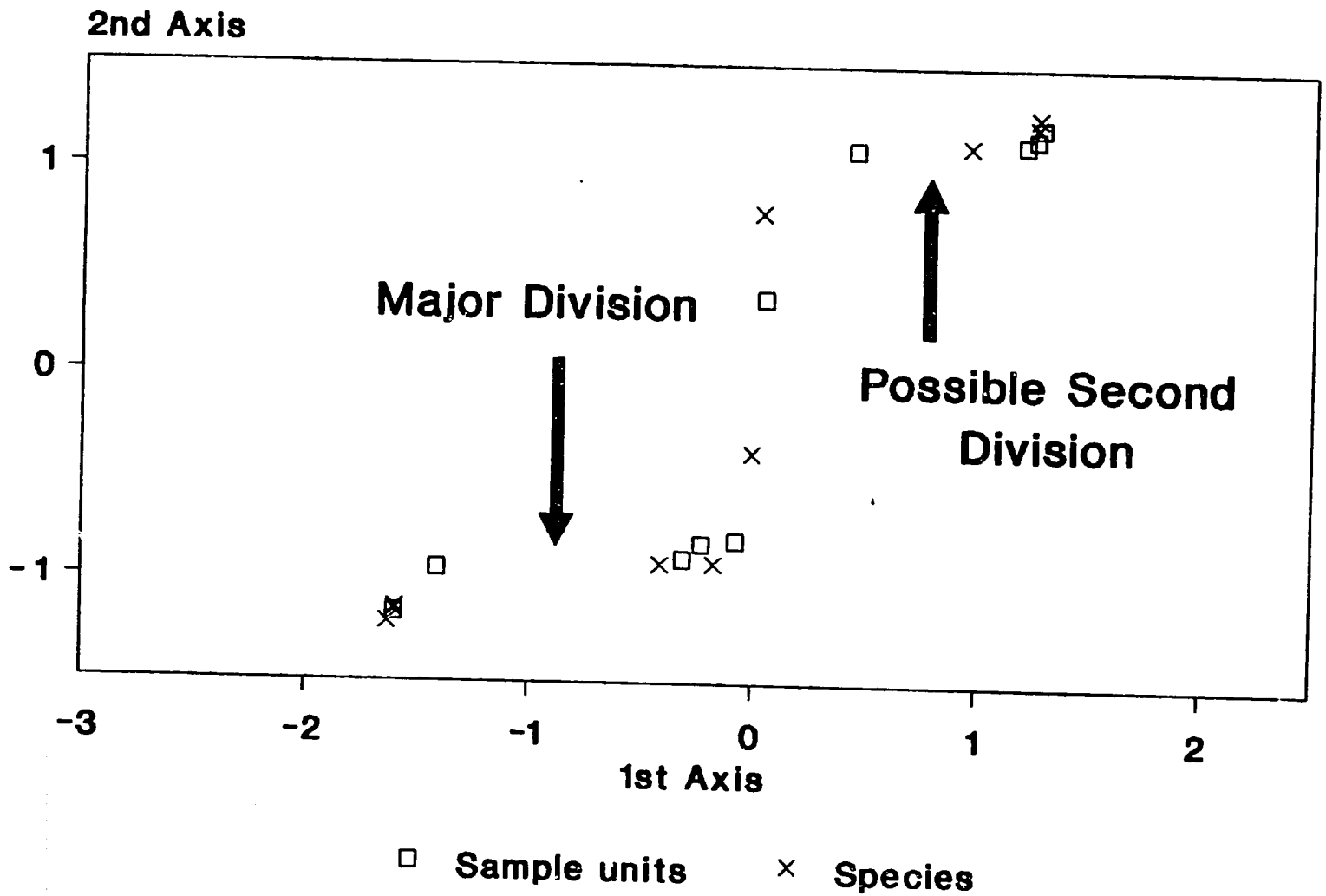


Figure 3. The 3-block data set projected on 2 correspondence axes.

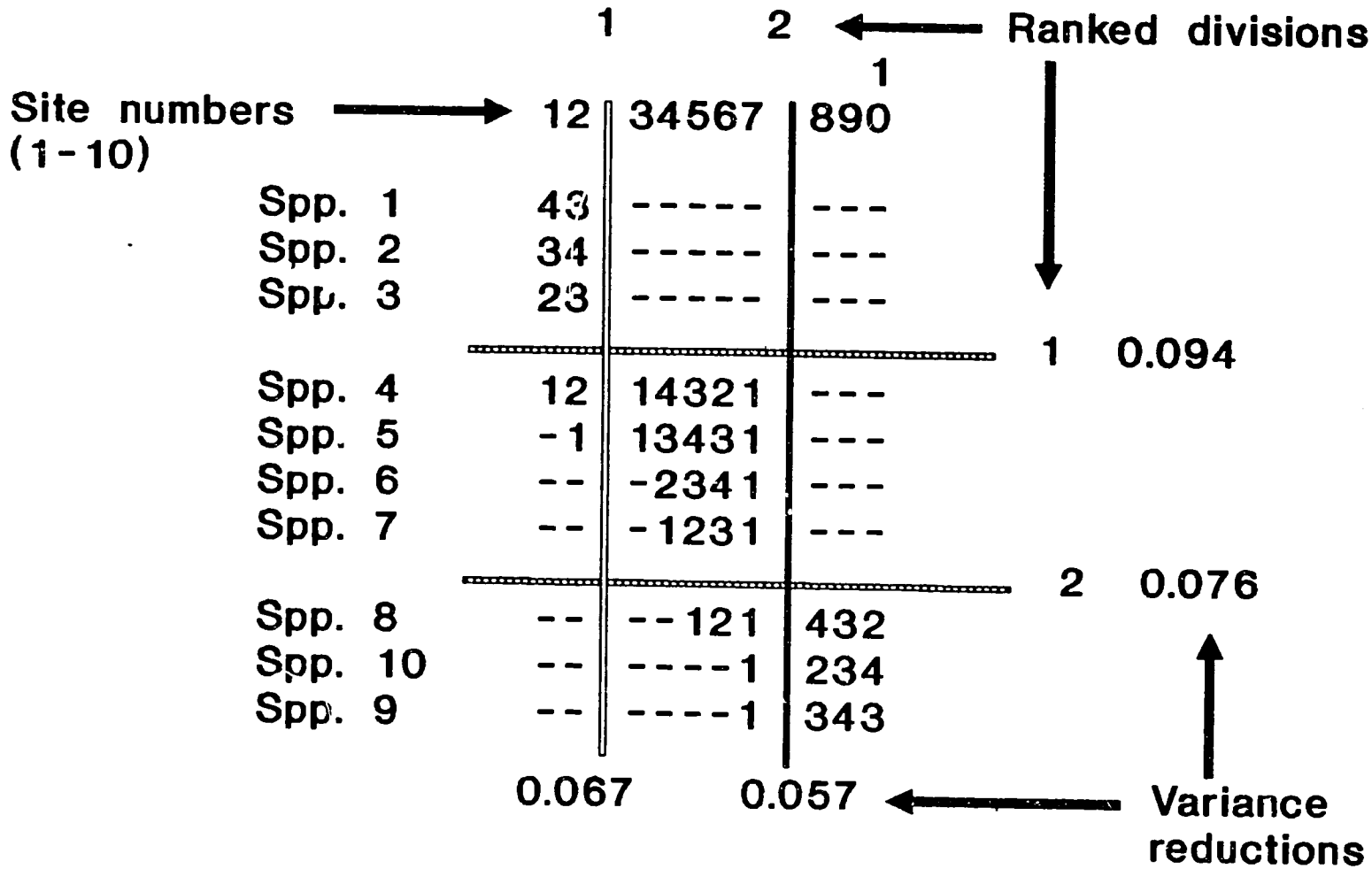


Figure 4. The 3-block data set with species and sites ordered as per the results of CA analysis. Divisions are indicated along with rank preferences and variance reduction values.

```
                    531246
                     \\\\\\
                           1
                  1234567890

1  Spp. 1      4321------
2  Spp. 2      34321-----
3  Spp. 3      234321----  --   5   0.059
4  Spp. 4      1234321---  --   3   0.078
5  Spp. 5      -1234321--  --   1   0.085
6  Spp. 6      --1234321-  --   2   0.078
7  Spp. 7      ---1234321  --   4   0.059
8  Spp. 8      ----123432  --   6   0.037
9  Spp. 9      -----12343
10 Spp. 10     ------1234

                  /\**0.059
                   /\**0.078
                    /\**0.085
                     /\**0.078
                      /\**0.059
                       /\**0.037
```

Figure 5. Divisions of a smooth gradient. Priority tends toward the center of the ordinated data.

Figure 6. Taco classifications based on succesive reductions of the object set with reanalysis (reduced area) and on divisions of multiple axes (successive axes). Data are from visual fish censuses on a forereef slope. Sites 1-4 are very shallow (approx. 5 m) and wave-swept.

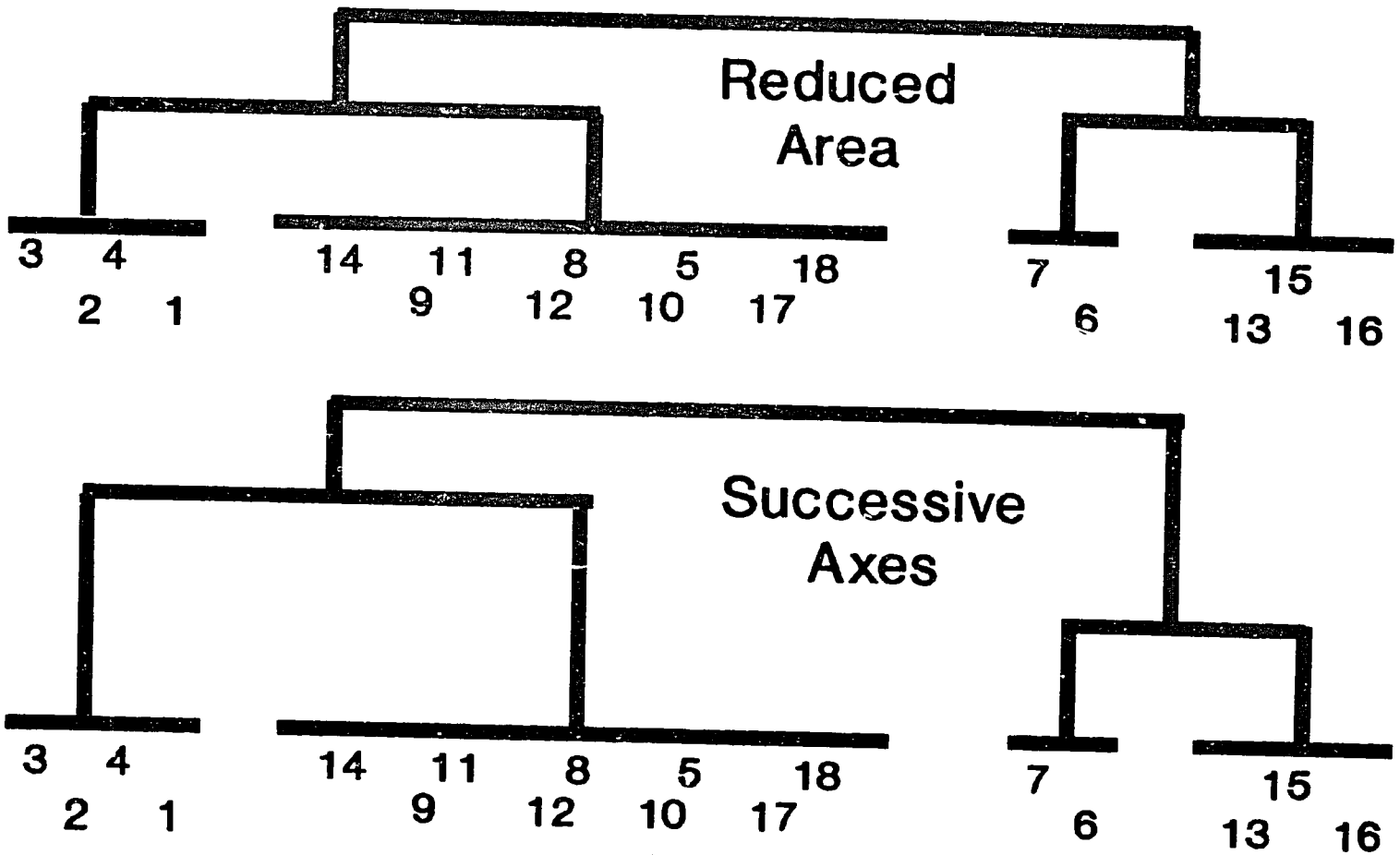
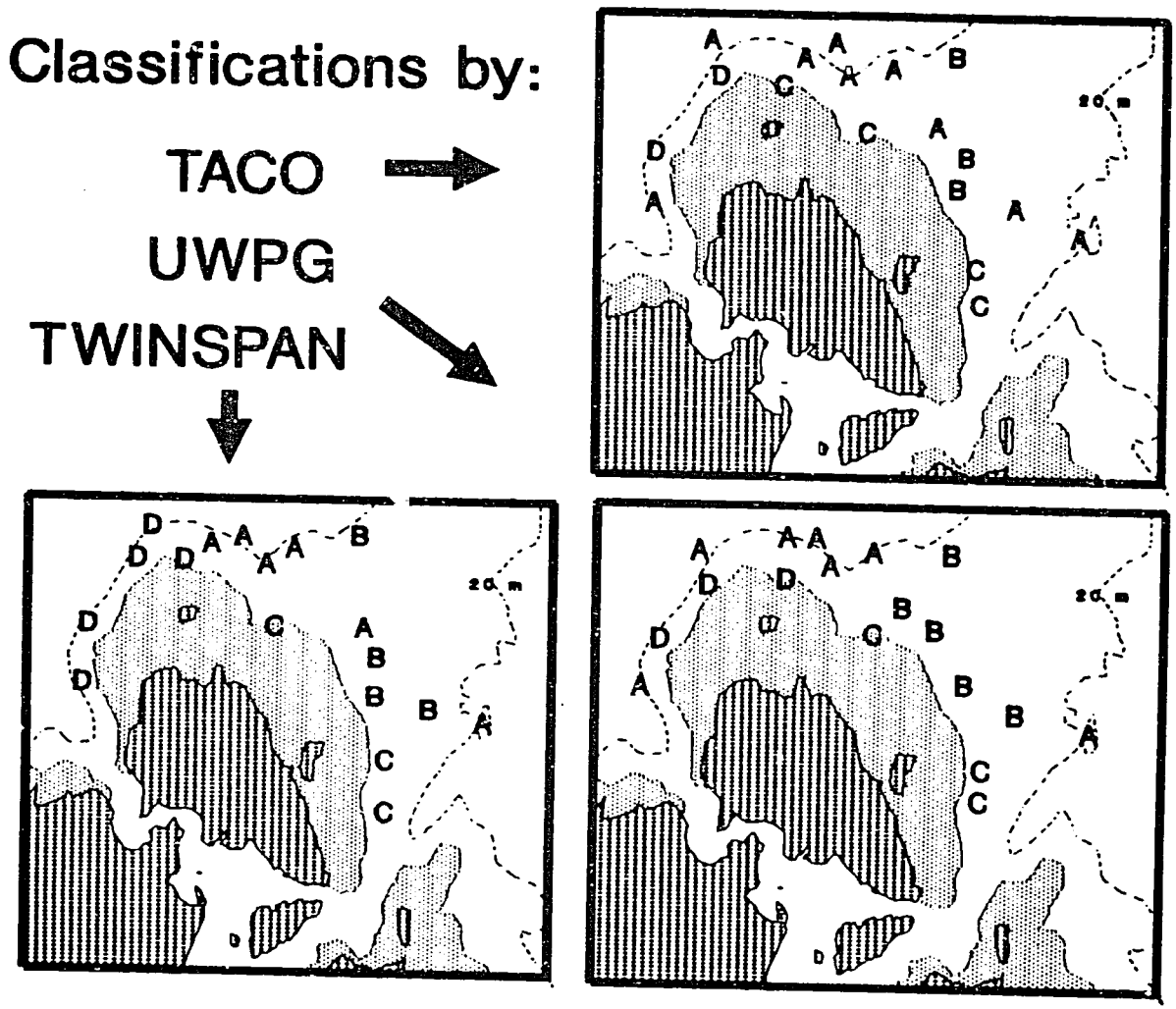



Figure 7. Classifications based on TACO, TWINSPAN, and the unweighted pair groups method (UWPG) of the 18 visual fish transect sites averaged over a year. The 109 top species were included, comprising 90% of the counts.