INVESTIGATIONS INTO THE PHENETIC RELATIONSHIPS BETWEEN SPECIES OF *ONONIS* L.

By R. B. Ivimey-Cook

Department of Botany, University of Exeter

Abstract

Several multivariate methods have been tested against Sirjaev's classification of *Ononis*, and their suitability for various purposes considered. It is concluded that while Association-analysis may provide the basis of a suitable key to OTUs, Principal Components Analysis offers one of the best starting points for polythetic classification. The difficulty of portraying the inter-OTU relationships can be resolved in part by using cluster analysis on the distance matrix calculated from the component specifications of the OTUs, over such a number of components as accounts for a suitable proportion of the variation within the population. Cluster analyses of distance matrices, using average linkage, calculated directly from the standardized data matrix did not, on the whole, give such satisfactory results. A comparison of weighted against unweighted cluster analyses showed in all cases that the latter seemed more appropriate.

1. Introduction

A principal aim of taxonomy must be the preparation of monographic accounts of genera. Any attempt in such monographs to discuss relationships between species ought to be made in the light of as much information as possible about the species concerned. Too often the published data are incomplete, often in material particulars. The consideration of relationships, which is a feature of many monographs, tends to be based on a limited, and often not wholly explicit, range of the more prominent characters. While it may be possible to infer phylogeny from fossil evidence in certain instances, in most cases, relationships can only be assessed between species as they occur at the present time.

The present paper enquires into the phenetic similarities between the species of the genus *Ononis* and considers these results by comparison with the arrangement presented by Sirjaev in his monograph of the genus (Sirjaev 1932); it allows some examination of the problems associated with quantitative aspects of Angiosperm taxonomy.

The tribe Trifolieae in the Papilionaceae contains 6 genera, among which *Ononis* may be distinguished by its relatively large pink, purple or yellow corolla and a legume which is neither coiled nor crescentic. Sirjaev divides the genus into two Sections; *Natrix*, which includes species with a paniculate inflorescence* and an elongated legume which is usually pendent, and *Bugranae*, whose species have a usually racemose inflorescence and an erect, rhomboid or ovoid legume. Within both Sections, a number of Sub-sections and Series are recognized (See Appendix 1, p 19).

The genus contains about 70 species, mostly distributed along both shores of the Mediterranean. The centre of distribution seems to have been southern Spain or North Africa, where nearly half the species occur. Very few penetrate north of the Alps, though 3 species occur in Britain. Several others can be found in the Middle East and south-west Asia, while *O. arvensis* is recorded as far east as India. Generally, the species flourish in fairly dry and frequently calcareous soils, but they are not especially adapted to extremes of either temperature or humidity.

There is a considerable range in the species concept within the genus. Three species in particular are very variable, *O. natrix*, *O. viscosa* and *O. spinosa*; in the last, some of

* The inflorescence in this and related genera is difficult to define. The term 'paniculate' is here used to describe an inflorescence composed of a number of axillary peduncles, which vary from 1 mm to several cm in length, each bearing 1–3 shortly pedicellate flowers. Those termed 'racemose' have no peduncle and the flowers are borne either sessile or very shortly pedicellate in the axils of the bracts, usually singly, occasionally in pairs.

the variants are commonly treated as having specific rank, but in general, the variation is recognized by the establishment of a number of subspecies and varieties. At the other extreme, there are a number of species pairs where the inter-specific distinctions are largely matters of size or other relatively minor features.

2. Methods

Various methods can be used to assess similarities between species; these range from the intuitive (subjective) assessments of the conventional or herbarium taxonomist to the so-called objective techniques of numerical taxonomy. The basic methodology of numerical taxonomy has been set out by Sokal & Sneath (1963) and it will suffice here to outline the techniques used.

The basis of numerical methods is the unit character, which must express some aspect of the genome; such characters may be divided into a number of states, which may be expressed in any units, and the numerical values of these states provide the basis for the calculation of similarity coefficients between the species. Various statistical techniques can then be used to investigate the relationships between the species, as measured by the characters employed.

Since it is not necessary to confine the investigations to categories of the same rank, such as species, the non-committal term 'operational taxonomic unit' (OTU) is normally employed. A total of 107 OTUs of *Ononis* were considered, comprising 57 species represented by one OTU each and 10 species represented by more than one OTU. For each OTU, 64 characters were investigated and their states determined. In Appendix 2 is given a list of these characters and an indication of the range of their states.

The determination of the states of the various characters was carried out chiefly on herbarium material in the British Museum (Natural History) (**BM**) and in Kew (**K**). Most of the quantitative characters have a considerable range, not only on a single individual, but also within the OTU. Since only a single value for each character can be inserted in the data matrix, the mean of the values measured, in some cases modified in the light of those given by Sirjaev, was generally adopted.

For certain characters, particularly lengths and breadths, small differences at the low end of the range may be as significant as much larger differences at the upper end. Kendrick (1964) considers resolving this problem by coding such states on a linear scale and incrementing the code increments. This may approximate to a logarithmic transformation of the data. In the present work, logarithmic transformations have been applied to the states of certain characters where it seemed that, in the untransformed data, the values of the states at the upper end of the range would have an over-riding influence on the results. The question of data transformations of this sort is a vexed one; while it is arguable that a logarithmic transformation is not appropriate in every case, there is no special virtue in the linear scale of measurement conventionally used.

Two main approaches may be used to obtain a classification of individuals—divisive-monothetic and agglomerative-polythetic (see Williams, Lambert & Lance 1966). The divisive-monothetic approach considers the entire population and splits it up hierarchically on the basis of the presence or absence of one attribute at every step of the hierarchy. Agglomerative-polythetic methods employ indices of similarity and combine those individuals which are most similar, again hierarchically, until the original population has been reconstructed.

One divisive-monothetic method was investigated (Association Analysis, see Williams & Lambert 1959, 1960, 1961). This method has been used with considerable success in ecological investigations and can utilize one of several possible association indices. These indices have been constructed for use with presence/absence data and, to apply this method, the characters must be recoded so that the states conform to this requirement. The avoidance of logical correlations as a result of this recoding usually involves the loss of a substantial amount of information; on these grounds alone, the association index is not an especially suitable tool for the investigation of taxonomic resemblance.

Agglomerative-polythetic methods are considerably more flexible. Mostly, they allow multistate characters to be used, though one, Information Analysis, uses data in a form identical with Association Analysis. Information Analysis is described by Williams, Lambert & Lance (1966) and involves the calculation of an information statistic which gives a measure of the disorder of a group of individuals. Successive combinations involve those pairs of individuals or groups giving the least increment in the value of the information statistic.

Multistate characters may be used to obtain either a correlation coefficient or a measure of taxonomic distance. Distance may be calculated as the distance between a pair of OTUs in a Euclidean hyperspace generated by treating the character states as a set of Cartesian co-ordinates. Since these states have been measured in a variety of units, some restriction on the lengths of the character axes is required. It has been suggested that the states could be constrained within axes of unit length, but it would seem that the more conventional procedure of standardization (reduction to zero mean and unit variance) would be preferable in several respects.

The product-moment correlation coefficient can also be used to measure phenetic resemblance though Eades (1965) considers that its use is not always appropriate. Two correlation matrices can be calculated, that between characters (the R-matrix) and that between OTUs (the Q-matrix). Although the R-matrix does not, of itself, give a direct measure of resemblance between OTUs, in general its calculation is to be preferred, since it can be processed to give the required information. The calculation of the Q-matrix is again affected by the variety of units in which the character states have been measured. While it is possible to standardize the character states (as was necessary for the calculation of distance) and then to obtain correlations between the OTUs using the resulting standard deviates, the procedure is not unexceptionable nor, in this instance, necessary.

The investigation of matrices of distances or correlations for structure and relationships is complex. For distance matrices, the technique of Cluster Analysis investigates directly the relationships between OTUs, considering only those OTUs which form a cluster at any one time. The same technique is involved in the analysis of the matrix of information statistics. The results of a Cluster Analysis can usually be portrayed as a two-dimensional dendrogram, the reduction in dimensions being indicative of the extent to which information on other relationships is missing.

There are two main versions of Cluster Analysis—the *unweighted* method in which the similarity between a cluster and the remaining individuals is dependent on the number of OTUs comprising the individuals concerned, and the *weighted* method, in which the similarity is not dependent in this way, each individual, however constituted, being treated as equivalent to any other individual. The constitution of a cluster may either be confined to a pair of individuals (Pair-group method) or other OTUs may be considered for addition to the cluster initial, to form a larger cluster (Variable-group method). Such additional OTUs must meet some suitable criterion of closeness. When they first proposed the method, Sokal & Michener (1958) used an arbitrary fixed limit which applied to all clusters. This seems insufficiently flexible to be applied to different stages of the clustering process and an alternative limit of 5 per cent. above the distance between the OTUs comprising the cluster initial was tried and gave reasonably satisfactory results.

For most purposes, however, the Pair-group method is preferable to the Variable-group method, for a number of reasons. Firstly, there is no arbitrary criterion of closeness—different results can be obtained according to how this is set; secondly, it was found that if this limit was set at 5 per cent., as mentioned above, only a few groups of more than two OTUs were obtained, while if the level were raised to any extent, large variable groups tended to be formed, especially in the later stages of the clustering process; thirdly, if a group of more than two OTUs was appropriate to a particular situation, this could be expected to appear as a number of closely related pairs; fourthly, computer programming with variably sized groups is much more elaborate.

The Pair-group method of Cluster Analysis thus involves the detection of that pair of individuals which are mutually closest together. This result is recorded and a revised distance matrix is calculated, treating the cluster as an individual, with or without weighting as appropriate. The process is repeated until one final cluster is obtained.

Quite apart from any weighting which may be involved, the calculation of distances between a cluster and the remaining individuals may be carried out in a number of ways. The one used here is by average linkage, whereby a cluster is located in the individual space by calculating, for each individual, the average of the distance between the components of the cluster and that individual. Lance & Williams (1966) have shown, however, that average linkage is only a special case of a general system in which the nature of the sorting strategy is determined by the values of four parameters; variations in these can result in different clustering patterns and the form of the cluster analysis results becomes a function of the clustering process and cannot be an inherent property of the data.

The methods available for the analysis of correlation matrices include a number coming under the general heading of Factor Analysis, one of the more widely used being Principal Components Analysis. This technique investigates the relationships between a multi-dimensional array of variables, some of which may be correlated and hence at least partially redundant. The points representing the variables form a roughly hyperellipsoidal swarm and the method extracts a set of orthogonal, and hence uncorrelated, components, these being the principal axes of the hyperellipsoid. The components are extracted in descending order of magnitude and much of the variance is accounted for by the first few components, in effect reducing the dimensions of the problem. The question of when to stop extracting components is one on which there is little agreement among the advocates of the method; in this work, all components greater than unity were extracted. Each component has associated with it a vector (the eigenvector) giving the relation between each variable and the component. These are termed the component loadings.

Principal Components Analysis only works efficiently where the swarm of variables is approximately multivariate normal. In certain cases, and especially where the variables represent OTUs, there may well be a tendency for two or more swarms to appear—*e.g.* a dumb-bell-shaped distribution which might correspond to two sub-generic taxa. Such a distribution is most likely to appear along the first component, which effectively finds the long axis of the dumb-bell. Under these circumstances, the second and subsequent components are likely to be taxonomically meaningless, as they will be constructed in the hyperspace common to the two swarms and will reflect some characters of both. This difficulty is not obvious in the matrix of component loadings when the R-correlation matrix is analysed, but these loadings do not, of themselves, give a direct measure of the similarities between the OTUs.

If the vector of component loadings so obtained is used to post-multiply the standardized data matrix, a measure is obtained of the extent to which each OTU contains characters as weighted by their representation on that particular component. These have been called the component specifications and enable the distribution of the OTUs to be plotted on the component axes.

Although Principal Components Analysis tends to reduce the dimensions of a problem by concentrating the variation in relatively few components, only rarely is it possible to consider only the first two or three components so that the results can be conveniently portrayed on paper. The results may be plotted for pairs of components as a series of scatter diagrams, but the results are difficult to visualize and unwieldy if more than four or five components are important. Such scatter diagrams are, in effect, displaying the component loadings or specifications on Cartesian co-ordinates and the results over the required number of components may be used to calculate distance matrices which can then be analysed by Cluster Analysis, as previously described.

3. Results

(i) *Association Analysis*

The result of an Association Analysis between OTUs using character states as attributes is shown, in the form of a hierarchy, in Fig. 1. The characters which effect the divisions

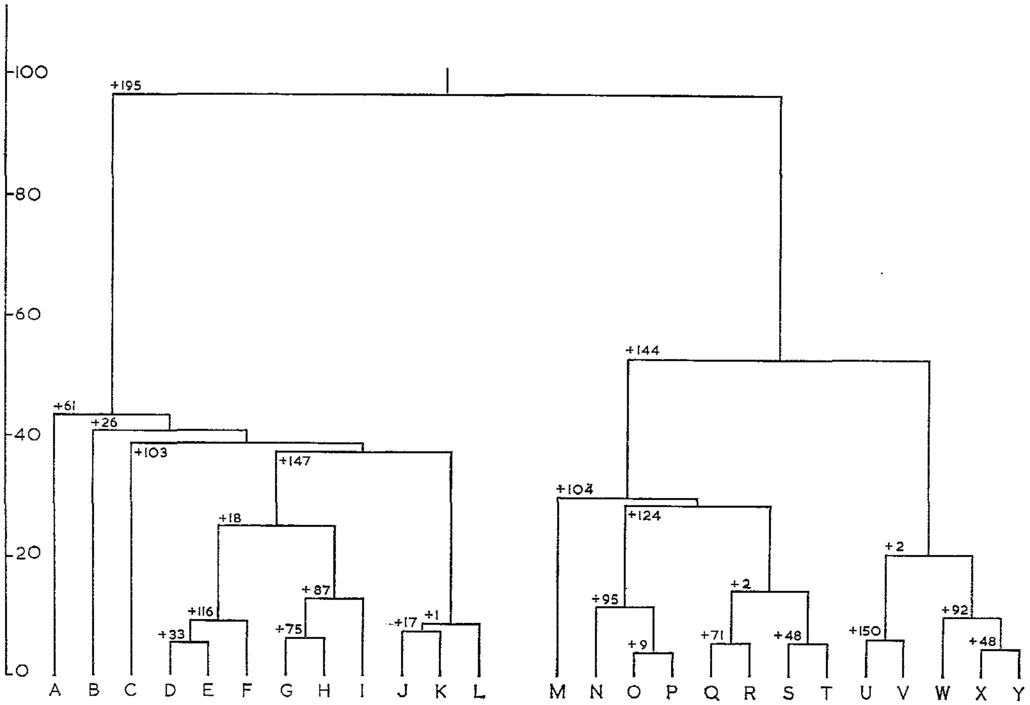

Fig. 1. The hierarchy resulting from an Association Analysis of the OTUs, using as attributes the character states as set out in Appendix 2.

at each stage are identified by the attribute number in the figure. The morphological features involved can be identified by reference to the last column of Appendix 2 and are set out in detail in Table 1. The names of the OTUs comprising the 25 final groups are set out in Table 2.

Since this method is inherently incapable of breaking down groups of fewer than 8 individuals when a 5 per cent. probability level of χ^2 is used to determine whether or not a group of individuals should be split, it is to be expected that the groups of OTUs will show varying degrees of heterogeneity. The greatest interest is likely to centre on the characters which have been chosen to effect the divisions. These characters are the attributes which when taken with every other attribute have the highest sum, over all the OTUs present, of the chosen index of association. They thus indicate the probable positions of the greatest discontinuities between the OTUs. The index of association used was χ^2/N, which theoretical considerations suggest is the most efficient. Whether it is preferable for taxonomic data has not been fully investigated. The level of maximum χ^2 at which the various divisions take place give some indications of the heterogeneity of the population under consideration, though it must be pointed out that a single significant association will suffice to cause the population to be divided.

In the standard procedure for Association Analysis, both positive and negative associations are given equal weight; in an ecological context, where the attributes are the species occurring in sample areas, a case can be made for this. Whether the same is true for taxonomic data, where the attributes represent various states of a character, is more questionable, and further investigation may show that a consideration of positive associations only may give a result which is taxonomically more meaningful.

It will be clear from an examination of Fig. 1 that the technique of Association Analysis produces a result which closely simulates a dichotomous key and since only those characters specifying the positions of the greatest discontinuities are mentioned, it is hardly to be

TABLE 1. Association analysis: subdivision characters (see Fig. 1).

195	Pod not erect
61	Leaflets with dense short hairs
26	Stipule base vaginate or infundibuliform
103	Calyx teeth with 1 nerve
147	Apex of standard not emarginate
18	Stipules longer than ca. 12 mm
116	Calyx with dense short glandular hairs
33	Petiole longer than ca. 7·5 mm
87	Arista present
75	Peduncle longer than ca. 12 mm
1	Plant not a herbaceous annual
17	Stipules longer than ca. 7·5 mm
144	Standard with glandular hairs
104	Calyx with more than 5 nerves/tooth
124	Standard purple
95	Calyx tube longer than 2 mm
9	Spines strong
2	Plant not annual
71	Bracts with 3 leaflets
48	Leaves in upper half of plant with more than 1 leaflet
2	Plant not annual
150	Standard apex not rounded, apiculate
92	Calyx not campanulate
48	Leaves in upper half of plant with more than 1 leaflet

expected that the OTUs comprising the 25 final groups are phenetically similar. If each OTU were to be replicated so that, on analysis, the final groups were composed of such replicates, then the characters chosen as division attributes should be those best suited for the construction of keys for the purposes of identification.

(ii) *Information Analysis*

The dendrogram constructed by the aggregation of OTUs using the information statistic as the criterion was not satisfactory. A number of groups could be detected readily, particularly groupings of subspecific taxa and some groups of closely related species, but the larger clusters did not appear to conform to any rational phenetic classification. This may reflect, to a considerable extent, the form in which the data were presented, particularly the technique used to convert quantitative data to presence/absence form; in any event, it is clear that further work will be needed before this technique can be applied to taxonomic situations.

(iii) *Principal Components Analysis*

This method was first applied to both the R- and Q-correlation matrices obtained from the complete 107×64 data matrix; from the former correlation matrix, the component specifications were obtained and the distribution of the OTUs plotted on the component axes. From the analysis of the Q-correlation matrix, the eigenvectors give a direct assessment of the relative positions of the OTUs. The two Sections into which *Ononis* is divided are quite readily recognizable from scatter diagrams of the OTUs plotted on the first two component axes. This is illustrated in Fig. 2 for the component specifications. Of the four OTUs which appear between the two principal groupings, *O. cintrana* (49) is recognized as being rather uncertainly placed in Section Bugranae, *O. tridentata* (2) and *O. fruticosa* (3)

TABLE 2. Association analysis: final species-groups (see Fig. 1).

A	rotundifolia
	tridentata
B	fruticosa
	cristata
	vaginalis
C	pubescens
	cintrana
D	sicula ssp. sicula
	sicula ssp. polyphylla
E	viscosa (all sspp.)
	crotalarioides
F	serotina
	pseudoserotina
	maweana
	natrix ssp. natrix
G	hebecarpa
	polysperma
	ornithopodioides
	peyerimhoffii
	antennata (all sspp.)
H	incisa
	adenotricha
	atlantica
I	reclinata
	dentata
	pendula
	laxiflora
	verae
	thomsonii
	christii
J	natrix ssp. ramosissima
	natrix ssp. stenophylla
	natrix ssp. angustissima
	natrix ssp. polyclada
	natrix ssp. filifolia
	natrix ssp. falcata
K	crispa
	natrix ssp. hispanica
	natrix ssp. mauritii
	natrix ssp. arganietorum
	natrix ssp. prostrata
L	biflora
M	megalostachys
N	repens (all vars.)
O	spinosa ssp. antiquorum (all vars.)
	spinosa ssp. leiosperma
P	spinosa ssp. spinosa (all vars.)
	spinosa ssp. foetens
Q	arvensis (all vars.)
R	masquillieri
	hispida ssp. hispida
	hispida ssp. arborescens
S	diffusa (all sspp.)
	serrata
	tournefortii
	cephalantha
	villosissima
T	alba
	monophylla
U	pinnata
	pusilla
	minutissima
	cephalotes
	saxicola
	striata
	leucotricha
V	aragonensis
	reuteri
	speciosa
W	mitissima
	alopecuroides
	baetica
	rosea
	avellana
	crinita
X	subspicata
	hirta
	cossoniana
	filicaulis
	subspicata
Y	euphrasiifolia
	variegata
	oligophylla

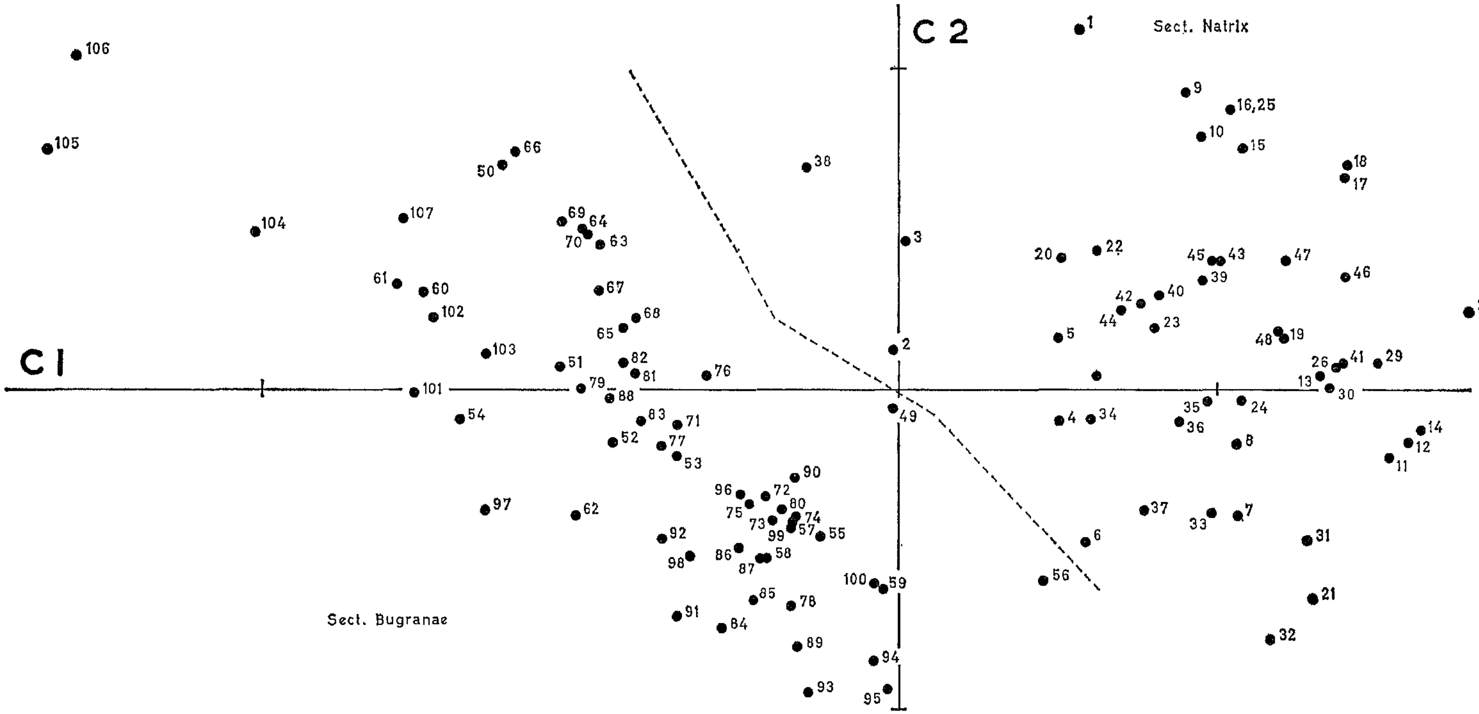

Fig. 2. A plot of the 107 OTUs on the first two component axes derived from the between-characters correlation matrix (component specifications). The two axes are drawn to the same scale and are marked at intervals of 10 units.

are classified more appropriately on component 3, as is, to a lesser extent, *O. pubescens* (38). The only species which appears misplaced in this figure is *O. minutissima* (56); again, however, it must be remembered that it is unlikely that a perfect separation can be obtained on the basis of two components which account for rather less than a quarter of the variation.

The characters which are especially involved in bringing about this separation are whether the pod is erect or pendent and the length of the peduncle, both characters which are used to differentiate the Sections. To a somewhat lesser extent, the lengths of the arista and of the calyx tube, the number or nerves in the calyx teeth and the number of seeds in a pod are all important characters in bringing about the separation along the first component. The second component is concerned chiefly with the size of the leaflets and stipules, OTUs with large foliage being found at the positive end of this axis.

Since it was felt that the immediate recognition of the two Sections would obscure the meaning of subsequent components to a considerable extent and since it was not necessary to use component analysis to make this initial separation anyway, the OTUs were split into two groups, essentially representing Section Natrix and Section Bugranae, but with a small overlap; in particular, *O. cintrana* was included in both groups.

The two groups (subsequently referred to as Group N and Group B respectively) were then re-analysed. This involved the calculation of a new R-correlation matrix for each and these were then subjected to Principal Components Analysis and the component specifications calculated. The omission of zero character scores reduced the total number of characters to 61 in Group N and 58 in Group B.

The component analyses of the two groups indicate certain differences in the shape of the hyperellipsoids. In Table 3 are given the variances of the first 12 components for Groups N and B and also, for comparison, of the original complete set of OTUs, together with the percentage of variation extracted by each component.

In Group N, the characters which contribute most to the direction of the first component are plant size, the texture and indumentum of the leaflet, the number of seeds per pod and the seed surface. The 3 OTUs from Sub-section Chrysanthae were included in Group N for purposes of comparison and this component very clearly separates these from the rest of the group. This is brought out very strongly in the plot of the component specifications. The characters with high loadings on the second component, and which thus are most concerned with distributing the OTUs along this axis, are the length of the stipules, the absence of long hairs on the calyx and the presence of short hairs on the stem, though quite a number of characters have loadings not much smaller than these. In Fig. 3 is shown the distribution of the OTUs along the component axes, using the component specifications for the first two components. Where a large proportion of the variation of a population is associated with the first two components, then such a diagram will give a good indication of the relationships between the OTUs. However, as Table 3 clearly shows, only about 25 per cent. of such variation has been considered and a plot of only the first two components cannot be considered an adequate representation of the distribution of the OTUs in the character space.

To investigate this distribution more adequately, the component specifications over the first 12 components were used to calculate a distance matrix which was then subjected to Cluster Analysis. Both the weighted and unweighted Pair-group methods were used and the resulting dendrograms were compared with Sirjaev's original classification using the cophenetic correlation coefficient. Values of 0·629 for the weighted and 0·635 for the unweighted methods suggest that there is little to choose between the two in this case. Fig. 4 shows the dendrogram for the unweighted method and a number of points are apparent:

1. In all cases where the OTUs are infraspecific taxa, these cluster together so that every species is eventually represented by a single stem. In the case of *O. natrix*, there is a marked distinction between a group comprising subspp. *natrix*, *ramosissima*, *stenophylla*, *angustissima*, *polyclada* and *filifolia* on the one hand and *hispanica*, *mauritii*, *arganietorum*, *prostrata* and *falcata* on the other. The best separation of these groups is found on component 3. though they are clearly distinguishable on the sizes of their leaves and stipules.

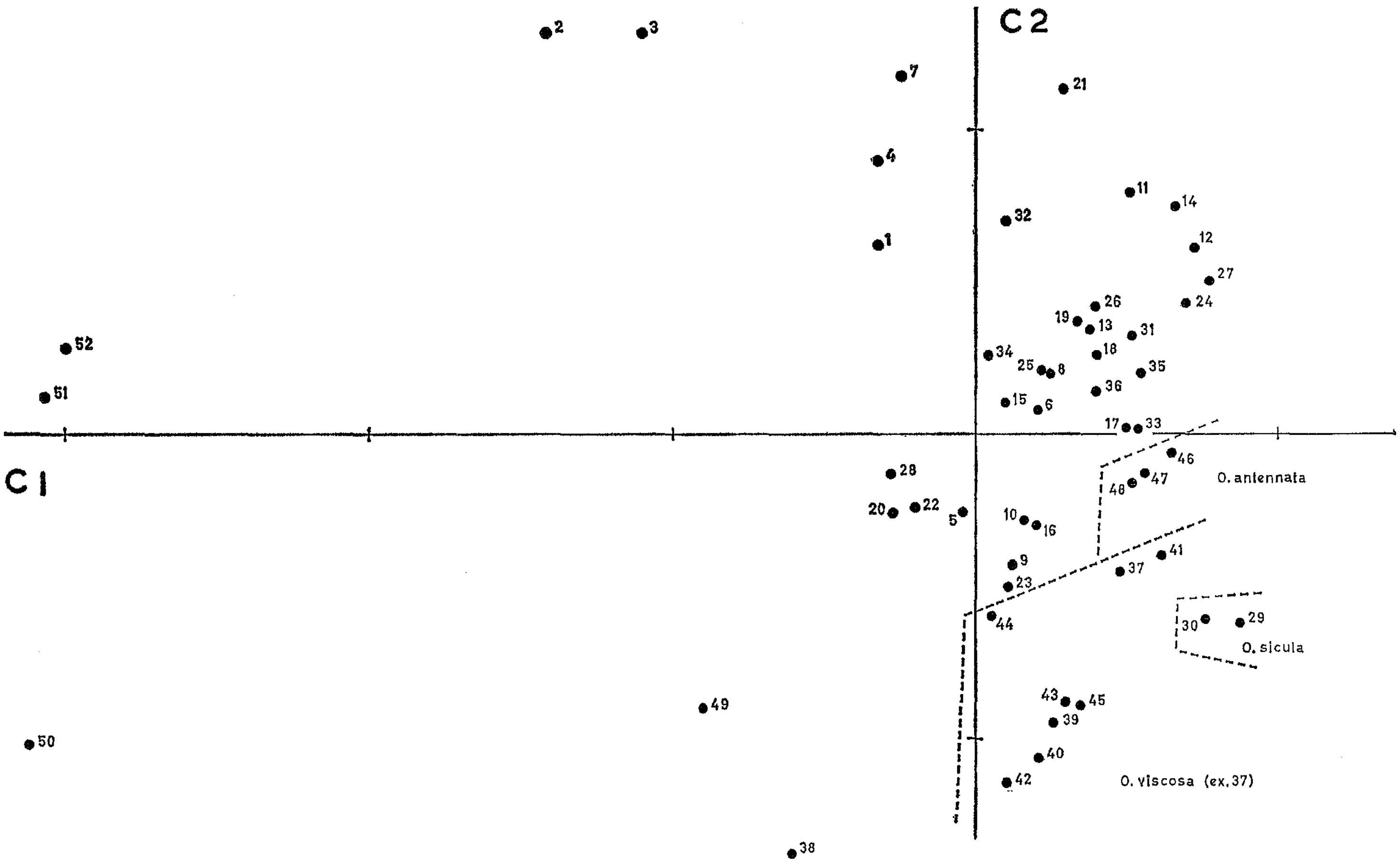

Fig. 3. A plot of the OTUs in group N on the first two component axes (component specifications). The axes are marked at intervals of 10 units.

TABLE 3. Principal Components Analysis: Variance of Components.

Component	Complete		Group N		Group B	
	Variance	Total % Variance	Variance	Total % Variance	Variance	Total % Variance
1	10·256	16·02	8·863	14·53	7·825	13·49
2	4·748	23·44	6·493	25·17	6·190	24·16
3	4·243	30·07	5·002	33·37	5·353	33·39
4	3·749	35·93	4·369	40·54	4·485	41·13
5	3·127	40·82	3·406	46·12	3·338	46·88
6	2·970	45·46	3·363	51·63	3·111	52·24
7	2·316	49·08	3·047	56·63	2·780	57·04
8	2·221	52·55	2·731	61·10	2·426	61·22
9	2·193	55·97	2·317	64·90	2·356	65·28
10	2·078	59·22	2·118	68·38	1·873	68·51
11	2·000	62·35	1·894	71·48	1·785	71·59
12	1·725	65·04	1·702	74·27	1·554	74·27
Total	41·624		45·306		43·077	

2. The recent recognition of *O. crotalarioides* (45) as a distinct species, chiefly on the basis of its much inflated pod, would not appear to be supported by these results. Overall, it seems no more dissimilar than the subspecies of *O. viscosa*, of which it was once one.

3. The most distinctive group is that comprising the 3 OTUs of Sub-section *Chrysanthae* (50–52), though the four OTUs of Sub-sections *Antiquae* and *Rhodanthae* are also well marked.

4. Sub-section *Reclinatae*, especially Series *Eu-reclinatae* (33–37) is well marked, with the addition of *O. christii* (5) and *O. thomsonii* (6) which differ chiefly in being shrubby. In

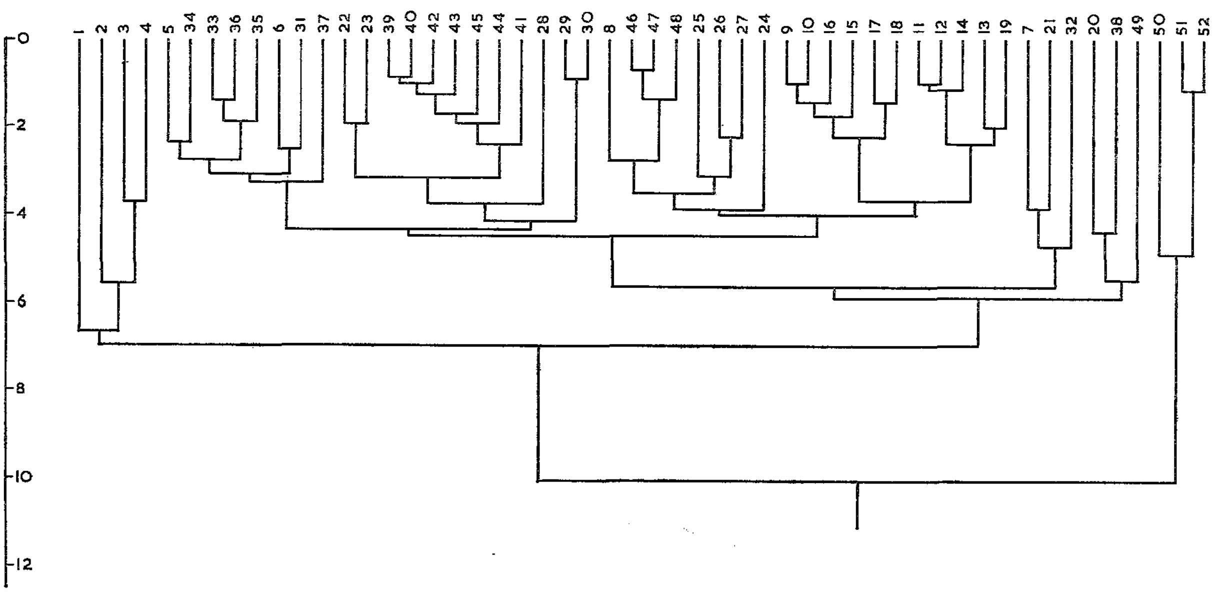

Fig. 4. The dendrogram resulting from a cluster analysis of the distance matrix calculated from the component specifications over the first 12 components, for group N. The vertical scale is indicated in distance units.

general, the aggregation of these species with the yellow-flowered *O. viscosa* group would seem not unreasonable. Although there is a tendency for flower colour to effect a separation within this group, it is clear that this is not the only significant character.

5. The conjunction of *O. antennata* (46–48) and *O. adenotricha* (8) is interesting in that the latter species is clearly rather different from the other members of Sub-section *Natrix*. The linking of these with Sub-section Biflorae (25–26), *O. ornithopodioides* and *O. polysperma*, would seem to bring together a group with aristate peduncles bearing usually two$\pm$yellow flowers and in several cases, numerous seeds in the pod. The *O. natrix* group joins on to this and there is a residue of 6 species, probably all of which, with the exception of *O. cintrana* (49), can be looked upon as having failed to cluster with their most appropriate groups. They mostly have one or a few unusual characteristics and these were probably sufficient to put them outside the range of the cluster-generating process (see p. 3).

An analogous series of operations were carried out on Group B. In this case, the characters most strongly associated with the first component are the breadth of the leaflet, the shape of the calyx and the length of the vexillum. OTUs with long vexilli, tubular or bilabiate calyces and relatively large leaflets are those in Sub-sections Verae and Crinitae and these are found at the negative end of this axis. For component 2, one character, the presence of dense glandular hairs on the vexillum, is especially prominent. Fig. 5 is a scatter diagram of the OTUs on the first two component axes, but again it is not possible to consider only the first two components in isolation and a distance matrix was calculated over the first 12 components and subjected to Cluster Analysis. The cophenetic correlation coefficient between the dendrograms and Sirjaev's classification were 0·443 for the weighted pair group method and 0·457 for the unweighted. Again, the unweighted method seems slightly preferable, though neither of these figures are high enough to be of particular significance.

Certain of the features of the unweighted dendrogram (Fig. 6) are similar to those previously mentioned; the infra-specific taxa tend to cluster together, *e.g.* of *O. spinosa* (69–76), *O. repens* (77–82) and *O. arvensis* (63–68). In the case of *O. spinosa*, there seems no clear-cut distinction between varietal and subspecific status—nos. 72–74 are varieties of ssp. *antiquorum* while no. 76 is ssp. *foetens*. Nos. 69, 70, and 71 are all varieties of ssp. *spinosa*. In both the other species such taxa are treated as varieties, and in general seem to be rather more different than some of the subspecies of *O. spinosa*. *O. masquillieri* (62) which has certain features, especially in the inflorescence, which distinguish it from the other members of this group, fails completely to cluster with it.

Other groups which are clearly indicated are Sub-section Chrysanthae (55–59) and Series Tuberculatae of Sub-section Diffusae (91–95). Smaller groups corresponding to Sirjaev's classification can also be found.

The evaluation of the larger groupings indicated in the dendrogram is rendered difficult by the fact that each tends to be formed by the addition of a reasonably homogeneous group to the population already generated. This 'chaining' tendency in the dendrogram may indeed reflect the fact that there seems to be a group of rather similar species (represented by those in the upper third of the dendrogram) while the remainder form relatively more dissimilar groups, though fairly similar within themselves.

(iv) *Cluster Analysis*

The application of Cluster Analysis to distance matrices calculated between the component specifications of OTUs is intended to simplify the interpretation of the results of Principal Components Analysis. The method is more usually applied to distance matrices obtained from the original data matrix, in which the OTUs are located in a character space by the values of their score for each character.

For several reasons, including those discussed under Principal Components Analysis, two separate distance matrices were calculated for the same Groups N and B, rather than a single matrix for the complete data. The results of both weighted and unweighted pair-group clustering on group N were compared with Sirjaev's classification and the cophenetic

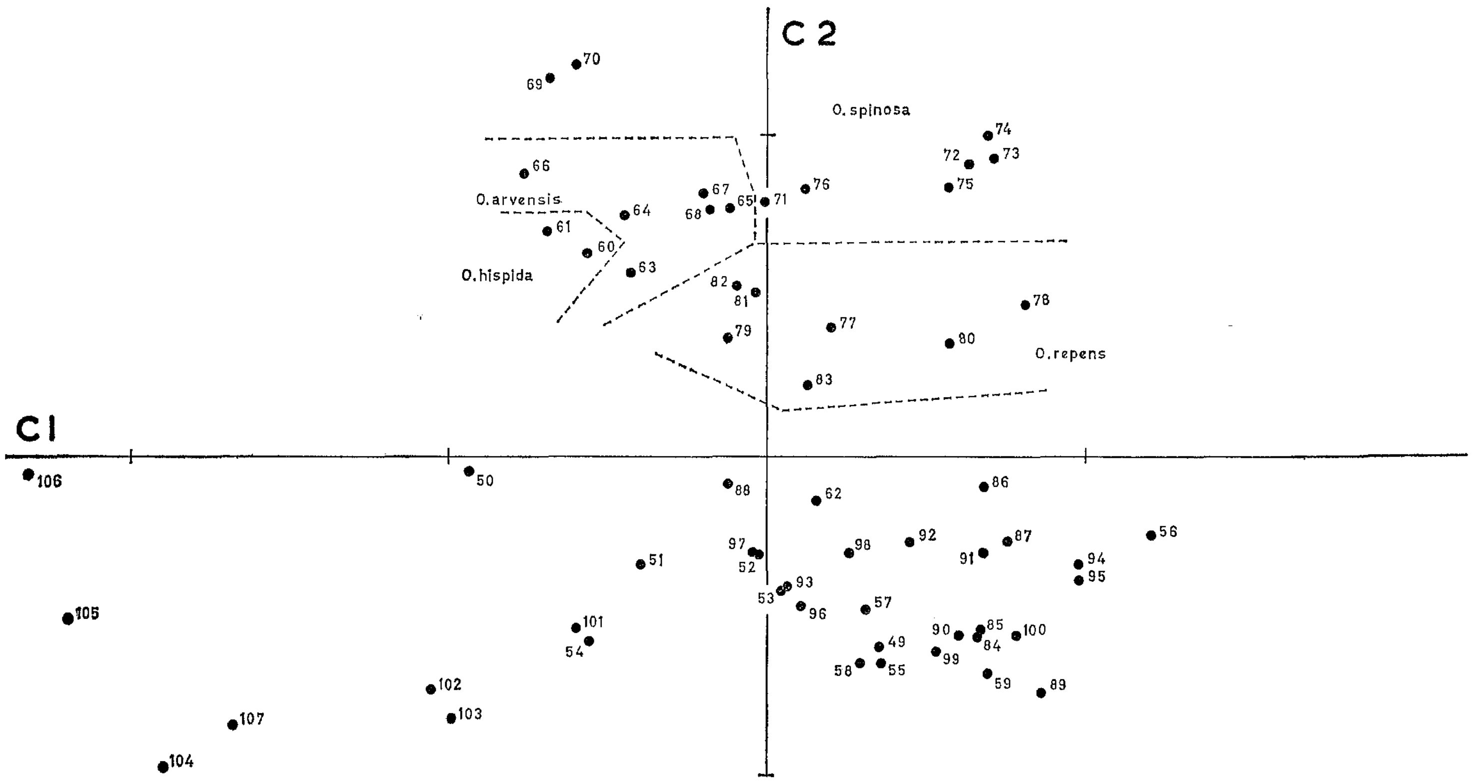

Fig. 5. A plot comparable to that in Fig. 3, but for group B.

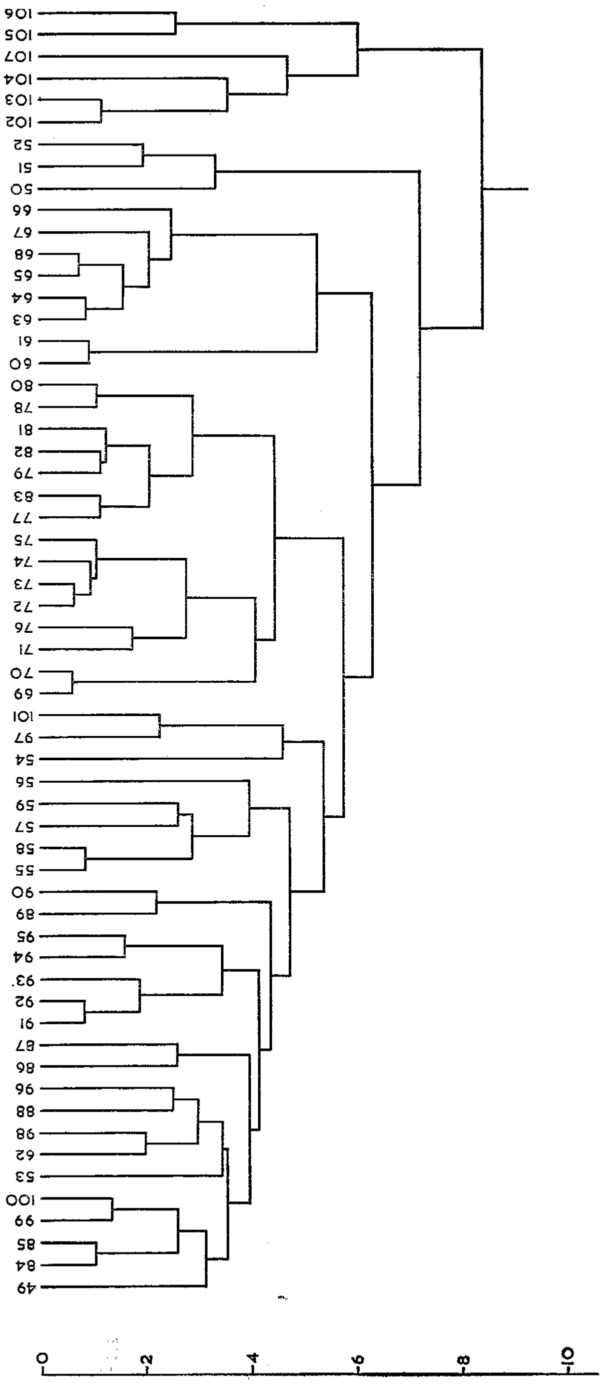

Fig. 6. The dendrogram comparable to that in Fig. 4, but for group B.

correlation coefficients were 0·523 for the weighted and 0·625 for the unweighted methods.

The dendrogram obtained by the unweighted method (Fig. 7) is remarkably similar to that obtained from the component specifications. Apart from differences in the distance scale and the fact that the OTUs are in a different order, the chief features of interest are the failure of *O. adenotricha* (8) to cluster with *O. antennata* (46–48) and the different (and perhaps more reasonable) position of *O. peyerimhoffii* (31). There is also a larger group of residual species which must be looked upon as joining where they do because of the exigencies of the method rather than as indications of phenetic similarity. This is hardly surprising when the large number of dimensions of the data matrix is borne in mind.

For Group B a similar comparison was made, the cophenetic correlation coefficients being 0·563 for the weighted and 0·698 for the unweighted methods.

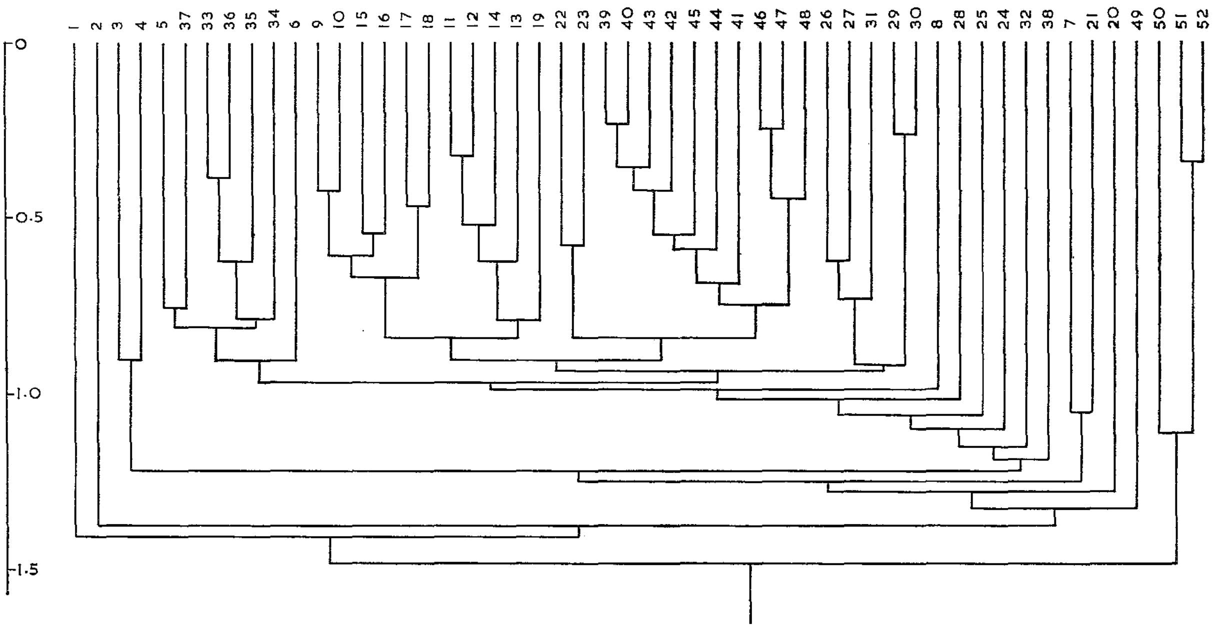

Fig. 7. The dendrogram resulting from the cluster analysis of the distance matrix calculated directly from the standardized data matrix, for group N. The vertical scale shows the distances at which the various groups cluster.

This dendrogram is shown in Fig. 8. The similarity between it and that from the component specifications is much less marked, although certain of the principal features recur: the aggregation of infra-specific taxa into species in the case of *O. repens*, *O. arvensis* and *O. spinosa*, the appearance of Sub-section Chrysanthae (50–52) and the aggregation of OTUs 102–107. Sub-section Bugranoides is represented lacking *O. striata* (59), but the recognition of other groupings is patchy, many of them being incomplete. The subsequent joining of these groups seems to have little meaning taxonomically and one is again forced to the conclusion that the clustering has taken place over too many dimensions to be meaningful.

4. Discussion

It is by now well realized that a computer is but the tool of the taxonomist and its ability to resolve classificatory problems is dependent not only upon the data presented to it but also upon the method employed to analyse those data. The work of the taxonomist

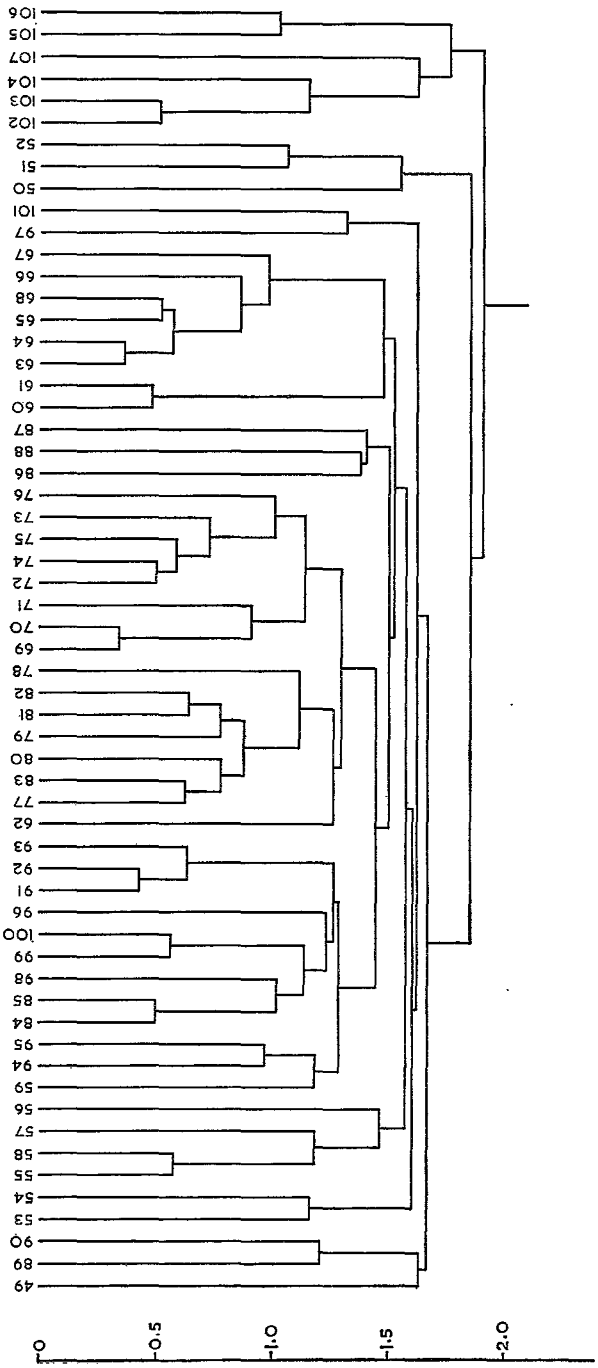

Fig. 8. The dendrogram comparable to that in Fig. 7, but for group B.

is firstly to select appropriate data, then to submit them to an acceptable analytical process or processes and, finally, to interpret the results.

It has been suggested that the introduction of numerical methods will remove from taxonomy the subjective element which has tended to obscure certain facets of classification. Rather, the subjective element has been diverted, though probably reduced somewhat in the process. It has become centred on the choice of characters from which an assessment of the similarity between species is to be made, rather than on the assessment itself. A poor selection of characters or an unsuitable choice of character states can change the entire course of an analysis.

The 64 characters used in this investigation would seem to be a minimum for this type of work, especially as 13 of them refer to the indumentum in some part of the plant; the results of the analyses thus tend to be orientated towards indumentum, which may account for some of the more surprising conclusions. Recent work on the classification of soils suggests, however, that a smaller number of relatively uncorrelated characters may be just as effective as a large group of unselected characters (Sarkar *et al.* 1966). There seems to be no obvious justification for an arbitrary weighting of floral over vegetative or other characters, but it is, of course, possible to introduce some measure of such weighting by increasing the number of floral characters used. A more detailed examination of flower structure would undoubtedly provide a considerable number of additional characters, *e.g.* pollen grain characters, which could be usefully employed. A further consideration is the extent to which flower colour, for example, is adequately described in terms of degrees of purpleness or yellowness or whether a series of states based on a chromatographic analysis of the pigments is not more appropriate. This opens up a considerable field utilizing biochemical characters which, unfortunately, are often only available if herbarium material does not have to be used.

The choice of character states also requires considerable care for, although standardization will largely remove the effect of different units of measurement, the use of unlimited categories for characters such as length, can introduce unwanted features. This was clearly seen in a preliminary analysis when untransformed length data were used—the very large leaflets and stipules characteristic of OTUs 102–107 distorted these dimensions of the character space so that they carried a disproportionate amount of weight in the analysis.

The question of transformation of the character states is one on which there has been little constructive comment from mathematicians. Theoretically, Principal Components Analysis requires that the correlation coefficients shall have been calculated from normally distributed data. Biological data are often not normally distributed and, whereas the application of logarithmic transformations to certain growth characters and of angular transformations to characters constrained to upper and lower limits are well known, there are certain types of character state which might require the most complex transformations to bring them to normality. In any event, it is impossible to transform any type of presence/absence data, which may make up a substantial part of a taxonomic data matrix.

Of the analytical methods used, Cluster Analysis is certainly the simplest agglomerative polythetic method and the construction of a two-dimensional dendrogram presents the information in an easily assimilable form. The reduction in the number of dimensions involves the loss of a substantial amount of information—indeed, all information concerning relationships other than those expressly indicated. It is basic to the method of Cluster Analysis used here that, at each stage, a revised distance matrix is calculated, treating clusters as new individuals and the distance between two such clusters as the mean of the distances between the individuals. The shape of the clusters so formed is important in assessing the efficiency of the method. Clustering by average linkage with only pairs of OTUs being considered for each cluster can lead to the centre of gravity of a cluster moving away from other OTUs which are similar to one of the individuals, yet not sufficiently similar to join the cluster immediately. Such OTUs are liable to become completely isolated and may eventually join a cluster to which they are not especially closely related. It was to meet this difficulty that the 'variable group' method of clustering was investigated, but its arbitrary limits are open to objection. The variable group method does prevent, at least in part, the tendency for OTUs to become separated from others to which they are related, though it is not able to deal with clusters which are not approximately hyperspherical. In the case of taxonomic data, the possibility of elongated clusters is a real one, though it is difficult to devise a method which will ensure that clusters of such a shape are located accurately without including unrelated OTUs as well. It is also worth noting that in every case the unweighted cluster analysis has given results more in keeping with the accepted classification than the weighted analysis.

One of the problems which arises when the results of different methods are to be compared is the mechanism of the comparison. The cophenetic correlation coefficient has been suggested as an appropriate measure for the comparison of dendrograms (Sokal & Sneath 1963). While this will always recognize two identical dendrograms, the amount of dissimilarity permitted before the dendrograms cease to be identical must depend on the precision of the comparison. The values of the correlation coefficient presented above were calculated on the assumption that the two Sections of *Ononis* represented taxa at the 50-phenon level and the range above this was divided into 10 parts, *i.e.* 5-phenon categories.

A general comparison of these results with the classification of *Ononis* put forward by Sirjaev suggests that there are few significant discrepancies. It would be unreasonable to expect a very close correlation and it is probably fair to conclude that there are few differences that need to be accounted for; this is not to say that some aspects of the classification might not be reconsidered. On the basis of the present work, which was intended chiefly to investigate methodology, it would be inappropriate to make any taxonomic changes, though a case could, perhaps, be made for the erection of two new Sections, one to contain Sub-section Antiquae (*O. rotundifolia*) and the other to contain Sub-sections *Salzmannianae*, *Verae* and *Crinitae*. Both new sections could be readily defined on morphological grounds and this would produce four Sections of roughly comparable standing.

A most promising method for taxonomic work would appear to be Principal Components Analysis, with the subsequent calculation of the component specifications. This seems to be more appropriate than the calculation of taxonomic distance, especially if the number of characters, and hence the dimensions of the distance matrix, is large. It also avoids the problem of the heterogeneity of column vectors. The use of Cluster Analysis to investigate the structure of the component specification matrix also seems useful and the loss of information should be considerably less than the direct investigation of the distance matrix calculated from the data, since there are many fewer dimensions to be considered.

Principal Components Analysis defines orthogonal axes; rotation of these axes can produce interesting and often useful results. Such rotation is standard practice with Factor Analysis and was employed in a taxonomic study with considerable success by Rohlf & Sokal (1962) to produce groups of individuals characterized by 'group factors' rather than the more generalized factors of the unrotated matrix. The rotation of axes defined by Principal Components Analysis has been carried out less frequently but an application to ecological data suggests that it may prove to be valuable; it has not, however, been carried out in connection with the analyses described in this paper.

References

EADES, D. C. (1965). The inappropriateness of the correlation coefficient as a measure of taxonomic resemblance. *Syst. Zoo.* **14**, 98–100.

KENDRICK, W. B. (1964). Quantitative characters in Computer Taxonomy. In *Phenetic and Phylogenetic Classification. Publ. Syst. Assn.* **6**, 105–114.

LANCE, G. N. & WILLIAMS, W. T. (1966). A generalized sorting strategy for computer classifications. *Nature* **212**, 218.

ROHLF, F. J. & SOKAL, R. R. (1962). The description of taxonomic relationship by factor analysis. *Syst. Zoo.* **11**, 1–16.

SARKAR, P. K., BIDWELL, O. W. & MARCUS, L. F. (1966). Selection of characteristics for numerical classification of soils. *Soil Sci. Soc. Amer. Proc.* **30**, 269–271.

SIRJAEV, G. (1932). Generis *Ononis* revisio critica. *Beih. Bot. Centr.* **49** (2), 381–665.

SOKAL, R. R. & SNEATH, P. H. A. (1963). *Principles of Numerical Taxonomy.* Freeman.

SOKAL, R. R. & MICHENER, C. D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **38**, 1409–1438.

WILLIAMS, W. T. & LAMBERT, J. M. (1959). Multivariate methods in plant ecology. I. Association analysis in plant communities. *J. Ecol.* **47**, 83–101.

WILLIAMS, W. T. & LAMBERT, J. M. (1960). Multivariate methods in plant ecology. II. The use of an electronic digital computer for association analysis. *J. Ecol.* **48**, 689–710.

WILLIAMS, W. T. & LAMBERT, J. M. (1961). Multivariate methods in plant ecology. III. Inverse association analysis. *J. Ecol.* **49**, 717–729.

WILLIAMS, W. T., LAMBERT, J. M. & LANCE, G. N. (1966). Multivariate methods in plant ecology. V. Similarity analyses and information-analysis. *J. Ecol.* **54**, 427–445.

APPENDIX 1. Conspectus of the Genus *Ononis* L. (after Sirjaev 1932).

		OTU Number
NATRIX		
Raceme pedunculate, 1–3 flowered; fruiting pedicel reflexed; pod linear or oblong		
Antiquae		
Subshrub, leaves and bracts 3-foliolate; vexillum hairy; 10th stamen adnate at the base only	*rotundifolia*	1
Rhodanthae		
Shrub or perennial; leaves ternate, coriaceous; peduncle shortly aristate; vexillum hairy		
Frutescentes		
Shrub; bracts bract-like; peduncle short, 1–3 flowered	*tridentata*	2
	fruticosa	3
Perennes		
Perennial; bracts leafy; peduncle mucronate, 1-flowered, occasionally with 2	*cristata* (*cenisia*)	4
Canarienses		
Shrub; leaves 3-foliolate; peduncle short, muticous; corolla rose, glabrous; alae without teeth; seeds tuberculate..	*christii*	5
Mauritanicae		
Shrub; leaves pinnate, bracts 3-foliolate; corolla rose, glabrous; alae without teeth	*thomsonii*	6
Eu-natrix		
Shrub or perennial; leaves usually 3-foliolate, but variable; bracts variable; corolla yellow, sometimes red-striped, glabrous; alae with teeth		
Atlanticae		
Peduncle spinous-aristate; stipules not connate	*atlantica*	7
Orientales		
10th stamen one-third adnate; lower leaves pinnate; peduncle usually 2-flowered	*adenotricha*	8
Polymorphae		
Peduncle unarmed; stipules not connate; 10th stamen two-thirds adnate	*natrix*	9–19
	crispa	20
Aegypticae		
Shrub; stipules connate; peduncle subspiny; leaves 3-foliolate	*vaginalis*	21
Serotinae		
Perennial; leaves 1-foliolate; peduncle 1-flowered, muticous or shortly aristate	*serotina*	22
	pseudoserotina	23
Torulosae		
Annual; corolla yellow, glabrous; pod torulose; seeds tuberculate ..	*ornithopodioides*	24

Biflorae		
Annual; peduncle 2-flowered, aristate; corolla yellow, glabrous		
Tuberculatae		
Seeds tuberculate	*biflora*	25
Laeves		
Seeds smooth	*hebecarpa*	26
Pilosae		
Leaves 3-1-foliolate; corolla rose and yellow; vexillum hairy; alae without teeth	*polysperma*	27
	maweana	28
Reclinatae		
Annual; peduncle 1-flowered; vexillum glabrous; alae without teeth		
Siculae		
Corolla yellow; peduncle aristate; seeds tuberculate	*sicula*	29–30
Sublaeves		
Corolla yellow; peduncle aristate; seeds nearly smooth	*peyerimhoffii*	31
Incisae		
Corolla yellow; peduncle shortly aristate; seeds rugose; stipules incised	*incisa*	32
Eu-reclinatae		
Corolla rose or whitish; peduncle muticous; seeds tuberculate	*reclinata*	33
	pendula	34
	laxiflora	35
	dentata	36
	verae	37
Viscosae		
Annual; leaves mostly 3-foliolate; corolla yellow, glabrous; alae with teeth		
Pubescentes		
Upper bracts bract-like; peduncle muticous; seeds smooth	*pubescens*	38
Eu-viscosae		
Bracts leafy; peduncle aristate; seeds tuberculate	*viscosa*	39–44
	crotalarioides	45
Antennatae		
Leaves all 3-foliolate; peduncle aristate; seeds tuberculate	*antennata*	46–48
BUGRANAE		
Peduncle absent; pod ovate, rarely reflexed		
Intermediae		
Annual; vexillum hairy; pod reflexed; seeds tuberculate	*cintrana*	49
Chrysanthae		
Shrub; leaves 3-foliolate, bracts bract-like; flowers yellow; seeds smooth		
Speciosae		
Leaflets subcoriaceous; raceme dense	*speciosa*	50
Aragonenses		
Leaflets coriaceous; raceme interrupted	*aragonensis*	51
	reuteri	52
Pinnatae		
Annual, shrubby; leaves pinnate; corolla rose, glabrous; seeds smooth	*leucotricha*	53
	pinnata	54
Bugranoides		
Shrub or perennial; leaves 3-foliolate; corolla yellow, glabrous.. ..	*pusilla*	55
	(*columnae*)	
	minutissima	56
	cephalotes	57
	saxicola	58
	striata	59

Acanthononis		
Shrub or perennial, often spiny; leaves variable; corolla rose, hairy; seeds rarely smooth		
Arborescentes		
Small unarmed shrub	*hispida*	60–61
Vulgares		
Shrub, often spiny	*masquillieri*	62
	arvensis	63–68
	(*hircina*)	
	spinosa	69–76
	repens	77–83
Villosissimae		
Annual; leaves 1-3-foliolate; corolla rose, glandular-hairy; seeds tuberculate		
Trifoliatae		
Leaves all 3-foliolate	*filicaulis*	84
	villosissima	85
Monophyllae		
Leaves all 1-foliolate	*monophylla*	86
	oligophylla	87
	alba	88
Variegatae		
Annual; leaves 1-foliolate; flowers in lax racemes; corolla yellow, hairy	*variegata*	89
	euphrasiifolia	90
Diffusae		
Annual; leaves 3-foliolate; corolla rose or white, usually hairy; raceme elongating after anthesis		
Tuberculatae		
Vexillum hairy; seeds tuberculate	*diffusa*	91–93
	serrata	94
	phyllocephala	
	tournefortii	95
Cossonianae		
Vexillum glabrous; seeds smooth	*hirta*	96
	cossoniana	97
	cephalantha	98
Subspicatae		
Vexillum glabrous; seeds tuberculate	*subspicata*	99–100
Mitissimae		
Annual; leaves 3-foliolate; upper bracts imbricate, coriaceous; corolla rose, glabrous; calyx tubular; seeds tuberculate	*mitissima*	101
Salzmannianae		
Annual; leaves 3-1-foliolate; calyx tubular; corolla rose, glabrous; seeds smooth	*alopecuroides*	102
	baetica	103
Verae		
Annual; leaves 3-foliolate; upper bracts bract-like; calyx tubular; corolla large, rose; vexillum hairy; seeds tuberculate	*rosea*	104
	avellana	105
	megalostachys	106
Crinitae		
Annual; leaves and bracts 3-foliolate; calyx bilabiate, tubular; corolla whitish, glabrous; seeds tuberculate	*crinita*	107

Appendix 2

	Character	No. of states	Range		Attributes*
			from	to	
1	Duration	5	herbaceous annual	shrubby perennial	1–4
2	Stature	4	erect	procumbent	5–7
3	Branch spines	3	absent	strong	8–9
4	Size (L)	5	under 15 cm	over 100 cm	10–14
5	Stipule length (L)	∞	0·5 mm	23 mm	15–19
6	Stipule shape	4	linear	rhomboid	10–22
7	Stipule base	6	free	infundibuliform	23–27
8	Stipule margin	5	entire	incised	28–31
9	Petiole length (L)	∞	1 mm	30 mm	32–34
10	Leaflet length (L)	∞	3 mm	45 mm	35–37
11	Leaflet breadth (L)	∞	1·3 mm	25 mm	38–40
12	Leaflet apex	4	acute	retuse	41–43
13	Leaflet margin	3	entire	dentate	44–45
14	Leaflet texture	3	coriaceous	fleshy	46–47
15	No. of leaflets (*upper leaves*)	∞	1 leaflet	5 leaflets	48–50
16	No. of leaflets (*lower leaves*)	∞	1 leaflet	7 leaflets	51–57
17	Short hairs on leaflet	5	none	dense	58–61
18	Glandular hairs on leaflet	5	none	dense	62–65
19	Sessile glands on leaflet	3	none	dense	66–67
20	Leaflet crisped	2	no	yes	68
21	Form of bracts	4	squamulose	3 leaflets	69–71
22	Peduncle length (L)	∞	absent	40 mm	72–76
23	Pedicel length (L)	∞	0·5 mm	10 mm	77–80
24	Flowers/peduncle	∞	1 flower	3 flowers	81–82
25	Raceme density	5	very lax	capitate	83–86
26	Arista length (L)	∞	absent	15 mm	87–90
27	Arista spinous	2	no	yes	91
28	Calyx shape	2	campanulate	tubular	92
29	Calyx teeth in fruit	3	spreading	accrescent	93–94
30	Calyx tube length	∞	1·5 mm	6·5 mm	95–96
31	Calyx teeth length	∞	2·5 mm	12·5 mm	97–99
32	Calyx teeth width at base	∞	0·5 mm	3·5 mm	100–102

	Character	No. of states	Range		Attributes*
			from	to	
33	Nerves/calyx tooth	∞	3 nerves	13 nerves	103–104
34	Short hairs on calyx	5	none	dense	105–108
35	Long hairs on calyx	5	none	dense	109–112
36	Short glandular hairs on calyx	5	none	dense	113–116
37	Long glandular hairs on calyx	5	none	dense	117–120
38	Anthocyanin in petals	5	none	purple	121–124
39	Flavonol in petals	5	none	golden	125–128
40	Standard length (L)	∞	6 mm	26 mm	129–133
41	Standard breadth (L)	∞	4 mm	16 mm	134–138
42	Hairs on standard	5	none	dense	139–142
43	Glandular hairs on standard	5	none	dense	143–146
44	Shape of standard apex	4	emarginate	acute	147–149
45	Form of standard apex	2	not apiculate	apiculate	150
46	Length/breadth ratio of alae	∞	1·3	4	151–153
47	Alar tooth	2	absent	present	154
48	Length/beak ratio of keel	∞	1·0	3·0	155–156
49	Keel beak angle	3	acute	obtuse	157–158
50	Stamen adnation	∞	free	joined	159–161
51	Pod length (L)	∞	4 mm	32 mm	162–167
52	Pod breadth (L)	∞	2 mm	8·5 mm	168–171
53	Hairs on pod	5	none	dense	172–175
54	Glandular hairs on pod	5	none	dense	176–179
55	Compression of pod	5	inflated	torulose	180–183
56	Seed size (L)	∞	0·75 mm	5 mm	184–186
57	Seeds/pod (L)	∞	1 seed	40 seeds	187–189
58	Seed surface	6	smooth & shining	echinate	190–194
59	Pod position	3	erect	pendent	195–196
60	Short hairs on stem	5	none	dense	197–200
61	Long hairs on stem	5	none	dense	201–204
62	Short glandular hairs on stem	5	none	dense	205–208
63	Long glandular hairs on stem	5	none	dense	209–212
64	Calyx form	3	teeth equal	bilabiate	213–214

* These numbers refer to the attributes used in both Association and Information Analyses.

∞ No constraint was applied to the number of states of these characters and they were measured in the units stated.

(L) These character states were subjected to a logarithmic transformation before analysis of the data.