ORDINATION AND CLASSIFICATION

SOME ORDINATION AND CLASSIFICATION METHODS IN PLANT ECOLOGY

By

CHIH-KANG CHEN, M.S.

A Project

submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

August 1976

MASTER OF SCIENCE (1976) McMASTER UNIVERSITY
(Statistics) Hamilton, Ontario

TITLE: Some Ordination and Classification Methods in Plant Ecology

AUTHOR: Chih-kang Chen, B.Sc. (Tamkang College, Taiwan, Republic of China)

M.B. (Tamkang College)

SUPERVISOR: Dr. Peter D.M. Macdonald

NUMBER OF PAGES: v, 71

ABSTRACT

This project studies the applicability of two ordination methods, principle component analysis and correspondence analysis, and one classification method, mode analysis, for a specific ecological data set. The differences between techniques are discussed and the results are compared.

ACKNOWLEDGEMENTS

Dr. P.D.M. Macdonald, my supervisor, has encouraged and assisted me through numerous discussions and invaluable suggestions. I appreciate also his aid in obtaining the computer access. The generous help of Dr. K.A. Kershaw with the ecological data made this project possible. Finally, I thank my wife, Shih-ling, for her cheerful typing of this manuscript and for her support throughout my stay at McMaster University.

TABLE OF CONTENTS

To My Wife

Shih-ling

1. INTRODUCTION

Multivariate statistical technique seeks a method of reducing the dimensionality of multivariate data in order to give the investigator a much simpler and clearer data structure to examine and interpret. In certain ecological studies, the data come in the form of sampled sites (plots), where each site is measured by the number or coverage or other ecological measurements of a set of species which appeared in it. The number of sites and the number of species sampled usually are very large. By using subjective judgement, an experienced ecologist may identify easily any obvious pattern response of vegetation along some environmental influence such as climate, topography, water supply, soil, etc. But for the more detailed analysis in order to reveal trend and classes of variation in the vegetation, an objective and scientific data analysis technique is needed. Recently, the multivariate statistical technique has been studied as a powerful and potential methodology in the field of ecology (Kershaw 1973 and Orloci 1975).

This project studies the applicability of two ordination methods, principle component analysis and correspondence analysis, and one classification method, mode analysis, for a specific set of ecological data.

The data have been analysed by Dr. Kershaw (1968b). One of his methods for analysis is the principle component analysis with the Domin scale measurement. Firstly, he classified the sample sites by classical subjective Braun-Blanquet association analysis (Blaun-Blanquet 1951) into eighteen associations. Then, he used these associations as a basis of comparison for the results of principle component analysis and other analyses. In this project, I applied the principle component analysis to the frequency measurement data instead of the Domin scale measurement data to examine the different results. Meanwhile, the principle component analysis was carried out by both correlation matrix and covariance matrix to compare the results. Transformed data were also used in the analyses.

Correspondence analysis has been proposed recently (Hill 1973) as an ordination methodology preferable to principle component analysis. Thus I studied this technique and used it on the data set to examine its properties and compare it with the principle component analysis. Mode analysis is one of the cluster analysis techniques for finding the natural grouping and is used here to examine its performance. The first few components extracted by principle component analysis and correspondence analysis are used as input data for the mode analysis as well as original data in order to compare the effect of the reduction of dimensionality. Since there are numerous multivariate statistical techniques available for different purposes and for a variety of sampling

methods, these three techniques in no way exhaust the techniques nor are they necessarily the best ones.

The purpose of this project is to clarify the properties and the usage of these three techniques. In this project, the basis used for evaluation and comparison of the different results is the Blaun-Blanquet association analysis. Since I am not familiar with the ecological context, the project will concentrate on demonstrating the use of these techniques but will not investigate critically the ecological meaning of the results.

2. DATA

The ecological data which are employed in this project are part of Dr. K. A. Kershaw's Nigerian savanna vegetation data. The data were sampled from Zaria province, Nigeria, and the detailed sampling procedure was described by Dr. Kershaw (1968a). There are four hundred and thirty-three sample sites and the area of each site is 100x400 feet. The sample sites were partially randomised with respect to obvious topographical features. For each sample site, the wood species were included in the enumeration by both the Domin scale measurement and the frequency measurement. The frequency measurement is the actual count of the number of occurrences of each species within each of the sample sites. The Domin scale is a simplified representation of the abundance of each species within each sample site. Its eleven scales are defined as follows:

	Domin scale
Cover about 100 percent	10
Cover > 75 percent	9
Cover 50-75 percent	8
Cover 33-50 percent	7
Cover 25-33 percent	6
Abundant, cover about 20 percent	5
Abundant, cover about 5 percent	4
Scattered, cover small	3
Very scattered, cover small	2
Scarce, cover small	1
Isolated, cover small	x

In this project, I used one hundred and ninety-eight sample sites as the input data, which were selected randomly by Dr. Kershaw from the original four hundred and thirty-three sample sites as a convenient size for analysis. Each site was measured by the number of occurences of each of fifty wood species. The species names are in Appendix A. Appendix B is the data list of the first five sample sites.

By referring to Dr. Kershaw's original sample note-book and the paper he published on the results of association analysis (Kershaw 1968a), firstly, I identified the associations of the one hundred and ninety-eight sample sites and then used it as the basis for later comparisons of the different results of analyses. There are four sample sites, site 42, 43, 57, and 58, in the one hundred and ninety-eight sites which do not belong to any of the associations, and the first association does not contain any of the one-hundred and ninety-eight sample sites. The one hundred and ninety-eight sites, therefore, are divided into eighteen groups. One group is the four sites mentioned above and the rest are the seventeen associations which are listed in Appendix C.

The first three associations have relatively common species present in each sample sites, and the dominant species is Dichrostrachys. Associations 4 to 8 have the common dominant species Isoberlinia doka. Associations 9 to 13 have Monotes as the common preeminent species. Associations 14, 15, 16 have Detarium and

Gardenia as the common species. The term Braun-Blanquet association will be used to refer these pre-determined associations in the subsequent sections.

3. METHODS

3.1 Principle Component Analysis

Consider a p by n data matrix $\underset{\sim}{X}$, representing n observations of a p-dimensional variable (X_1, X_2,..., X_p). Let $\underset{\sim}{S}$ denote the sample covariance matrix. Principle component analysis seeks an orthogonal transformation of the original variable to a set of new variables, which are called the principle components (Y_1, Y_2,..., Y_p), such that

$$\underset{\sim}{Y} = \underset{\sim}{A}\,\underset{\sim}{X} \quad .$$

From the orthogonal constraint, the coefficient matrix $\underset{\sim}{A}$ satisfies $\underset{\sim}{A}^T\underset{\sim}{A} = \underset{\sim}{A}\,\underset{\sim}{A}^T = \underset{\sim}{I}$. If we arrange the new variable set by their variance in descending order, then the jth principle component of the sample observations is the linear compound

$$Y_j = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p \quad .$$

The coefficients of this linear compound are the elements of the standardized eigen vector of the sample covariance matrix $\underset{\sim}{S}$ corresponding to the jth largest eigen value c_j. The variance of the jth component Y_j is equal to c_j, and the total system's variance is $c_1 + c_2 + \ldots + c_p = \text{trace } \underset{\sim}{S}$.

Since matrix $\underset{\sim}{S}$ is non-negative definite, the eigen values are all real and non-negative. This transformation reconstructs the original dispersion matrix by a new

set of uncorrelated variables and we may consider them separately. In geometric representation, the principle components are the new variables specified by the axes of a rigid rotation of the original coordinate system into an orientation corresponding to the directions of maximum variance in the sample scatter configuration. These component axes are the least-square solutions of the best fitted lines for the sample. The complete mathematical treatment and proofs can be found in most of the multivariate textbooks.(Cf. Morrison, chap. 7, 1967; Kendall & Stuart, chap. 43, 1967).

The main reason for the above transformation is to try to find a means to display the population structure as economically as possible. If after completing the transformation, the first m components take account of most of the system's variation, then it may be reasonable to discard the remainder and hence reduce the number of variables to be considered. The relative contribution of the jth component toward the total variation can be measured by

$$\frac{c_i}{\text{trace } \underset{\sim}{S}} \quad .$$

Bartlett (1954) developed a testing scheme for deciding on the number of components to be preserved for the multinormal observations. In practice, if the first four

or five components do not take account of a large portion of the system's variation, then it is very hard to interpret the result and the value of the transformation will be in doubt.

Sometimes the components are extracted from the correlation matrix instead of the covariance matrix. In this case, the sum of the eigen values will be equal to the number of dimension p. The general rules for the choice of which matrix to be used in the principle component analysis are: If the responses are in widely different measurement units, the correlation matrix should be employed. Otherwise the covariance matrix generally is more meaningful and has greater statistical appeal. Principle component analysis is the most popular multivariate statistical technique and has applications in many field.(Cf. Kendall 1939 in Agriculture; Craddock 1964 in Meteorology). For some applications in Ecology, it has been used by Orloci (1966) and Kershaw (1968 b).

3.2 Correspondence Analysis

This technique was proposed by Dr. M.O. Hill (1973, 1974) as a extention of Whittaker's gradient analysis (1967). Consider a p by n data matrix $\underset{\sim}{X}$, where the elements x_{ij} represent, for example, the number of occurrences of species i at site j (i=1,...,p; j=1,...; n). Whittaker's gradient analysis assumes that there is a well-marked physcial gradient,

and assigns scores t_j ($j=1, 2, \ldots, n$) comforming with the gradient to each of the n sites. For each species, one calculates the mean site scores s_i ($i=1, \ldots, p$) to indicate its preference

$$s_i = \sum_j t_j x_{ij} / x_{i.} \; ; \qquad \text{where } x_{i.} \text{ is the row total}$$

The derived species scores then are used to calculate a new set of the site scores and get the gradient analysis result t'_j of the sites

$$t'_j = \sum_i s_i x_{ij} / x_{.j} \; ; \qquad \text{where } x_{.j} \text{ is the column total}$$

Hill suggested the reciprocal average procedure which iterates the above procedure successively by replacing the old scores with new ones. It can be proved (Hill 1974) by matrix algebra that this procedure will converge and is equivalent to a singular value decomposition problem as follows.

Define $\underset{\sim}{R} \equiv \text{diag}(x_{i.})$ and $\underset{\sim}{C} \equiv \text{diag}(x_{.j})$, then $(\rho, \underset{\sim}{s}, \underset{\sim}{t})$ is a solution of correspondence analysis of $\underset{\sim}{X}$ if and only if

$$\rho \underset{\sim}{s} = \underset{\sim}{R}^{-1} \underset{\sim}{X} \underset{\sim}{t} \quad , \quad \rho \underset{\sim}{t} = \underset{\sim}{C}^{-1} \underset{\sim}{X}^T \underset{\sim}{s}$$

or

$$\rho (\underset{\sim}{R}^{\frac{1}{2}} \underset{\sim}{s}) = (\underset{\sim}{R}^{-\frac{1}{2}} \underset{\sim}{X} \underset{\sim}{C}^{-\frac{1}{2}})(\underset{\sim}{C}^{\frac{1}{2}} \underset{\sim}{t})$$

$$\rho (\underset{\sim}{C}^{\frac{1}{2}} \underset{\sim}{t}) = (\underset{\sim}{R}^{-\frac{1}{2}} \underset{\sim}{X} \underset{\sim}{C}^{-\frac{1}{2}})^T (\underset{\sim}{R}^{\frac{1}{2}} \underset{\sim}{s})$$

Where $\underset{\sim}{R}^{\frac{1}{2}}$ takes the square-root of $\underset{\sim}{R}$, solving for $\underset{\sim}{s}$, we have

$$\rho^2(\underset{\sim}{R}^{\frac{1}{2}}\underset{\sim}{s}) = (\underset{\sim}{R}^{-\frac{1}{2}}\underset{\sim}{X}\,\underset{\sim}{C}^{-\frac{1}{2}})(\underset{\sim}{R}^{-\frac{1}{2}}\underset{\sim}{X}\,\underset{\sim}{C}^{-\frac{1}{2}})^{T}(\underset{\sim}{R}^{\frac{1}{2}}\underset{\sim}{s}) \quad .$$

$\underset{\sim}{R}^{\frac{1}{2}}\underset{\sim}{s}$ is the eigen vector of a non-negative definite matrix of the form $\underset{\sim}{B}\,\underset{\sim}{B}^{T}$, ρ^2 is the eigen value of the solution. The solution of the axes will be ordered by their eigen-values. The unit vector corresponds to the eigen value 1 and is the trivial solution of the system.

In the same paper, Hill also proved several properties of correspondence analysis: that it is equivalent to Fisher's (1940) canonical analysis of the contingency table where ρ is the correlation of $\underset{\sim}{s}$ and $\underset{\sim}{t}$ with respect to matrix $\underset{\sim}{X}$, and that it is equivalent also to a special case of Hotelling's canonical correlation analysis. He indicated that correspondence analysis is a potential multivariate technique for ecological data. It also has several advantages over principle component analysis. Correspondence analysis uses a good species ordination to derive the site ordination which closely resembles that obtained by principle component analysis with standardized data. But for unstandardized data, principle component analysis tends to emphasize species-rich sites and lead to a less satisfactory species ordination. Another difference between these two techniques is that the axes derived from the correspondence

analysis are not orthogonal to each other, which probably conforms more with the ecological gradients where they are seldom orthogonal to each other.

3.3 Mode Analysis

For a multivariate p by n data matrix, where p is the number of variables and n is the sample size, cluster analysis attempts to devise a classification scheme for grouping the n individuals into g group, where the individuals assigned to the same group are "similar" but individuals from various groups are "different". The number of groups g and the characteristic of the group are to be determined by the classification scheme.

There are a number of different techniques in cluster analysis (Everitt 1974), each having advamtages and disadvantages depending on the kind of data and the purpose of the analysis. Mode analysis (Wishart 1969a) belongs to one of the hierarchical clustering techniques. It was developed to avoid the "chaining effect" of the nearest neighbor method (Williams, Lambert and Lance, 1966) and to detect the natural grouping of the individuals. The procedure is firstly to decide whether the data are multimodal. In the single variate case, this can be done by constructing a histogram and removing the low frequency regions (saddle regions) temporarily and assigning a group to each of the modal regions. Each point which falls in a saddle region is then

assigned to its nearest mode. But for the p-variate case, the construction of a histogram of this kind involves too many calculations. Wishart (1969a) suggested the use of the spherical regions instead of the rectanguloid regions to simplify the calculation. His algorithm is as follows.

(i) Select the frequency (density) threshold k, compute the similarity matrix for the individuals and determine the distance PD from each point to its kth nearest point. (The distance is calculated in the similarity matrix.)

(ii) Order the distances PD so that the smallest is first and use KP as an index. KP defines the order in which the data points become dense; point KP(1) has the smallest kth. distance PD(1) and is first to become dense when PMIN=PD(1), point KP(2) is second at PD(2), and so on.

(iii) Select distance thresholds PMIN from successive PD values, initialising a new dense point at each cycle. There are three possible fusion phases as the second and each subsequent dense point is introduced.

(a) The new point does not lie within PMIN of another dense point, then it initialises a new group.

(b) The point lies within PMIN of dense points

from one group only. The point, therefore, is fused directly to that group.

(c) The point falls within PMIN of dense points from separate group. Then the groups concerned are fused.

(iv) A note must be kept of the nearest-neighbor distance DMIN between dense points of different groups. When PMIN exceeds DMIN, immediately the two groups separated by DMIN is fused.

At each intermediate cycle of the above algorithm, there are different sets of groups classified, except the first and the last cycles where only one cluster is defined. Wishart suggested that the solution with the largest number of groups should be considered as the most significant and hence adopted as the natural grouping. The analysis may never reveal more than one group which indicates that the data are unimodal. There is a computer program available for mode analysis (Wishart 1969b). Appendix D is the flow chart for the mode analysis computer program. For a specified k, it calculates the array PD, distance in ascending order of each sample point to its kth nearest point. Array KP indicates the sample points corresponding to PD. The rest of the procedure simply follows the algorithm listed in page 13.

4. PRINCIPLE COMPONENT ANALYSIS

4.1 Raw Data

Principle component analysis (hereafter referred to as P.C.A.) is an eigenvalue extraction technique. Since the data matrix is fifty by one hundred and ninety-eight, a computer is necessary for the computation. The IMSL and SSP library were checked for their eigen-value extraction ability and the results conformed to six decimal places. In this project, two decimal places would be accurate enough. The following computation are based on IMSL library program.

The P.C.A. calculation could be based on either the correlation matrix or the covariance matrix. Both are computed. P.C.A. is applied to the correlation matrix of the untransformed data, the square-root transformed data, and the natural-logarithm transformed data. The first ten eigenvalues and their cumulative percentage of the total variation are presented in Table 1.

Table 1 Extracted eigenvalues from correlation matrix

	eigenvalues & cumulative percentages									
	1	2	3	4	5	6	7	8	9	10
x	6.18	4.16	3.35	2.66	2.20	1.98	1.81	1.76	1.66	1.63
	12	21	27	33	37	41	45	48	52	55
$\sqrt{x}$	7.96	4.76	3.78	2.90	2.24	1.86	1.66	1.60	1.54	1.40
	16	25	33	39	43	47	50	54	57	59
log(x+1)	8.13	4.77	3.79	2.89	2.28	1.85	1.68	1.61	1.51	1.39
	16	26	33	39	44	47	51	54	57	60

From the computation, we find that the use of correlation matrix for P.C.A. calculation is not very efficient. It takes about ten components to account for fifty percent of the total variation and almost all of the components to explain the whole system's variation. In this situation, the value of the analysis may be questioned, since it cannot achieve the purpose of reducing the dimensionality without considerable loss of information. From Table 1 we can also see that square-root transformation and logarithm transformation make little improvement to the results and are not much different from each other. While the main purpose of these transformations is to stabilize the variances, by using the correlation matrix we standardize the variables by their standard deviation and give them an equal scale. Thus, the results from original data and transformed data are not much different.

The computation results derived from covariance matrix are presented in Table 2. From the table we can see that the first three components all take more than ten percent of the total variation and after the sixth component they drop down to less than four percent. Again the square-root and logarithm transformations give similar results. In the following analysis, we will consider only the square-root transformation. It is very clear that the P.C.A. calculation results for this data set are much better if one employs the covariance matrix rather than correlation matrix. The sample sites can be ordered

Table 2 Extracted first ten eigenvalues from the covariance matrix

	eigenvalues & cumulative percentages									
	1	2	3	4	5	6	7	8	9	10
x	545	284	226	116	89	71	53	30	26	20
	35	52	67	74	80	84	88	90	91	92
$\sqrt{x}$	14.8	9.4	5.9	3.8	2.2	1.8	1.5	1.3	1.2	1.1
	27	45	55	62	66	70	73	75	77	79
log(x+1)	6.4	3.9	2.4	1.6	0.97	0.84	0.75	0.63	0.58	0.55
	26	42	51	58	62	65	68	71	73	75

by the obtained principle components. If we can attach some environmental gradients to the first few components and explain the system's variation satisficatorily by using them alone, then the dimensionality will be reduced and the redundancy in species will be removed by the P.C.A. transformation.

We now can plot the sample points against the derived first few principle components by projecting them into the 2-dimension plane. (e.g. component 1 vs. 2, component 2 vs. 3, etc.) Based on the pre-determined Braun-Blanquet associations of the sample sites, we can evaluate the performance of the different P.C.A. calculations. Fig. 4.1 is the square-root transformed data plots on the first 2 principle component axes derived by correlation matrix. The points (sample sites) are

Fig. 4.1

X axis-- first principle component axis

Y axis-- second principle component axis

Square-root transformed data

Use the correlation matrix

Origin and scale are arbitary

(≠ denotes coincident points)

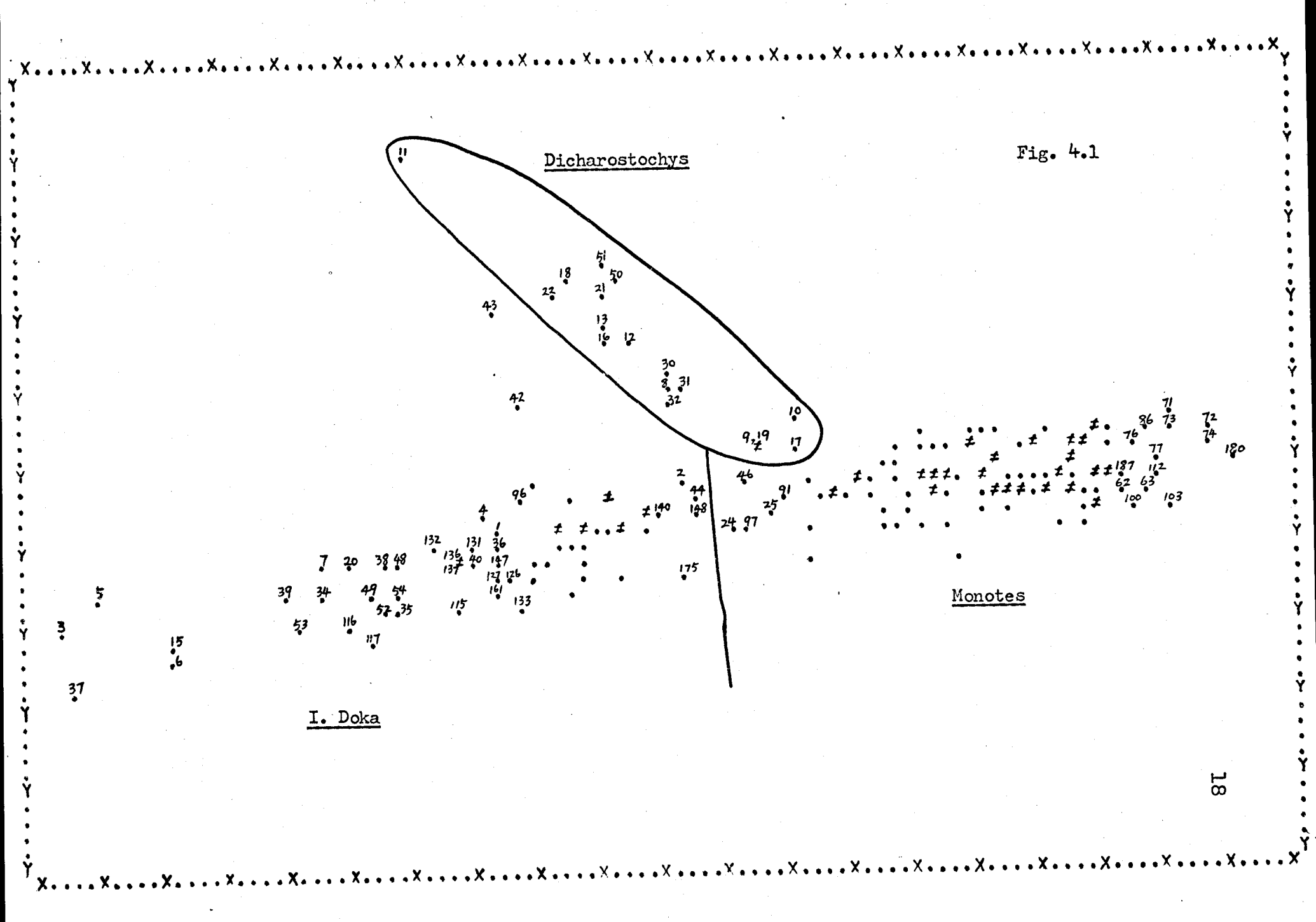


Fig. 4.1

all mixed up and can be divided barely into the three groups where each group contain species: Isoberlinia doka, Monotes, and Dichrostachys respectively. The plots on the other component axes derived by correlation matrix yield the same sort of unidentifiable result.

Fig. 4.2 is the plot of original data on the first two component axes derived by covariance matrix. In the middle of the graph, there is an obvious vacancy between the different groups. By listing and examining the sample sites under the order of the magnitude of their first principle component, we may conclude that the first component axis is an axis showing the species variation, where the left half is sample sites with species Isoberlinia doka and the right half is sites with Monotes. It takes account of thirty-five percent of the total variation and divides the sites into two different groups. Except this, however, we cannot recover any of the Braun-Blanquet associations. A study of the plots on other sets of principle components also reveals the same difficulty of recovering the associations. For P.C.A. to be effective, there should be no complete discontinuity between sample extremes. This can be illustrated by the following two-dimension situation.

Fig. 4.2

X axis-- first principle component axis

Y axis-- second principle component axis

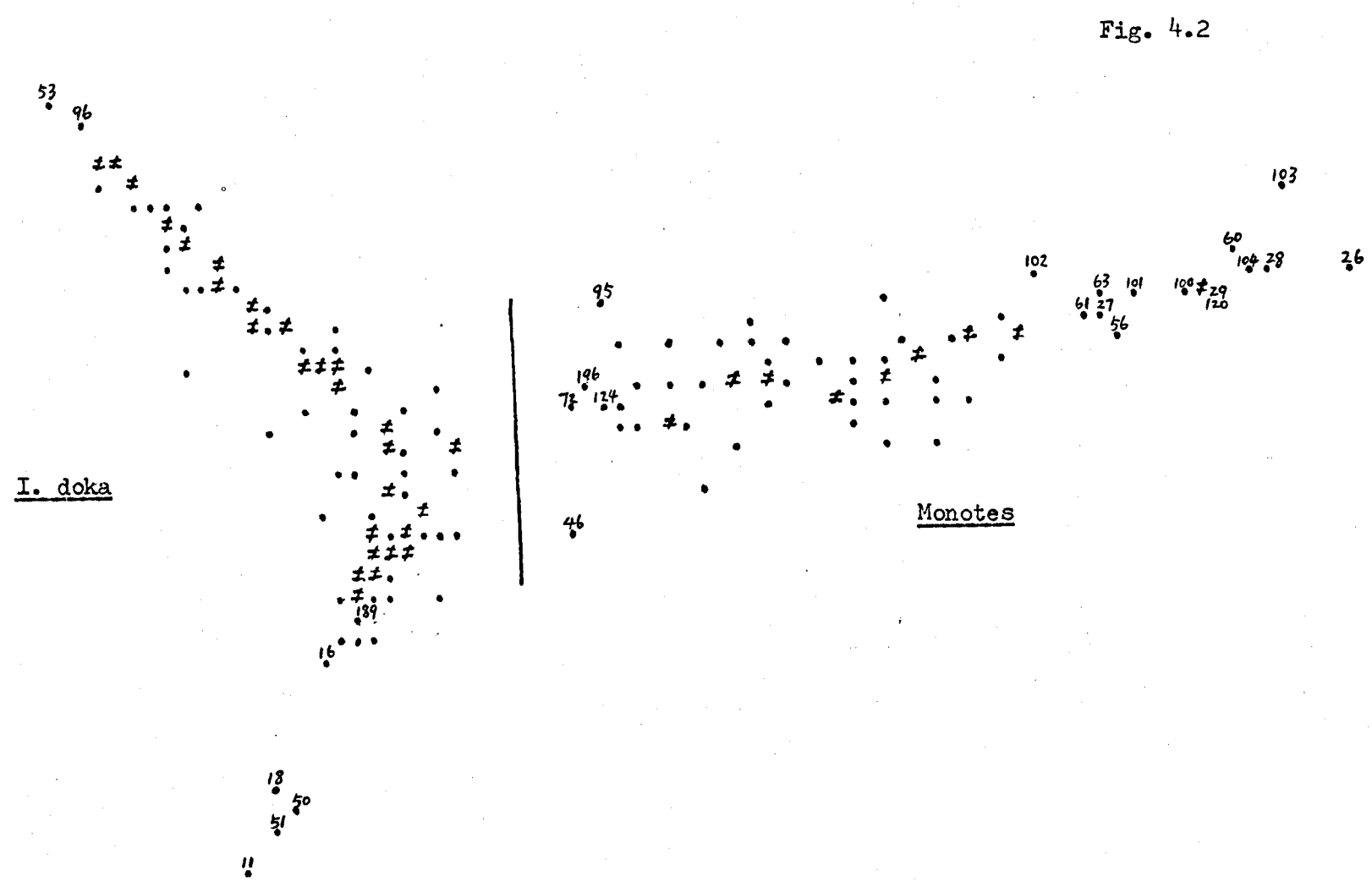
Untransformed data

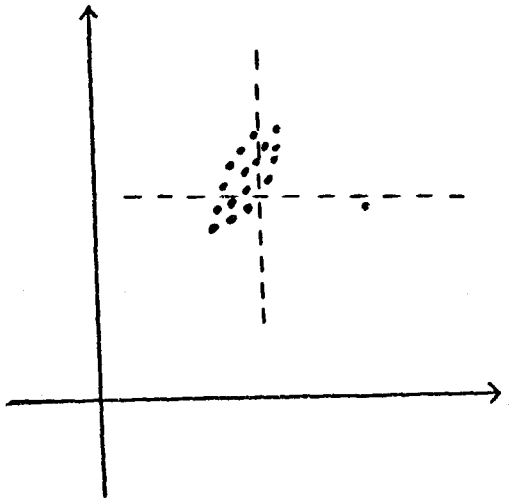
Use the covariance matrix

Origin and scale are arbitarary

($\neq$ denotes coincident points)

Fig. 4.2





Unsatisfactory P.C.A. transformation due to the extreme point

The use of frequency measurement probably introduces sample extreme points which create the distortion mentioned above and make the P.C.A. result unsatisfactory. The square-root transformation attempts to reduce the effect of extreme points.

Fig. 4.3 is the plot of square-root transformed data on the first two principle components derived by covariance matrix. By identifying the sample sites, we see that associations 2, 3, 8, 12, 13, 14, 15, and 16 can be divided relatively well. The associations 4, 5, 6, with the common species Isoberlinia doka are mixed in the left hand side, and associations 9, 10, 11 with species Monotes are in the right side. Fig. 4.4 is the plot on the first and the third component axes. It separates several associations as in Fig. 4.3. We can distinguish, moreover, beteen association 5 and associations 4 and 6. Those associations with the common species Monotes are mixed together still on the right side

Fig. 4.3

X axis-- first principle component axis

Y axis-- second principle component axis

Square-root transformed data

Use the covariance matrix

Origin and scale are arbitrary

($\neq$ denotes coincident points)

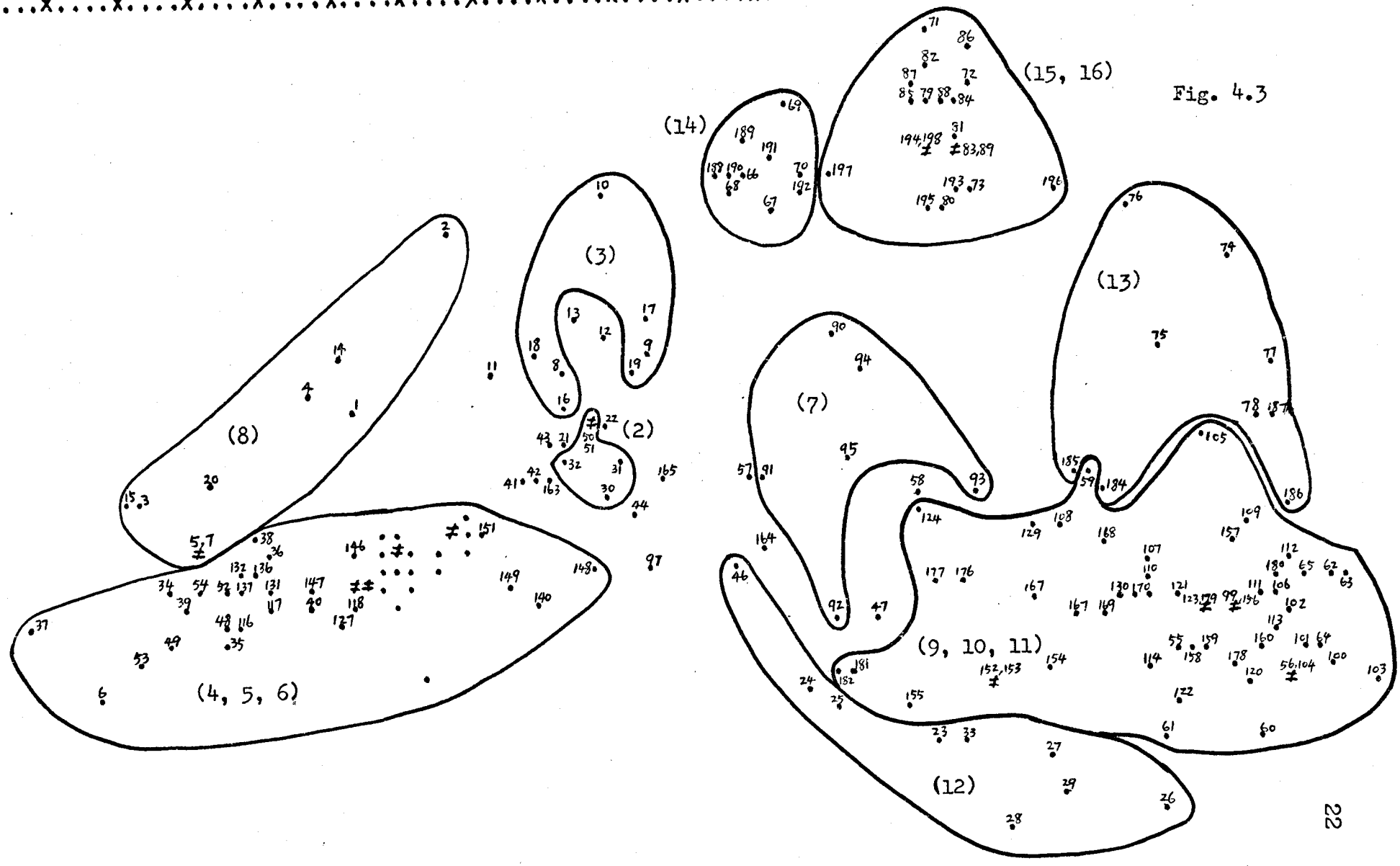


Fig. 4.3

Fig. 4.4

X axis-- first principle component axis

Y axis-- third principle component axis

Square-root transformed data

Use the covariance matrix

Origin and scale are arbitrary

($\neq$ denotes coincident points)

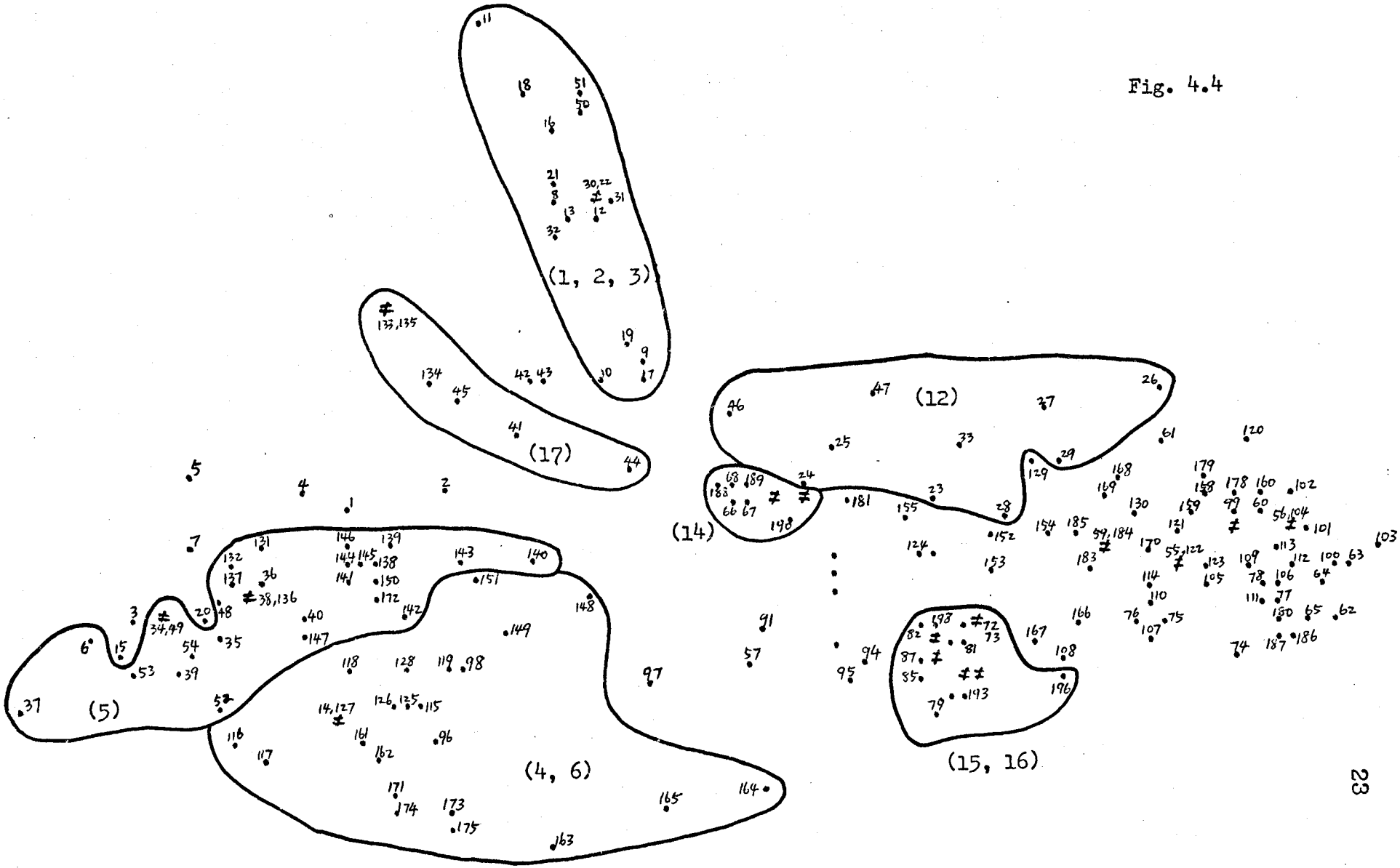


Fig. 4.4

and cannot be divided properly.

The above technique of plotting the sample sites by projection to two of their principle component axes has serious drawbacks. When we project the spatial clusters onto the plane, the structure of the data is completely distorted and obscured. Hence, sometimes precious information concerning the real structure is concealed. One means to overcome this disadvantage is to attempt to increase the plot to three dimensions. Kershaw (1968b) devised a method to project the spatial points onto the three plane surface. For a small sample size, this would be a better approach, since we can reconstruct the position of the point in three dimensional space and gain a much clearer picture of the data structure. But for a large sample, it has little advantage over the plotting of three two-dimensional projections separately. Perspective drawing (Kershaw & Shephard, 1972) is another graphic technique which illustrates the sample points hanging on the three dimensional space. But it also has the shortcoming of deciding the relative depth of points in the graph. By drawing the consecutive rotation of the graph or using an interactive displaying device to show the rotation, one can overcome this problem and gain a clearer picture of the data structure. This is, however, much more expensive and time-consuming.

Among the above computations, square-root transformed data with covariance matrix seems to have the best result for P.C.A. of this frequency measured ecological data set. A comparison with Dr. Kershaw's previous analysis, in which he employed the Domin scale measurement (Kershaw 1968b), shows that frequency measurement has not given much improvement over the Domin scale measurement in return for the amount of extra work required for sample collection and treatment. Part of this probably may be attributed to the fact that the Braun-Blanquet associations are derived by using the Domin scale measurement.

4.2 The Use of the Most Abundant Species

From section 4.1, we learn that due to the larger data variation introduced by the use of the frequency measurement, the P.C.A. can not yield good results. Examining the original data list, we find out that there are species which are present in less than six sites out of the total of one hundred and ninety-eight sample sites. These rare species sometimes present more noise than information for they can form the extreme points in the sample and hence distort the P.C.A. results. Comparing the total number of species and the presence of each species in the sites, one may discard fifteen species which are found in no more than twenty-five sites and the total number of

individual species in less than sixty. The discarded fifteen species are 1, 2, 3, 5, 6, 7, 8, 14, 18, 24, 26, 27, 34, 43, and 50. (The corresponding species names are in Appendix.)

P.C.A. has been performed on the reduced data matrix. Both original and square-root transformed data are calculated using the covariance matrix. The first ten eigenvalues and their cumulative percentage of total variation are in Table 3.

Table 3 Extracted eigenvalues from the reduced data set

	eigenvalues and cumulative percentage									
	1	2	3	4	5	6	7	8	9	10
x	545	282	226	116	89	71	53	30	26	19
	35	53	67	75	81	85	89	91	92	93
$\sqrt{x}$	14.7	9.4	5.7	3.8	2.2	1.8	1.5	1.3	1.1	1.0
	28	46	57	65	69	72	75	78	80	82

Comparing Table 2 and Table 3, we are aware that they are very similar. The sample sites are plotted against the derived components as before in order to examine the interpreting ability. Fig. 4.5 is the plotted original data on the first two principle component axes. It is in effect identical to Fig. 4.2. The plots on the other sets of principle components also are identical to their correspondence derived in the previous section. This means that the discarding of the fifteen rare species has no effect on

Fig. 4.5

X axis-- first principle component axis

Y axis-- second principle component axis

Use thirty-five species

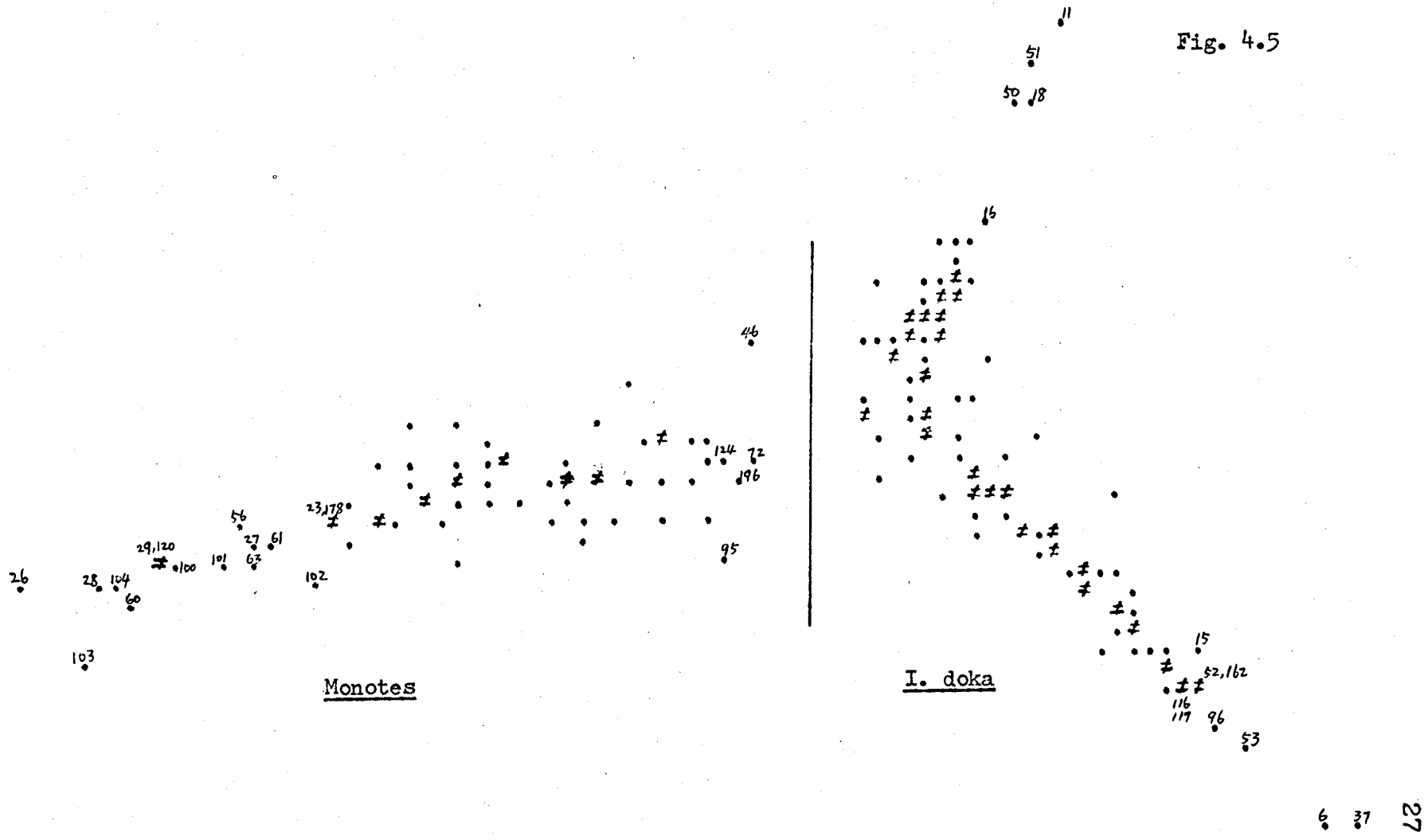
Untransformed data

Use the covariance matrix

Origin and scale are arbitrary

($\neq$ denotes coincident points)

Fig. 4.5



the P.C.A. ordination of the sample sites. Fig. 4.6 is the square-root transformed data plotted on the first two principle component axes and also is similar to Fig. 4.3. Thus for the efficient ordination of the sample sites by P.C.A., thirty-five species could achieve the same results as fifty species. The information used in ordination appears to be carried by the abundant species. Although discarding the rare species has not improved the results, it certainly could simplify the data collection and editing. The fact that P.C.A. is insensitive to rare species has been pointed out by Austin & Greig-Smith (1968). For their ecological data, the efficient P.C.A. ordination could be attained by using just twenty-five percent of the total species.

4.3 Principle Component Analysis after Mode Analysis

P.C.A. works most effectively when the data cluster approximately conforms to a multinormal distribution. Sometimes, however, the data present a dumb-bell shaped distribution, which forms two or more than two clusters. In this case, the first principle component of P.C.A. certainly could find the long axis of the dumb-bell and reveal this variation. Since the second and the subsequent components are constructed in the hyperspace common to the clusters and reflect some characteristics of the clusters, they usually do not have a clear meaning. This situation may be shown

Fig. 4.6

X axis-- first principle component axis

Y axis-- second principle component axis

Use thirty-five species

Square-root transformed data

Use the covariance matrix

Origin and scale are arbitrary

($\neq$ denotes coincident points)

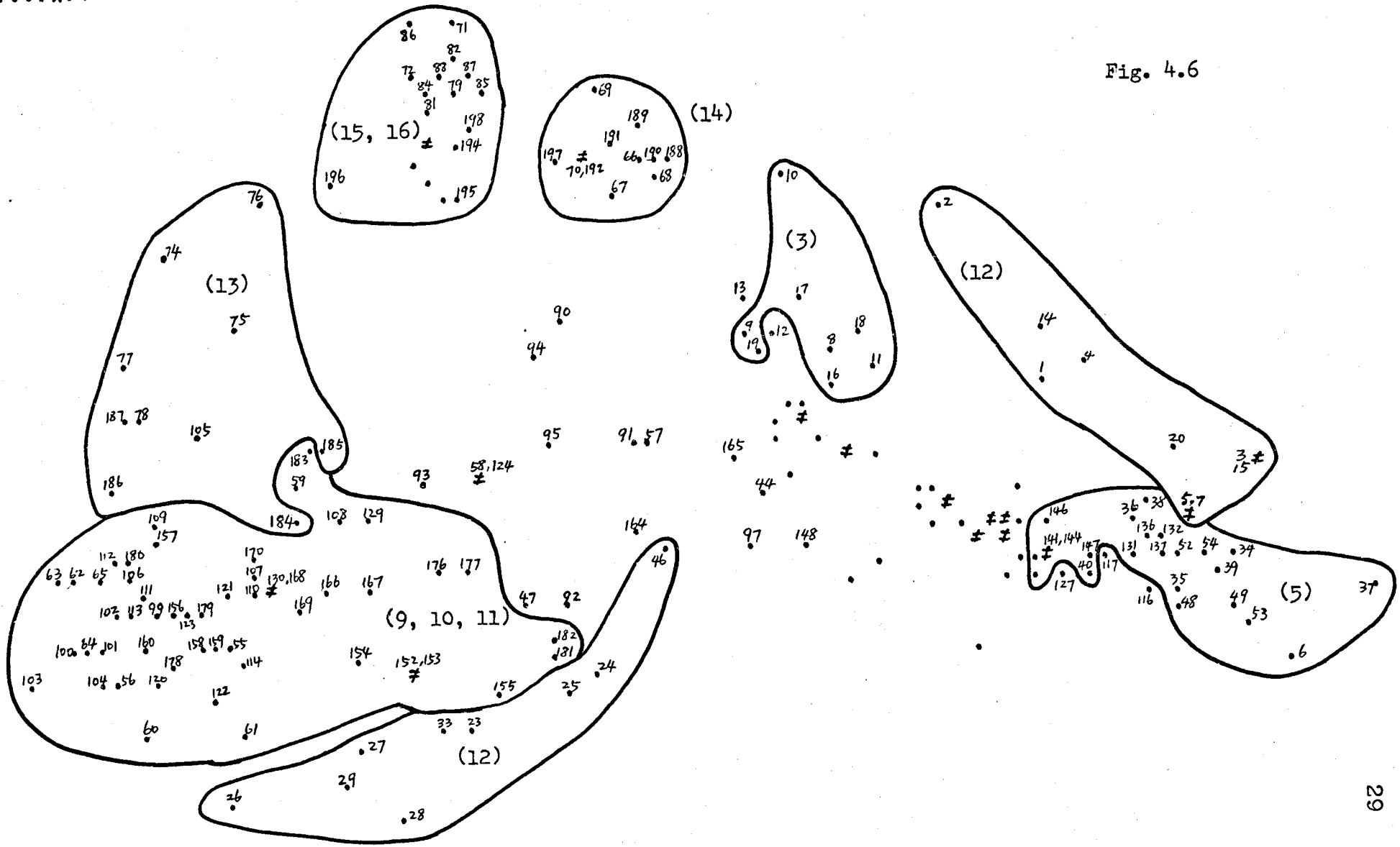
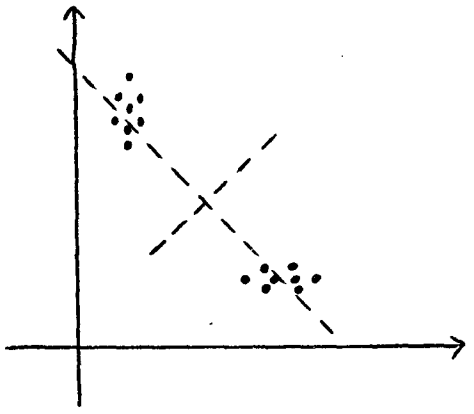


Fig. 4.6

by the following extreme case in two-dimensions.



The second component of the P.C.A. transformation is meaningless

From the previous analysis, it illustrated that the data probably have two or more clusters and hence the P.C.A. could not work efficiently. To overcome this problem, I employed the mode analysis first in order to reveal the natural grouping. Then I subject each group to the P.C.A. calculation when it is necessary. Using the results of mode analysis in section 6.1, which form six groups, the biggest two of them are used here for P.C.A. The results are as follows.

The first group has sixty-two sample sites, all but two of these sites have Isoberlinia doka as the common species. The number of species measured is reduced to forty-eight, since the other two species are not present in any of the sixty-two sites. The eigenvalues extracted from the covariance matrix are in Table 4.

Table 4 Extracted eigenvalues
for the sixty-two sites

	eigenvalues and cumulative percentage									
	1	2	3	4	5	6	7	8	9	10
x	274	256	131	51	46	37	21	17	13	9
	30	58	73	78	83	87	90	92	93	94
$\sqrt{x}$	8.7	4.2	3.3	2.8	2.3	1.7	1.5	1.2	1.0	0.9
	25	36	46	54	60	65	69	73	76	78

The results of P.C.A. with original data are not satisfactory. When one plots the components graph and checks the data list, it is revealed that the first and the second components are indicators of the species' variation of Isoberlinia doka and Terminalia respectively. The plots on the other sets of components show that the sample sites are gathered as a central mass and can not be divided into groups comforming to the Braun-Blanquet associations. For the square-root transformed data, the second component divides out sample sites 42, 43 from the other sites. These two sites do not belong to any of the associations and the species presented in them are very different from the other sixty sample sites. Hence one could consider them as outliers. Fig. 4.7 shows the sixty-two sample sites plotted on the first and the third components. It can be divided into four group conforming to the Braun-Blanquet associations 4, 5, 6, and 8.

Fig. 4.7

X axis-- first principle component axis

Y axis-- third principle component axis

Use sixty-two sample sites

Square-root transformed data

Use the covariance matrix

Origin and scale are arbitrary

($\neq$ denotes coincident points)

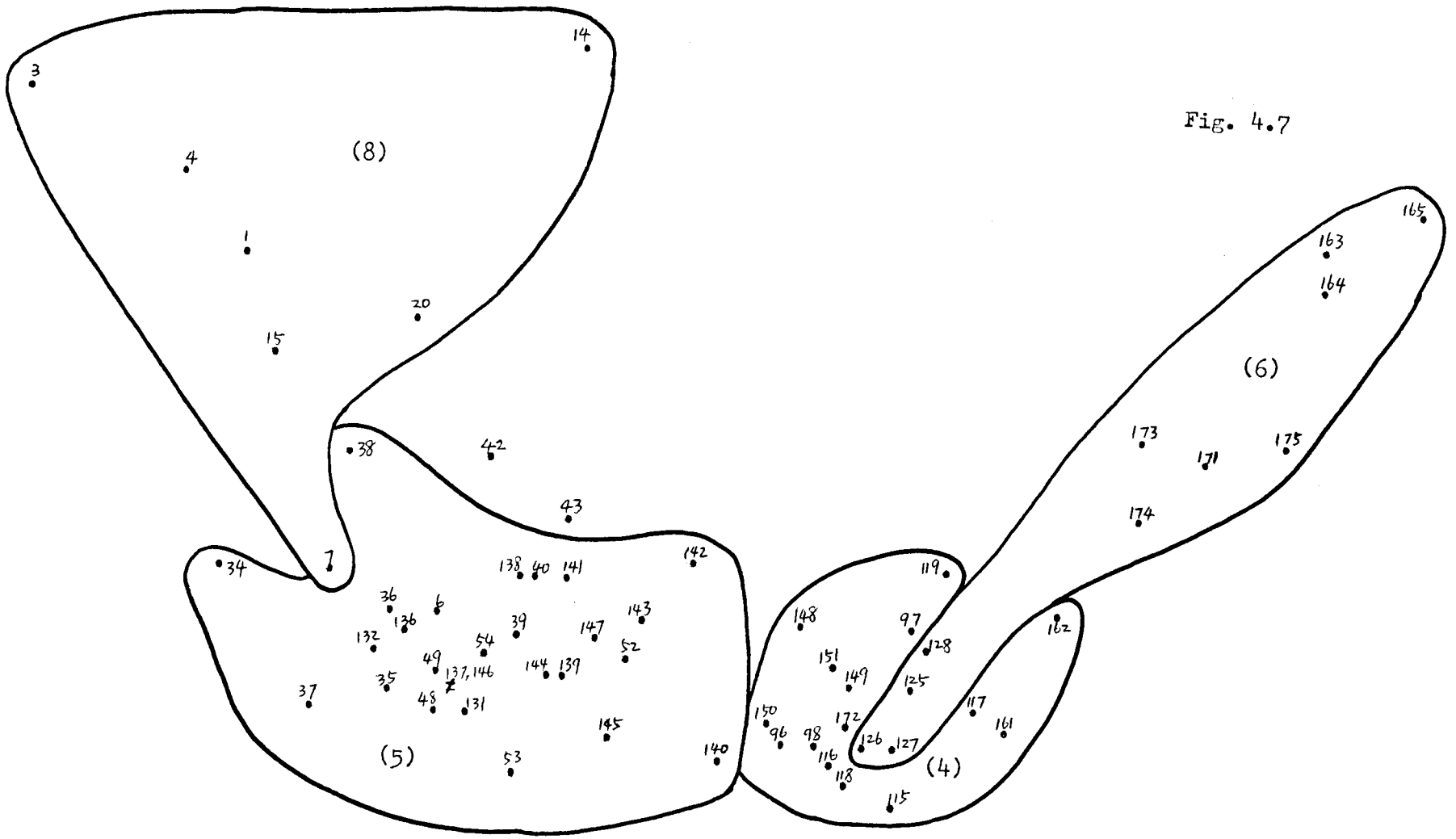

(8)
(6)
(5)
(4)
3
14
4
1
20
15
38
42
43
34
7
142
138
40
141
119
36
6
136
39
143
148
97
132
54
147
52
128
151
49
137,146
144
139
149
35
125
37
48
131
117
150
172
145
161
96
98
126
127
140
116
53
118
115
165
163
164
173
175
171
174
162

Fig. 4.7

The second group has eighty-seven sample sites. Most of the sites contain species Monotes. The measured species number is reduced to forty-five. The P.C.A. on untransformed data also is distorted by the species variation. Fig. 4.8 is the eighty-seven sample sites plotted against the first two component axes derived from the covariance matrix with the square-root transformed data. It can divide out the associations 7, 12, 13, and 17, but associations 9, 10, and 11 are mixed in the left hand side and thus are difficult to divide properly. On the whole, however, the P.C.A. after mode analysis leads to a rather satisfactory result.

Fig. 4.8

X axis-- first principle component axis

Y axis-- second principle component axis

Use eighty-seven sample sites

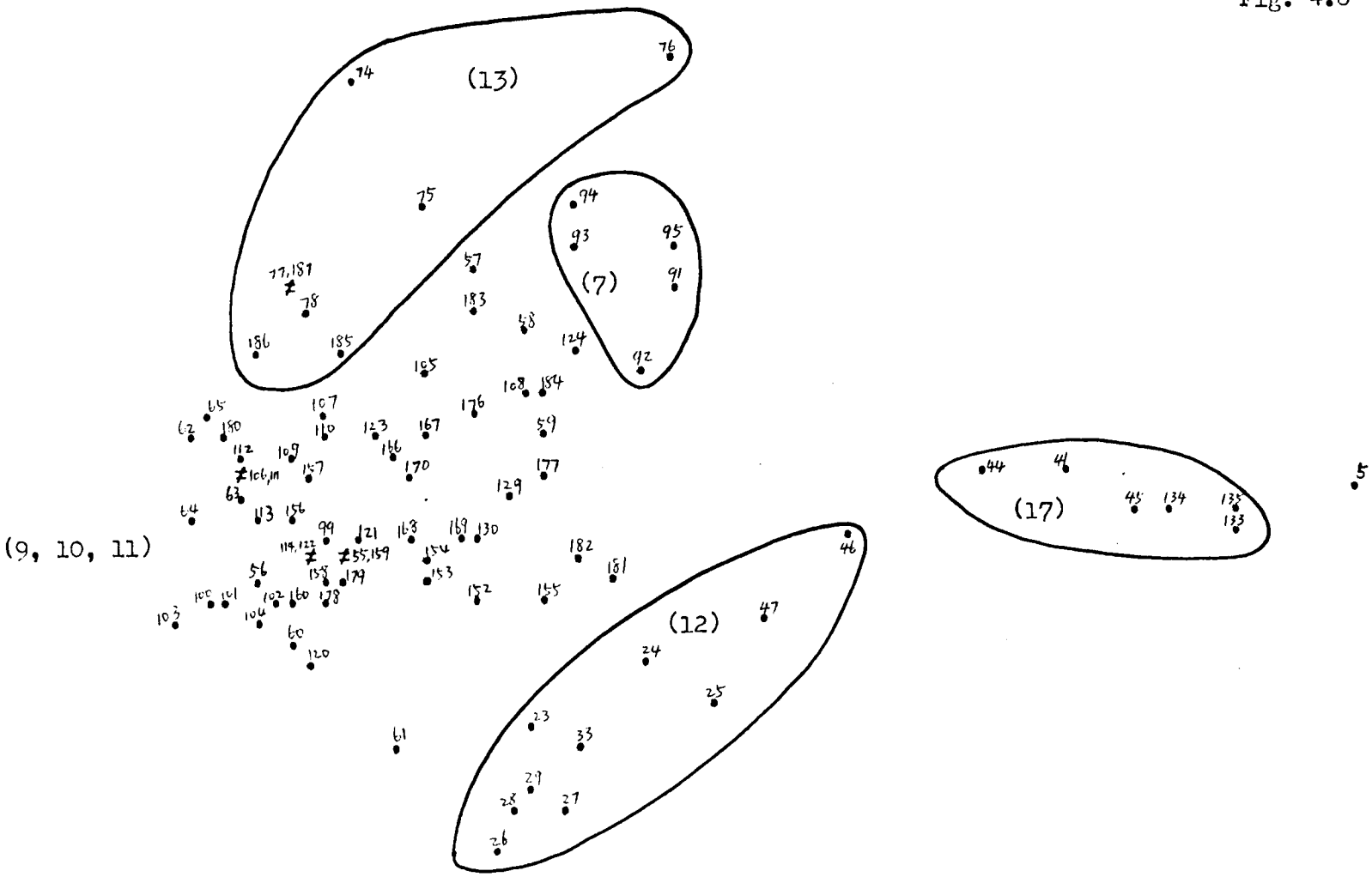
Square-root transformed data

Use the covariance matrix

Origin and scale are arbitrary

($\neq$ denotes coincident points)

Fig. 4.8



5. CORRESPONDENCE ANALYSIS

5.1 Raw Data

The most significant feature of correspondence analysis as an ordination technique is its duality of sites and species scores, a property shared by no other technique. For every set of sites scores, there corresponds a unique set of species scores and vice versa, where the correspondence between them is determined by the correlation ρ. The correlation could be used as a measurement of the ability of species scores to order the sites. The formula in section 3.2 illustrated that we may extract as many as r=min(p,n) sets of scores. But in practical usage only the sets of scores with high correlations are of interest.

Applying the correspondence analysis to the untransformed data, the square-root transformed data and the presence-absence data, the first ten nontrivial correlations were derived. They are shown in Table 5.

Table 5 First ten correlations between sites and species scores

	1	2	3	4	5	6	7	8	9	10
x	0.89	0.80	0.72	0.65	0.58	0.55	0.51	0.48	0.43	0.41
$\sqrt{x}$	0.79	0.67	0.59	0.54	0.46	0.42	0.36	0.35	0.33	0.31
(+,-)	0.72	0.56	0.49	0.44	0.38	0.37	0.35	0.33	0.32	0.30

The derived sites scores and species scores are in the form of standardized eigenvectors. Since correlation is a scale-free measurement, we can rescale the scores into more convenient units, such as from score 0 to score 100. The derived species scores or site scores can be used as rectanglar coordinates to plot the species or sites for further studies and interpretation in environmental terms. Techniques such as overlaying the environmental data for each sampling site over the corresponding points on the plot are employed generally for this purpose. There are three broad types of derived score axes--seriation, nodal, and polynomial axis. The seriation axes are those which arrange the sites (or species) according to some natural gradient in the data structure from one end to the other evenly. The nodal axes have a clear gap in the score seriation. Hence they divide the sites (or species) into some natural groupings. The polynomial axes depend on the other axes and thus form polynomial relations with them.

From section 4.2, it is shown that principle component analysis pays very little attention to the rare species. The main information for the calculation of the principle component analysis is supplied by the abundant species. Hill (1973) considered this as a disadvantage of principle component analysis and pointed out that correspondence analysis can improve this condition. By an

examination of the derived species scores, it has been ascertained that the correspondence analysis stresses the rare species too much: in the ordination rare species usually obtain the extreme scores, that is, the highest or lowest scores. In the first twenty sets of species scores derived from untransformed data, there are seventeen sets with those fifteen rare species mentioned in section 4.2 as extreme scores. The square-root transformed data and presence-absence data ~~they~~ both have nineteen sets which have the fifteen rare species as extreme scores. This reveals that the correspondence analysis, in contrast to principle component analysis, has the opposite property. Although this is not necessarily a disadvantage, it certainly implies the difficulty of interpreting ordination results. Since the correspondence analysis is the successive calculation of gradient analysis, we might think that if we start with a set of species or site scores which conform with a specific physical gradient and carry on the correspondence analysis calculation until the scores converged, then the derived final scores should be related with this physical gradient and have its environmental meaning. But it is not the case, for no matter what scores are used to start the calculation, they converge to the maximum non-trivial solution of the correspondence analysis. Employing the same notation as section 3.2, the reciprocal averaging procedure is as

follows.

$$\underset{\sim}{s}^{(i)} = \underset{\sim}{R}^{-1}\underset{\sim}{X}\,\underset{\sim}{t}^{(i)} \quad , \qquad \underset{\sim}{t}^{(i+1)} = \underset{\sim}{C}^{-1}\underset{\sim}{X}^{T}\underset{\sim}{S}^{(i)}. \qquad (i=0,\ 1,\ 2,\ldots)$$

Viewed from the point of species scores, the above procedure can be represented as

$$\underset{\sim}{s}^{(i+1)} = \underset{\sim}{R}^{-1}\underset{\sim}{X}\,\underset{\sim}{C}^{-1}\underset{\sim}{X}^{T}\underset{\sim}{s}^{(i)} \quad , \quad (i=0,\ 1,\ 2,\ldots)$$

where $\underset{\sim}{s}^{(i+1)}$ is the new species score. The final solution would converge to the maximum nontrivial solution of the eigenvector. This characteristic shows that the correspondence analysis is just an ordination method which has a dual property of the site scores and species scores measured by their correlation. The interpretation of the ordination results have to depend on further analysis such as a knowledge of the environment or the species itself. Since there is no information concerning species readily available, we turn again to the Braun-Blanquet association as a basis to compare the different results obtained.

From Table 5 we learn that the first three correlations for untransformed data are 0.89, 0.80, and 0.72 respectively, all of which are relatively high. Examining the derived site scores reveals that they are ordered according to some underlying structure. For example, the first axis divides the sample sites into three groups which are sites with Dichrostachys, sites with Isoberlinia doka, and sites

with Monotes. The second axis has a polynomial shape on the first axis. This is shown in Fig. 5.1. By identifying the sample sites on the plotted graph, we see that associations 1, 2, 3, 8, and 17 can be divided out relatively well. In the lower left associations 7, 9, 10, 11, 12, 13, 14, 15 and 16 are mixed together. The upper left part includes associations 4, 5, and 6. Fig. 5.2 shows the sample sites plotted along the second and the third sets of site scores. It can divide out associations 3, 8, 14, 15, 16, and 17, but the rest are mixed together. The plot on the first and the third sets of site scores does not provide more information on dividing of associations. If there is environmental information available, then we can continue this kind of examination for the other sets of scores and attach some meaning to them.

For the square-root transformed data, the species with extreme scores also are those rare species, but there are few changes in the middle part of ordination. The correlations between species and site scores are smaller than those of the untransformed data. Plotting the sample sites against the derived score sets shows shapes similar to those obtained with the untransformed data. This is because correspondence analysis does not require data centering within species in the same manner as required in P.C.A.. Thus the transformation has less influence on

Fig. 5.1

X axis-- site scores with highest correlation

Y axis-- site scores with second highest correlation

Untransformed data

Origin and scale are arbitrary

($\neq$ denotes coincident points)

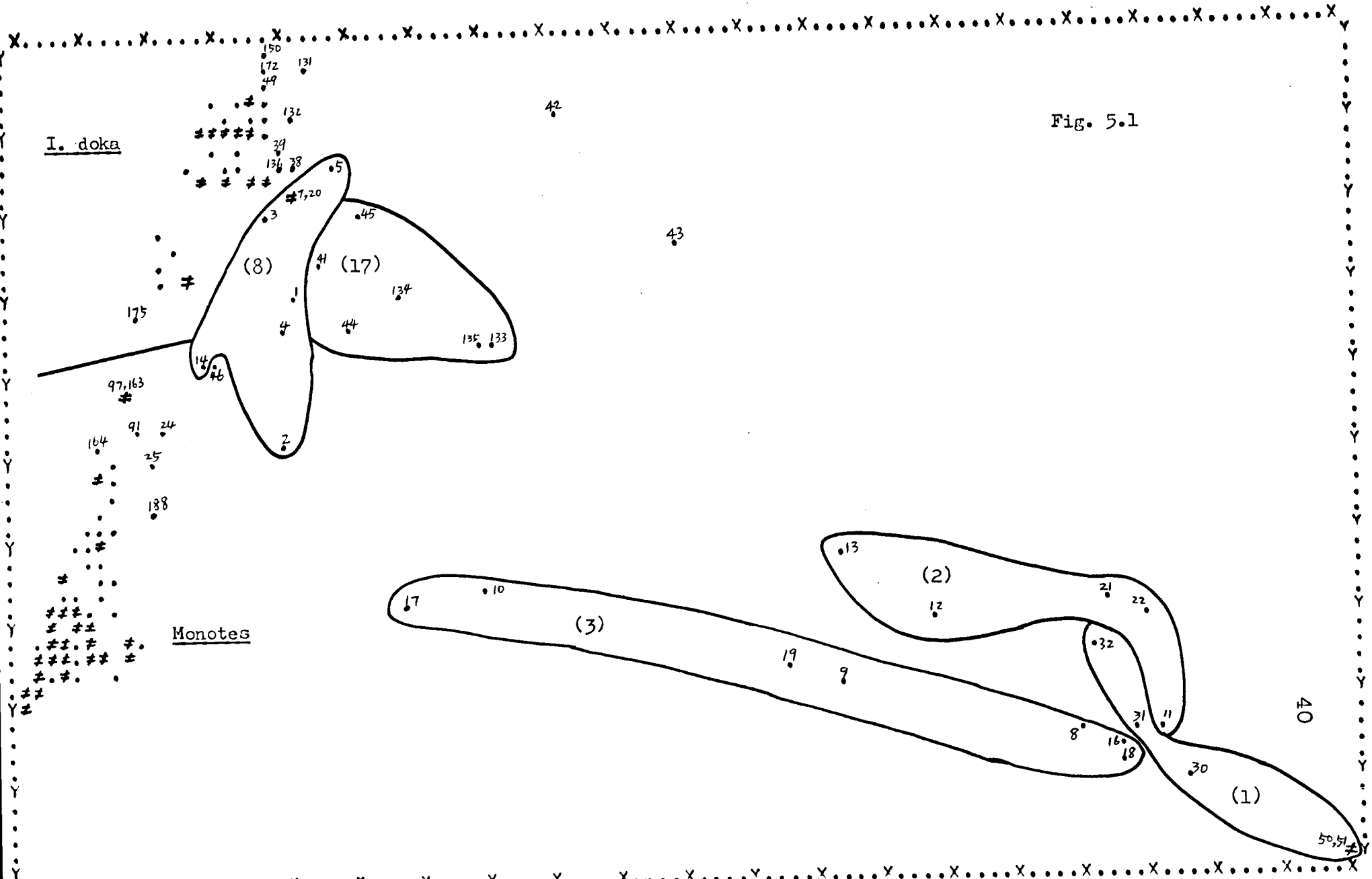

I. doka
Monotes
(8)
(17)
(3)
(2)
(1)
150
172
131
49
132
42
39
136
38
5
7,20
3
45
43
41
1
134
175
4
44
135
133
14
46
97,163
91
24
164
25
188
7
13
10
17
12
21
22
32
19
9
31
11
8
16
18
30
50,51

Fig. 5.1

Fig. 5.2

X axis-- site scores with second highest correlation

Y axis-- site scores with third highest correlation

Untransformed data

Origin and scale are arbitrary

(≠ denotes coincident points)

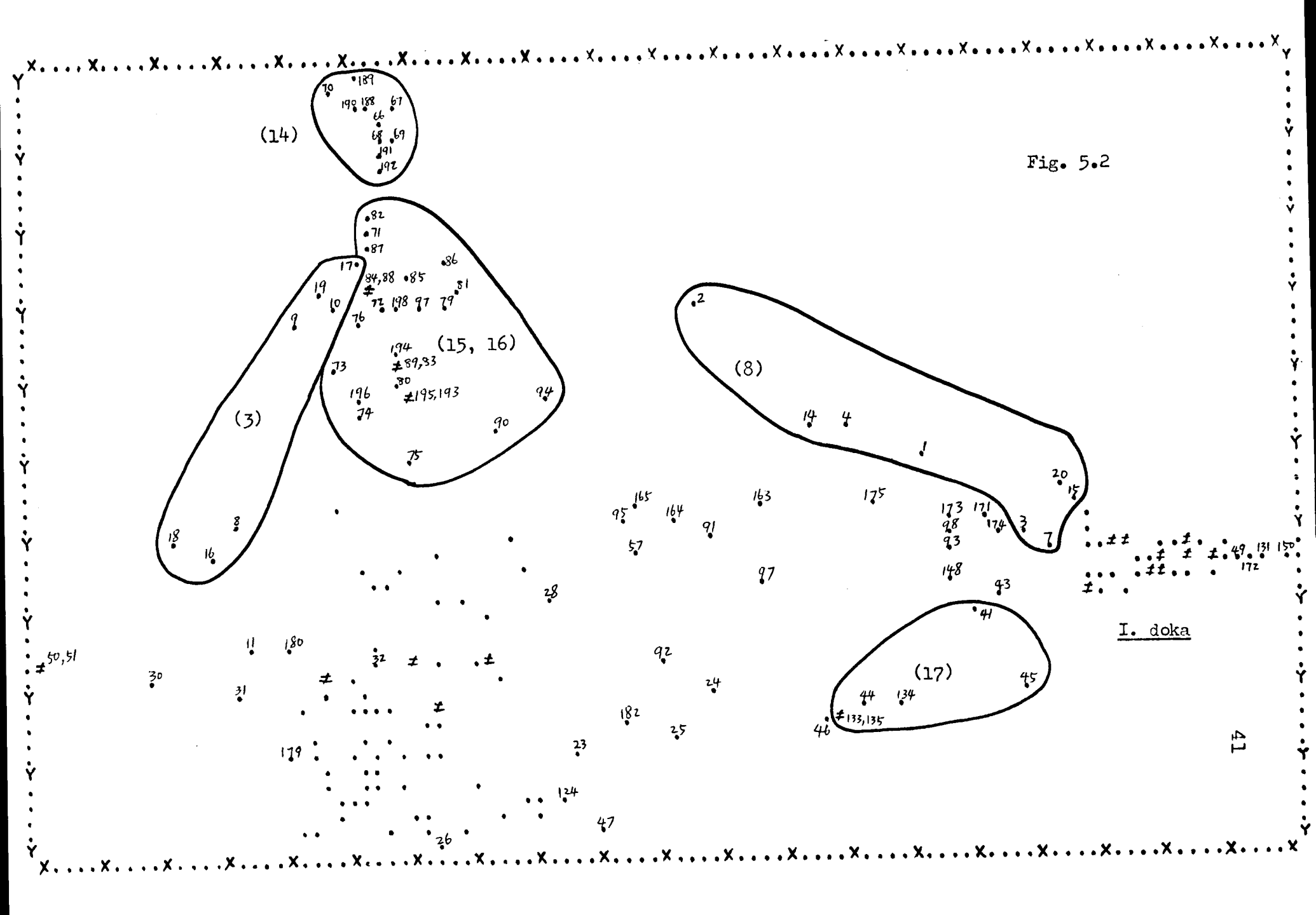


Fig. 5.2

the results. Fig. 5.3 shows the sample sites plotted against the first two sets of sites scores. It has the same shape as Fig. 5.1, but it enlarges the differences between site scores and leads to a more comprehensible graph. Identifying the sites, we can see that each site falls fairly well in the appropriate associations according to the ordered scores. It clearly divides association 5 from associations 4 and 6. The associations 14, 15, and 16 with Detarium as common species are grouped in the lower right part, above them are the associations with Monotes formed into one large group.

Performing the correspondence analysis calculation on the presence-absence data instead of on the frequency data to examine its performance shows that the rare species are even more emphasized in this case. The first three sets of species scores actually start with species 1, 2, 3, and 7. The correlations between site scores and species scores also are reduced: the maximum nontrivial one being just 0.72. For the presence-absence data, the most valuable information is all gathered in the first two sets of scores. If we plot the sites against the remaining score sets, they just show sample sites huddled in one large cluster. Thus very little is revealed about the data structure. Fig. 5.4 is the sample sites plotted on the first two sets of site scores. Identifying the sample sites illustrated that it is very similar to Fig. 5.3. Although the ordination is less satisfactory

Fig. 5.3

X axis-- site scores with highest correlation

Y axis-- site scores with second highest correlation

Square-root transformed data

Origin and scale are arbitrary

($\neq$ denotes coincident points)

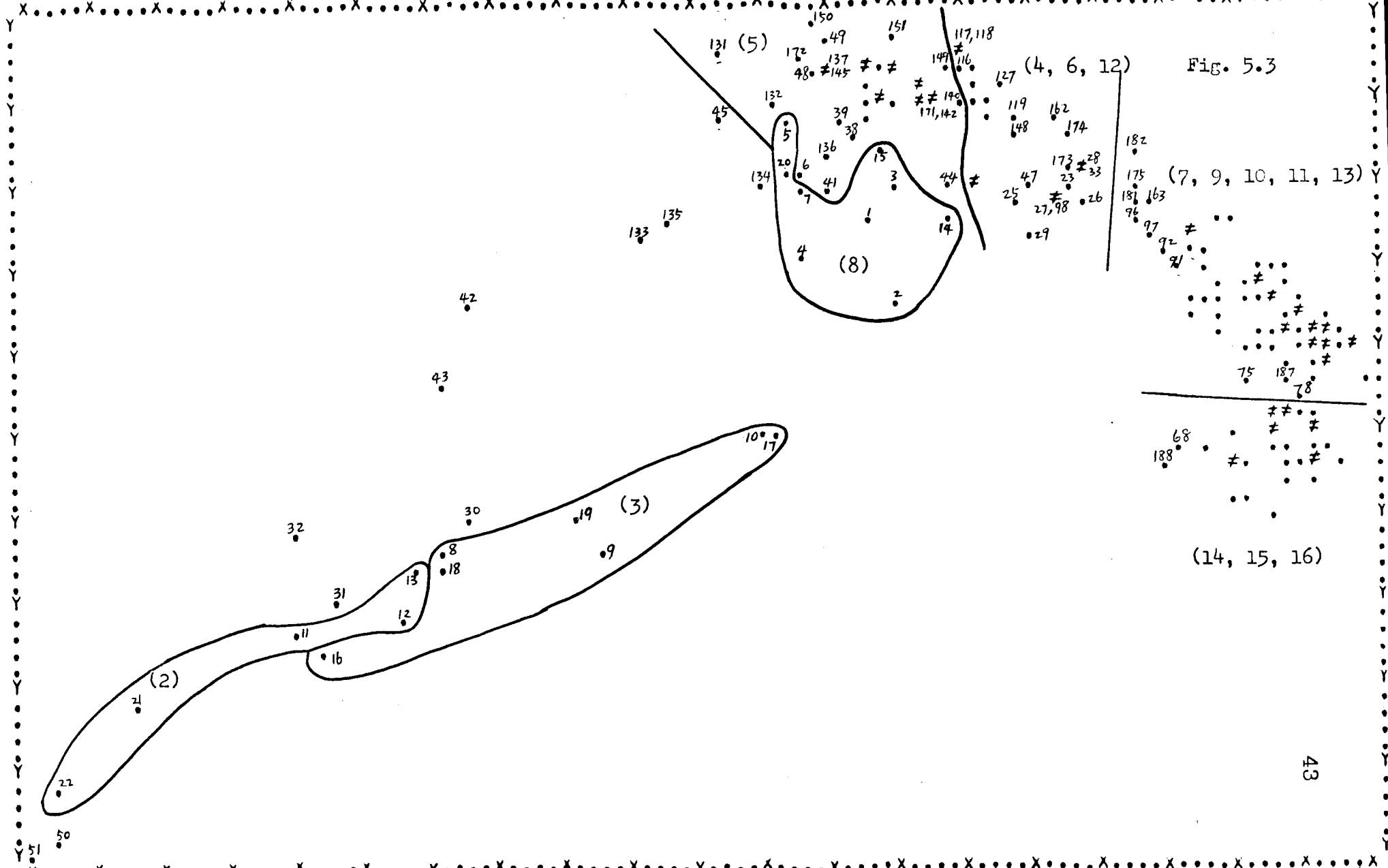


Fig. 5.3

Fig. 5.4

X axis-- site scores with highest correlation

Y axis-- site scores with second highest correlation

Presence-absence data

Origin and scale are arbitrary

($\neq$ denotes coincident points)

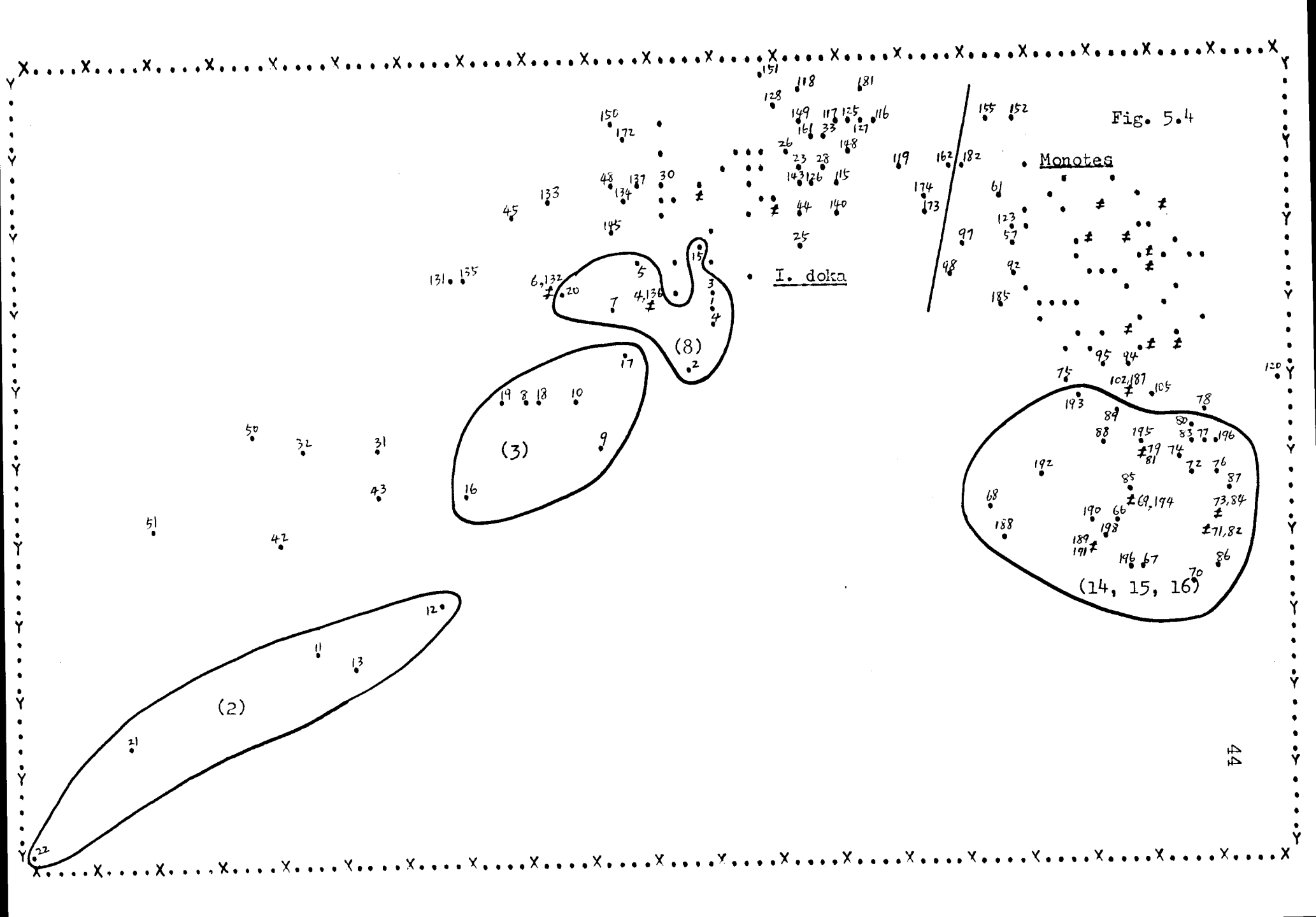


Fig. 5.4
Monotes
I. doka
(2)
(3)
(8)
(14, 15, 16)

than that derived from square-root transformed data, it certainly indicates that correspondence analysis is not sensitive to the scale of measurement.

5.2 The Use of the Most Abundant Species

Since the correspondence analysis stresses the rare species, we next discard the fifteen rare species mentioned in section 4.2 and use the remaining thirty-five abundant species to calculate the correspondence analysis and compare the results with the previous results. The first ten non-trivial correlations derived are shown in Table 6.

Table 6 First ten correlations
derived from the reduced data set

	1	2	3	4	5	6	7	8	9	10
x	0.89	0.80	0.72	0.64	0.57	0.55	0.50	0.43	0.41	0.37
$\sqrt{x}$	0.77	0.66	0.59	0.44	0.42	0.42	0.36	0.31	0.30	0.28
(+,-)	0.69	0.54	0.42	0.37	0.34	0.33	0.32	0.30	0.28	0.27

The derived correlations are very similar to those shown in Table 5. The ordinated scores are also similar. The discarding of rare species also has the effect of enlarging the difference between scores. Fig. 5.5 shows the sample sites plotted against the first two sets of site scores of the untransformed data. Fig. 5.6 is the plot of the first two sets of site scores for presence-absence data.

Fig. 5.5

X axis-- site scores with highest correlation

Y axis-- site scores with second highest correlation

Use thirty-five species

Untransformed data

Origin and scale are arbitrary

($\neq$ denotes coincident points)

Fig. 5.5

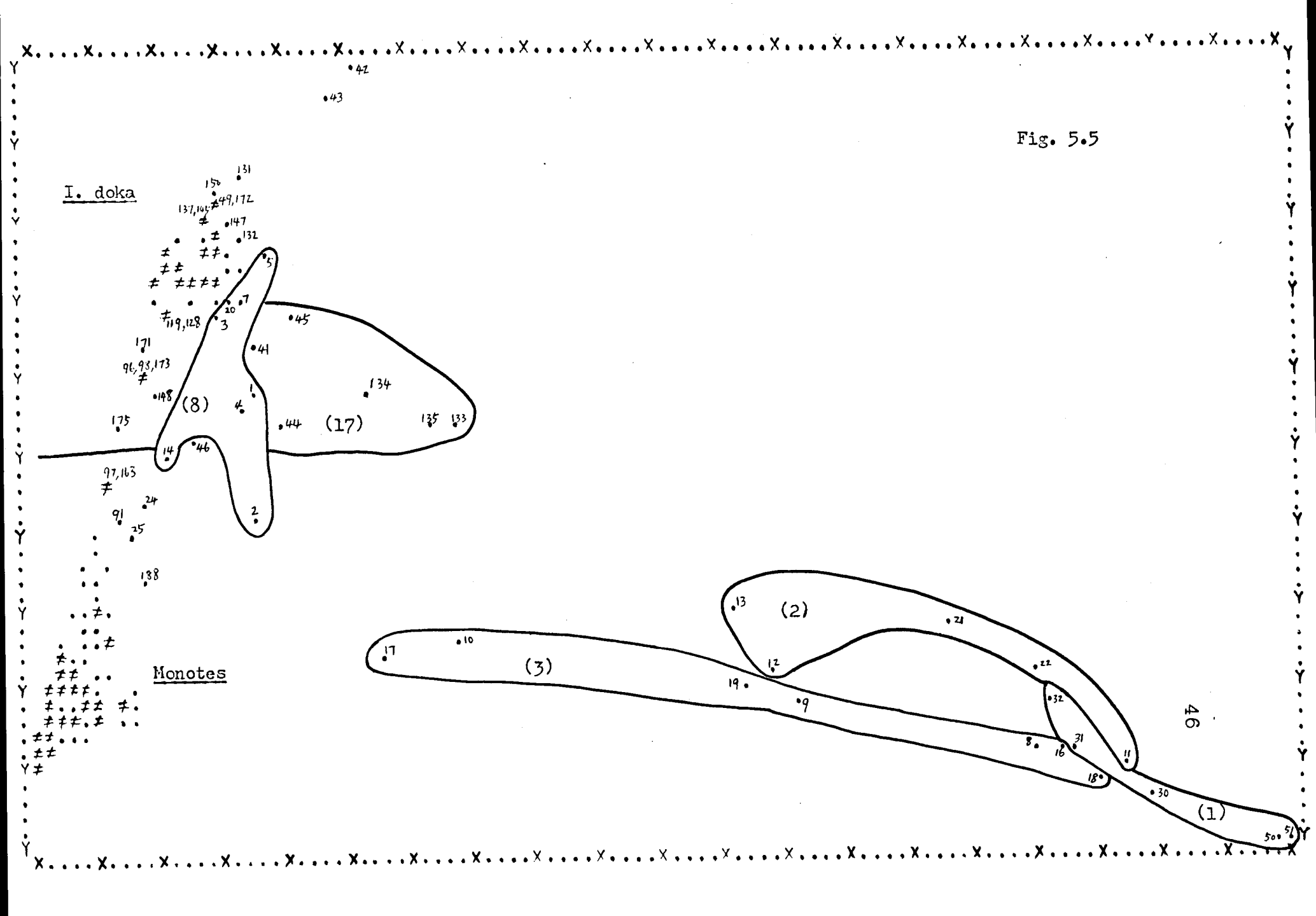


Fig. 5.6

X axis-- site scores with highest correlation

Y axis-- site scores with second highest correlation

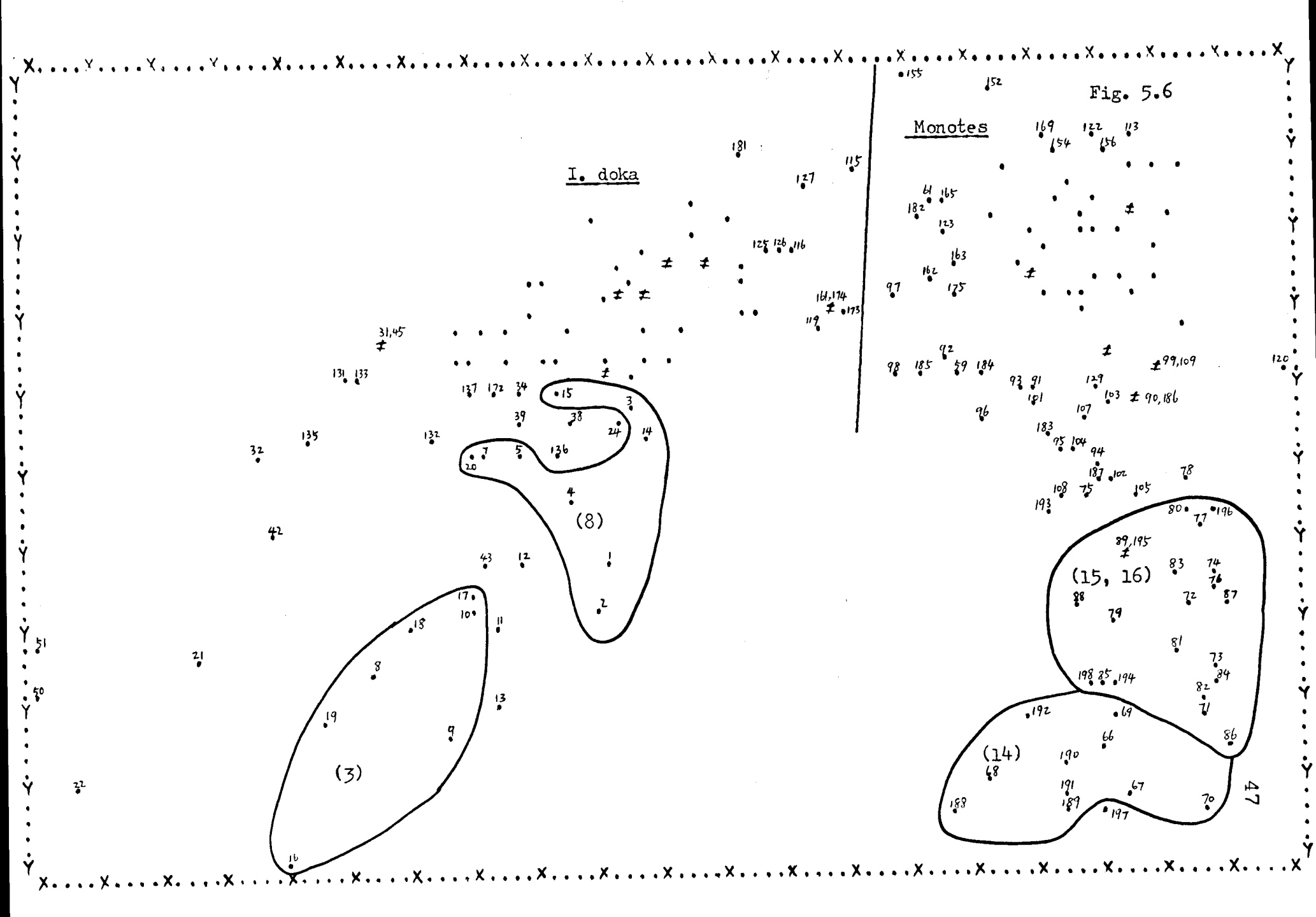
Use thirty-five species

Presence-absence data

Origin and scale are arbitrary

($\neq$ denotes coincident points)

Fig. 5.6



They conform with Fig. 5.1 and Fig. 5.4 respectively. For this specific set of ecological data, although correspondence analysis emphasizes rare species, it does share one property with P.C.A.. Employing the most abundant species from about two-thirds of the total species, we derive results similar to those obtained from the complete data set.

6. MODE ANALYSIS

6.1 Raw Data

Wishart (1969b) developed the computer package CLUSTAN for general cluster analysis. The mode analysis programs are used for the following analyses. The user can either employ the Euclidean distance between individuals or the product moment correlation between individuals to construct the similarity matrix for calculation. The frequency threshold k has to be decided before the calculation. The programs work according to the algorithm listed in section 3.3. It introduces the dense point consecutively according to the order of KD and performs one of the three fuse actions as a cycle. Since the really critical phases are those at which existing groups are fused together (see section 3.3, iii(c) and iv), the programs only output those groupings obtained immediately before such a fusion. In his study of several real data matrices(sample size ranging from 30 to 350), Wishart (1969a) discovered that mode analysis has the following unique features:

(1) The useful range of the frequency threshold k is about 1 to 6 depending on the sample size. For large data sets (n greater than 200), empirical trials indicate that values of k in the range 3 to 5 yield practically identical results. When k takes the value 1, the algorithm degenerates to the nearest neighbor method.

(2) During the empirical trials, the largest number of outputs was 24, while the average was about 11.

(3) In each output of the empirical trials, the number of groups formed is never more than 9, and the average maximum was about 6.

Mode analysis was first tried by using both Euclidean distance and product moment correlation on the untransformed data and the square-root transformed data with frequency threshold 3 to compare their performence. They do not present very different results. The Euclidean distance was employed on the following analyses. For the untransformed data with input frequency threshold 3, the mode analysis yielded nine outputs, each output gives two to seven divided groups. Wishart suggested that the selection of the output with the maximum number of groups was the best solution. Everitt (1974, p. 87), on the other hand, pointed out that it was not always the case, and that the decision of number of groups was heavily dependent on the investigator's evaluation. Based on the Braun-Blanquet associations, the most meaningful output grouping, conforming best to those divided associations, is in the fifth output. It has four groups which are listed below.

Group 1 contains seventeen sites: 2, 10, 14, 17, 66, 67, 68, 69, 70, 71, 76, 94, 188, 189, 190, 191, 192.

Group 2 contains twenty-three sites: 72, 73, 74, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 163, 164, 165, 193, 194, 195, 196, 197, 198.

Group 3 contains seventy-three sites: This group is composed mainly of associations 9, 10, 11, 12, and 13, which have common species Monotes.

Group 4 contains eighty-five sites: This group is composed mainly of associations 1, 2, 3, 4, 5, 6, 7, 8, and 17. They have common species Isoberlinia doka.

The above analysis shows that there probably are four natural groups. But compared with the Braun-Blanquet associations, they are not in close enough conformity both in the number of groups and the sites presented in the groups. Hence further analysis is needed in order to reveal more information.

The mode analysis for the square-root transformed data with input threshold 3 yielded better results. There are twelve outputs, each presenting two to seven groups. The output which conforms best is the ninth output. Its six groups are listed below.

Group 1 contains seven sites: 8, 9, 10, 16, 17, 18, 19.

Group 2 contains ten sites: 11, 12, 13, 21, 22, 30, 31, 32, 50, 51.

Group 3 contains eleven sites: 2, 66, 67, 68, 69,

70, 188, 189, 190, 191, 192.

Group 4 contains twenty-one sites: 71, 72, 73, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 193, 194, 195, 196, 197, 198.

Group 5 contains sixty-two sites: 1, 3, 4, 6, 7, 14, 15, 20, 34, 35, 36, 37, 38, 39, 40, 42, 43, 48, 49, 52, 53, 54, 96, 97, 98, 115, 116, 117, 118, 119, 125, 126, 127, 128, 131, 132, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 161, 162, 163, 164, 165, 171, 172, 173, 174, 175.

Group 6 contains eighty-seven sites: 5, 23, 24, 25, 26, 27, 28, 29, 33, 41, 44, 45, 46, 47, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 74, 75, 76, 77, 78, 91, 92, 93, 94, 95, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 120, 121, 122, 123, 124, 129, 130, 133, 134, 135, 152, 153, 154, 155, 156, 157, 158, 159, 160, 166, 167, 168, 169, 170, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186, 187.

The above groups conformed successfully with the Braun-Blanquet associations. Group 1 is association 3; group 2 is associations 1 and 2; group 3 is association 14; group 4 is associations 15 and 16; group 5 is associations 4, 5, 6, and 8; group 6 is associations 9, 10, 11, 12, 13, and 17. Aside from association 7, which is dispersed through groups 5 and 6, there is virtually no mixing present in the six

groups. The larger groups 5 and 6 could be subjected to furhter analyses in order to obtain more detailed information. In section 4.3, they were used as input data for P.C.A. I also tried to employ the mode analysis again toward both group 5 and group 6. But they could not derive a finer grouping. The frequency thresholds 2 and 3 were used in these analyses. Among the outputs obtained in the four mode analyses (the two thresholds tried with the untransformed data and with the square-root transformed data) carried out for group 5, only one output which formed the group with sample sites 115, 116, 117, 118, 119, 148, 149, 150, 151, 161, 171, 172, and 173 conforms with association 4. This is also the case in the four mode analyses outputs for group 6. The best-conforming output has three groups. One group has sites 5, 41, 45, 133, 134, and 135, and conforms with association 17, another group contains sites 23, 24, 25, 26, 27, 28, 29, 33, 44, 46, 47, and conforms with association 12, the other seventy sites are formed into the third group. The rest of the outputs are just groups of totally mixed sample sites.

The above results can be derived by applying the frequency threshold 2 directly to the complete data set with square-root transformation. It gives eight outputs, each of them contains five to seven groups. The most meaningful output is the seventh. Its seven groups are as follows.

Group 1 contains six sites: 5, 41, 45, 133, 134, 135.

Group 2 contains eleven sites: 23, 24, 25, 26, 27, 28, 29, 33, 44, 46, 47.

Group 3 contains thirteen sites: 8, 11, 12, 13, 16, 18, 21, 22, 30, 31, 32, 50, 51.

Group 4 contains fifteen sites: 2, 9, 10, 17, 19, 66, 67, 68, 69, 70, 188, 189, 190, 191, 192.

Group 5 contains twenty-one sites: 71, 72, 73, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 193, 194, 195, 196, 197, 198.

Group 6 contains sixty-two sites: It is exactly the same sample sites which were on page 52 as listed previously as group 5.

Group 7 contains seventy sites: This group is composed mainly of associations 9, 10, 11, and 13.

Comparing the above groups to those groups derived by using frequency threshold 3, we find that two more associations have been divided out, namely, association 12 and association 17. The similarity of these two results obtained by using different thresholds also confirmed the existance of natural groups in the data.

6.2 The Use of the Most Abundant Species

For most of the cluster analysis, the number of

variables involved will decide the necessary amount of calculation. It is of interest to examine the performance of mode analysis in the reduced data matrix. Firstly, mode analysis was applied to the untransformed thirty-five species data matrix with threshold 3. There were nine outputs, each with three to nine groups. The sixth output was considered to be the most significant grouping. It contains five groups.

Group 1 contains seven sites: 5, 41, 44, 45, 133, 134, 135.

Group 2 contains fourteen sites: 8, 11, 12, 13, 16, 18, 21, 22, 30, 31, 32, 42, 50, 51.

Group 3 contains seventeen sites: 2, 10, 14, 17, 66, 67, 68, 69, 70, 71, 76, 94, 188, 189, 190, 191, 192.

Group 4 contains twenty-two sites: 72, 73, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 163, 164, 165, 193, 194, 195, 196, 197, 198.

Group 5 is formed by the remaining one hundred and thirty-eight sites.

Those smaller groups roughly conform to the Braun-Blanquet associations, but no association is completely recovered. The overall results are similar to those derived from the complete data set.

Again, the results obtained by using the square-root transformed data are much better. There are twelve

outputs and each of them forms three to nine groups. The ninth output is chosen and its seven groups are as follows.

Group 1 contains six sites: 8, 9, 10, 16, 18, 19.

Group 2 contains twelve sites: 11, 12, 13, 21, 22, 30, 31, 32, 42, 43, 50, 51.

Group 3 contains twelve sites: 2, 17, 66, 67, 68, 69, 70, 188, 189, 190, 191, 192.

Groups 4 contains seventeen sites: 5, 23, 24, 25, 26, 27, 28, 29, 33, 41, 44, 45, 46, 47, 133, 134, 135.

Group 5 contain twenty-one sites: 71, 72, 73, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 193, 194, 195, 196, 197, 198.

The remaining sample sites form two large groups: one with sixty sites, most of them sites with Isoberlinia doka, the other with seventy sites, most of them sites with Monotes. The discarding of the rare species makes little difference in the results of mode analysis. For the untransformed data, discarding the rare species enables the mode analysis to divide out both the three small associations, 1, 2, and 3 to form group 2, and the association 17 to form group 1. But both do not conform perfectly to the Braun-Blanquet associations. For the square-root transformed data, although the discarding of the rare species gives an extra group 4 which is composed primarily of associations 12 and 17, we can notice that a few sites have been misclassified. Except

these minor changes, reducing the number of species does not make much difference in the results.

From the mode analysis algorithm, we see that the reduction of the number of variables can simplify only the preperations of the similarity matrix. The major computation needed is dependent on the number of individuals involved. In terms of saving the amount of calculation, the discarding of rare species makes no great contribution and should not be taken into much consideration.

6.3 Mode Analysis after P.C.A. and Correspondence Analysis

Sometimes the ordination results were used as the input data instead of the original data for cluster analysis to economize the calculation time and to reduce the data variation. The methods, however, project the original data into fewer dimensions. This provides a chance for distorting the data structure. The derived groups might be completely wrong, especially when the ordination itself is not efficient. Here, I employ the first ten principle components derived from the square-root transformed data as input data for mode analysis with frequency threshold 3. The results contain fifteen outputs and each one has three to nine divided groups. Referring to the Braun-Blanquet associations, the ninth output is regarded as the best conforming one. It's eight groups are as follows.

Group 1 contains seven sites: 41, 44, 45, 46, 133, 134, 135.

Group 2 contains nine sites: 23, 24, 25, 26, 27, 28, 29, 33, 47.

Group 3 contains ten sites: 74, 75, 76, 77, 78, 93, 105, 185, 186, 187.

Group 4 contains fourteen sites: 8, 9, 11, 12, 13, 16, 18, 21, 22, 30, 31, 32, 50, 51.

Group 5 contains fifteen sites: 2, 10, 17, 19, 66, 67, 68, 69, 70, 94, 188, 189, 190, 191, 192.

Group 6 contains twenty-one sites: 71, 72, 73, 79, 80, 81, 82, 83, 84, 85, 86, 86, 87, 88, 89, 90, 193, 194, 195, 196, 197, 198.

Group 7 contains fifty-nine sites: This group is composed mainly of associations 4, 5, 6, and 8.

Group 8 contains sixty-three sites: This group is composed mainly of associations 9, 10, and 11.

The above analysis shows that the results derived from using the first ten principle components divide out more associations than that of the complete data set. It has three extra groups, associations 12, 13, and 17. But more sites are misclassified. On the whole, the utilization of P.C.A. results makes the mode analysis more effective but less accurate in recovery of the associations.

By using the first ten sets of species scores which

were derived by correspondence analysis from the square-root transformed data as input data, again, I performed the mode analysis with threshold 3. It gives eleven outputs and each one has three to seven groups. Most of the outputs do not conform to the Braun-Blanquet associations. The best one is the tenth, which has four groups. They are as follows.

Group 1 contains six sites: 41, 44, 45, 133, 134, 135.

Group 2 contains nine sites: 8, 11, 16, 18, 30, 31, 32, 50, 51.

Group 3 contains ten sites: 66, 67, 68, 69, 70, 188, 190, 191, 192.

The rest of the sites form a large group. It is seen easily that using the correspondence analysis results as input data for mode analysis is much less satisfactory than using of the P.C.A. results.

7. CONCLUSION

Multivariate data are collected in many different disciplines and each data set has its own structure and meaning. The numerous multivariate statistical techniques available were developed in an effort to satisfy many different purposes and there is no general method for the investigator to use. Thus the choice of an appropriate technique and the meaningful interpretation of the results depends heavily on the researcher's experiences and knowledge in his specific field. The use of Braun-Blanquet associations in this project as basis for comparing analysis is only an attempt to attach ecological meanings and is not necessarily the best approach. Perhaps someone with greater ecological experience could extract considerably more information from the various analyses.

P.C.A. is an effective methodology for the condensation of the data structure. But when the data are multimodal, its interpretation is in doubt. Section 4.3 demonstrated an attempt to avoid this difficulty. The results obtained are rather satisfactory. P.C.A. is also hampered by its linearity assumptions and cannot satisfactory deal with the data which are not linearily correlated. Transformations,

such as the square-root transformation or logarithm transformation, cannot always correct this non-linearity. If there is evidence of a non-linear data structure, other methods which are capable of dealing with this situation should be employed.

When the scale of measurement is the same, the use of the covariance matrix for P.C.A. calculation yield more meaningful results than does the use of the correlation matrix. The first three principle components derived by P.C.A. with untransformed data take account of more variation than that of the P.C.A. with transformed data. The results of untransformed data only reflect the species variation and over simplify the real data structure. The results of P.C.A. derived from transformed data are more meaningful.

As pointed out in section 4.1, correspondence analysis is just an ordination method with the dual property of site scores and species scores but with no readily interpretable meaning. It has no advantages over P.C.A. and further studies are needed for the interpretation of derived scores. P.C.A. and correspondence analysis have the similar ability to recover the Braun-Blanquet associations. Both techniques are insensitive to the discarding of rare species and this can be considered in sampling procedure to save time and effort. In the condensation of the data structure, as has been seen in section 6.3, P.C.A. is much better than correspondence analysis.

Mode analysis is also insensitive to the discarding of rare species. But the reduction of dimensionality of variables cannot reduce effectively the amount of calculation needed for mode analysis. The major computation needed is dependent on the number of individuals involved. The ouput selected as the solution in mode analysis depends heavily on the investigator's own judgement. In this project, I chose the output most conforming to the Braun-Blanquet associations as the best solution. In many cases, the results successfully recover the Brayn-Blanquet associations. Thus, mode analysis appears to be a powerful technique for the discovery of the natural groups from the ecological data. In conclusion, the three multivariate statistical techniques employed in this project reveal, in some degree, the aspects of the data structure. For the purpose of obtaining a objective insight into the data, they are the useful techniques for the ecologists.

APPENDIX A

Key to Species Numbers

1. Diospyros mespiliformis
2. Grewia mollis
3. Anogeissus leiocarpus
4. Lannea schimperi
5. Sterospermum kunthianum
6. Cussonia barteri
7. Khaya senegalensis
8. Steganotaenia araliacea
9. Entada africana
10. Annona senegalensis
11. Terminalia avicennioides
12. Lannea microcarpa
13. Isoberlinia doka
14. Pterocarpus erinaceus
15. Bridelia ferruginea
16. Afrormosia laxiflora
17. Parinari curatellifolia
18. Vites doniana
19. Strychnos spinosa
20. Daniellia oliveri
21. Swartzia madagascariensis
22. Monotes kerstingii
23. Strychnos innocua
24. Parkia clappertoniana
25. Uapaca togensis
26. Nauclea latifolia
27. Trichilia emetica
28. Hymenocardia acida
29. Ixora bauchiensis
30. Ximenia americana
31. Terminalia laxiflora
32. Detarium microcarpum
33. Psorospermum corybiferum
34. Ochna afzelii
35. Piliostigma thonningii
36. Lophira lanceolata
37. Securidaca longepedunculata
38. Butyrospermum paradoxum
39. Protea elliottii
40. Gardenia ternifolia

41. G. erubescens
42. Faurea speciosa
43. Combretum molle
44. Crossopteryx febrifuga
45. Combretum binderianum
46. Isoberlinia tomentosa
47. Vitex simplicifolia
48. Combretum glutinosum
49. Dichrostachys cinerea
50. Cassia singueana

APPENDIX B

Data List of the First Five Sample Sites

species	sample sites					species	sample sites				
	1	2	3	4	5		1	2	3	4	5
1	0	0	0	0	0	26	0	0	0	0	0
2	0	0	0	4	0	27	0	0	0	0	0
3	0	0	0	0	0	28	0	0	16	0	0
4	4	1	9	0	10	29	0	0	0	0	4
5	0	0	2	0	0	30	3	0	1	3	3
6	0	0	0	0	1	31	0	1	0	0	0
7	0	0	0	0	0	32	13	26	3	23	2
8	0	1	0	2	13	33	0	0	1	0	1
9	0	6	5	4	1	34	0	0	0	0	0
10	1	4	9	5	17	35	0	0	5	0	17
11	21	17	77	31	8	36	0	0	0	0	0
12	1	1	1	4	1	37	4	1	2	0	1
13	10	3	22	12	10	38	6	1	3	3	14
14	0	0	0	1	0	39	0	0	0	0	0
15	0	0	5	0	14	40	0	0	1	1	0
16	0	0	9	0	0	41	1	18	12	2	0
17	0	0	1	0	0	42	0	0	0	0	0
18	0	0	0	0	2	43	1	3	0	0	0
19	16	8	11	2	4	44	1	0	0	0	0
20	0	0	0	0	0	45	0	0	0	0	0
21	0	0	0	1	0	46	0	0	2	1	27
22	0	0	0	0	0	47	1	3	0	2	1
23	1	0	0	1	0	48	6	9	0	9	0
24	1	0	0	0	0	49	0	0	0	0	0
25	0	0	0	0	0	50	0	0	0	0	0

APPENDIX C

The Seventeen Associations

(1) *Dichrostachys* erosion complex association (five sites): 30, 31, 32, 50, 51.

(2) *Anogeissus-Feretia* river bank association (five sites): 11, 12, 13, 21, 22.

(3) *Combretum glutinosum-Dichrostachys-Entada* ironstone pavement association(seven sites): 8, 9, 10, 16, 17, 18, 19.

(4) *Isoberlinia doka-Pterocarpus* fringing woodland association(thirteen sites): 115, 116, 117, 118, 119, 148, 149, 150, 151, 161, 162, 171, 172.

(5) *Isoberlinia doka-Annona-Terminalia-Ximenia* association(twenty-seven sites): 6, 34, 35, 36, 37, 38, 39, 40, 48, 49, 52, 53, 54, 131, 132, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147.

(6) *Isoberlinia doka-Parinari* association(ten sites): 125, 126, 127, 128, 163, 164, 165, 173, 174, 175.

(7) *Isoberlinia doka-Uapaca* association(eight sites): 91, 92, 93, 94, 95, 96, 97, 98.

(8) *Isoberlinia doka-Detarium* association(nine sites): 1, 2, 3, 4, 5, 7, 14, 15, 20.

(9) *Monotes-Isoberlinia doka* association(seventeen sities): 55, 107, 108, 110, 114, 121, 122, 152, 153, 154,

155, 166, 167, 176, 177, 181, 182.

(10) Monotes-Parinari association(seventeen sites): 56, 59, 60, 61, 62, 63, 64, 65, 103, 104, 105, 106, 109, 123, 124, 130, 183.

(11) Monotes southern association(twenty sites): 99, 100, 101, 102, 111, 112, 113, 120, 129, 156, 157, 158, 159, 160, 168, 169, 170, 178, 179, 180.

(12) Monotes northern association(ten sites): 23, 24, 25, 26, 27, 28, 29, 33, 46, 47.

(13) Monotes-Parinari-Detarium association (nine sites): 74, 75, 76, 77, 78, 184, 185, 186, 187.

(14) Detarium-Gardenia ironstone association (ten sites): 66, 67, 68, 69, 70, 188, 189, 190, 191, 192.

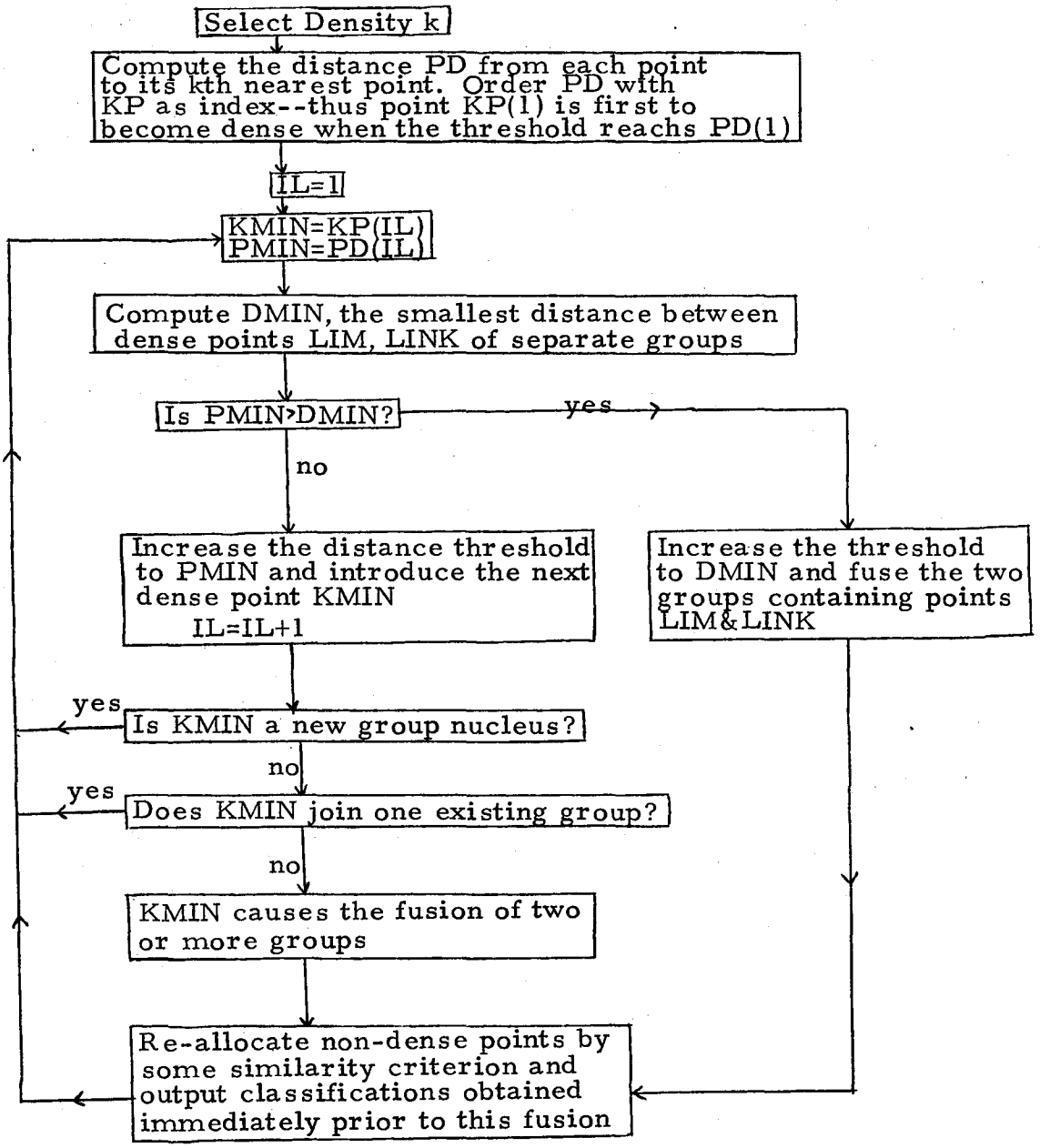
(15) Parinari-Gardenia-Detarium association (seventeen sites): 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 193, 194, 195, 196, 197, 198.

(16) Daniellia-Gardenia-Detarium association (four sites): 71, 72, 73, 90.

(17) Isoberlinia tomentosa-Isoberlinia doka association(six sites): 41, 44, 45, 133, 134, 135.

APPENDIX D

Flow Chart for Mode Analysis Computer Program



BIBLIGRAPHY

Austin, M.P. and Greig-Smith, P. (1968). "The Application of Quantitative Methods to Vegetation Survey, II: Some Methodological Problems of Data from Rain Forest," Journal of Ecology, 56, pp. 827-44.

Bartlett, M.S. (1954). "A Note on the Multiplying Factors for Various Chi-squared Approximations," Journal of Royal Statistical Society, series B, 16, pp. 296-8.

Braun-Blanquet, J. (1951). Pflanzensoziologie, 2nd edn. Vienna: Springer.

Craddock, J.M. (1965). "A Meteorological Application of Principle Component Analysis," The Statistician, 15, pp. 143-56.

Duran, B.S. and Odell, P.L. (1974). Cluster Analysis: A Survey. New York: Springer-Verlag.

Everitt, B. (1974). Cluster Analysis. New York: John Wiley & Sons.

Fisher, R.A. (1940). "The Precision of Discriminant Functions," Annals of Eugenics, 10, pp. 422-30.

Greig-Smith, P. (1964). Quantitative Plant Ecology. 2nd edn. New York: Plenum Press.

Hill, M.O. (1973). "Reciprocal Averaging: An Eigenvector Method of Ordination," Journal of Ecology, 61, pp. 237-49.

_____. (1974). "Correspondence Analysis," Journal of Royal Statistical Society, series C, 23, pp. 340-54.

Kendall, M.G. (1939). "The Geographical Distribution of Crop Productivity in England," Journal of Royal Statistical Society, 102, pp. 21-30.

_____. (1957). A Course in Multivariate Analysis. New York: Hafner Publishing Co.

_____ and Stuart, A. (1967). The Advanced Theory of Statistics. 2nd edn. London: Charles Griffin & Co. Ltd.

Kershaw, K.A. (1968a). "A Survey of Vegetation in Zaria Province, North Nigeria," Vegetatio, 30, pp. 244-68.

_____. (1968b). "Classification and ordination of Nigerian Savanna Vegetation," Journal of Ecology, 56, pp. 467-82.

_____ and Shepherd, R. (1972). "Computer Display Graphics for Principle Component Analysis and Vegetation Ordination Studies," Canadian Journal of Botany, 50, pp. 2239-50.

_____. (1973). Quantitative and Dynamic Plant Ecology. 2nd edn. New York: American Elsevier Publishing Co., Inc.

Marriott, F.H.C. (1974). The Interpretation of Multiple Observations. London: Academic Press.

Morrison, D.F. (1969). Multivariate Statistical Methods. New York: McGraw-Hill Book Co.

Morrison, D.G. (1967). "Measurement Problems in Cluster Analysis," Management Science, 13, pp. B-775-80.

Orloci, L. (1966). "Geometric Models in Ecology: The Theory and Application of Some Ordination Methods," Journal of Ecology, 54, pp. 193-215.

_____. (1975). Multivariate Analysis in Vegetation Research. Netherland: The Hague.

Seal, H.L. (1966). Multivariate Statistical Analysis for Biologists. London: Methuen & Co. Ltd.

Whittaker, R.H. (1967). "Gradient Analysis of Vegetation," Biological Review, 42, pp. 207-64.

Williams, W.T., Lambert, J.M., and Lance, G.N. (1966). "Multivariate Methods in Plant Ecology: Similarity Analyses and Information-analyses," Journal of Ecology, 54, pp. 427-45.

Wishart, D. (1969a). "Mode Analysis: A Generalization of Nearest Neighbor which reduces Chaining Effects," in Numerical Taxonomy, ed. by A.J. Cole, New York: Academic Press.

_____. (1969b). Fortran II Programs for 8 Methods of Cluster Analysis (Clustan I), Computer Contribution 38, Lawrence: The University of Kansas Press.

Noy-Meir, I. (1973). "Data Transformations in Ecological Ordination, I: Some Advantages of Non-centering," Journal of Ecology, 61, pp. 329-41.

_____, Walker, D., and Williams, W.T. (1975). "Data Transformationsin Ecological Ordination, II: On the Meaning of Data Standardization," Journal of Ecology, 63, pp. 779-99.