Vidyodaya J., Arts, Sci., & Lett., Vol. 14, No. 1 —*January* 1986 —*pp*. 81 - 90

A COMPARATIVE STUDY OF HIERARCHICAL AGGLOMERATIVE CLUSTERING STRATEGIES

By

P. WICKRAMAGAMAGE

(*B.A. Sri Lanka, M.Sc., Ph.D London*) *University of Peradeniya*
Department of Geography

Introduction

The hierarachical agglomerative strategies have been by far the most widely used clustering methods, especially in biological taxonomy (Sneath and Sokal 1973, p. 214). Although these methods have been developed for classifying hierarchical populations, they can still be useful as a strategy to obtain an initial partition of a population for non-hierarchical classification. Therefore, it is important to examine the nature and properties of these methods and their inter-relationships. The agglomerative sorting strategies require an inter-individual similarity (dissimilarity) matrix and involve sorting of similar individuals into groups (clusters) by successive fusion. This process is generally continued until all individuals and groups are fused to form a hierarchical tree which is graphically represented by a dendrogram. Different agglomerative strategies differ from each other in the way fusion of groups occur.

The agglomerative clustering strategies can be compared in two important ways:

(*a*) optimality of classification with respect to a statistical criterion

(*b*) goodness-of-fit (distortion).

The global optimality of a given classification may not be determined by existing methods, but classifications obtained from different methods can be compared to determine the "best classification" amongst them (Webster 1971)

Goodness-of-fit of a given classification depends upon to the extent to which the original relationships are preserved. The inter-individual, similarity matrix undergoes distortion in the course of fusion and the degree of this distortion varies from strategy to strategy. Different agglomerative strategies applied to the same inter-individual similarity matrix may produce mutually incompatible classifications. This study attempts to explain the nature of this variation and its effect on the resulting classification.

2. Methods and Data

Squared Euclidean distance was used to generate an inter-individual distance matrix. This measure was preferred to other such methods for its simplicity and the allowance for missing data. The distance between ith and jth individuals (dij) is given by the following function ;

$$d_{ij} = \frac{1}{p}\sum_{k=1}^{P} (x_{ik} - x_{jk})^2$$

Where P — number of attributes common to both individuals

X — attribute value (standardized to unit variance and zero mean)

Squared Euclidean distance between ith and jth individuals was calculated using attributes common to both individuals and the effect of missing values was eliminated by dividing distance coefficients by the number of attributes common to both individuals (p.)

The classification was done by seven hierarchical agglomerative strategies (Table 1). These clustering strategies were applied on the distance matrix calculated using Squared Eulclidean distance. The classification process begins by fusing the most similar individuals or groups and works upward through the heirarchy until all individuals and groups are joined together to form a single super group. The seven strategies (Table 1) differ from each other in the way in which the similarity between groups is calculated. However, all the strategies are combinatorial in that the original distance matrix need not be preserved throughout the classificatory process.

A general formula for all Seven agglomerative strategies has been proposed by Lance and Williams (1967). according to this formula, the distance between a group just formed by the fusion of ith and jth individuals and any other group k is defined as follow.

$$d_{k(ij)} = \propto_i d_{ik} + \propto_j d_{jk} + \beta\, d_{ij} + \gamma \mid d_{ki} - d_{kj} \mid$$

Where, $\propto_i$, $\propto_j$, β and γ are parameters specifying a paiticular clustering strategy. The values of these parameters for the seven clustering strategies used in this study are listed in Table 1.

Table 1

Values of the parameters for seven Strategies

	Strategy	$\propto i$	$\propto i$	β	γ
1.	Single Linkage Method	0.5	0.5	0	$-.5$
2.	Complete Linkage Method	0.5	0.5	0	.5
3.	Average Linkage Method	$n_i/n_i + n_j$	$n_j/n_i + n_j$	0	0
4.	Centroid Method	$n_i/n_i + n_j$	$n_j/n_i + n_j$	$-n_i n_j/(n_i + n_j)^2$	0
5.	Median Sort Method	0·5	0·5	0·25	0
6.	Ward's Error Sum-of-Squares Method	$(n_k + n_i)/N$	$(n_k + n_j)/N$	$-n_k/N$	0
7.	Lance—Williams Method	$1-(\beta + \propto j)$	$\propto i$	$1-\propto i + \propto j$	0

$N = n_{i+}n_j \quad n_k$

A large number of clustering strategies can be derived by changing the values of the parameters.

Goodness-of-fit or the distortion of each clustering strategy was examined by cophenetic correlation defined by Sokal and Rohlf (1962). This is in fact product-moment correlation between the original similarity matrix and the new similarity matrix obtained from dendrograms. The relative similarity between individuals undergoes changes during the classificatory process, and the degree of this distortion varies from strategy to strategy. Product-moment correlation was calculated using a sample of 30 distance coefficients randomly drawn to reduce the computational load involved in the use of the whole population.

Data on ten soil properties (Table 2) for 32 soil profiles obtained from the published records of USDA (1975) were used in this analysis. The selection of soil profiles was done on the basis of availability of quantitative data for the ten soil properties used.

Table 2

Soil properties used to compute inter-individual distance matrix

1. Percentage Silt
2. Percentage Clay
3. Percentage Organic Carbon
4. Percentage Dithionite Extractable Iron as Fe
5. pH (1:1 soil/water suspension)

6. Exchangeable Ca me/100g soil
7. Exchangeable Mg me/100g soil
8. Exchangeable Na me/100g soil
9. Exchangeable K me/100g soil
10. Cation Exchange Capacity (C.E.C.) me/100g soil

The collection of data on soil profiles is normally done after dividing it into a series of horizons. In the United States, the soil profile is divided into three Master horizons, and each of them is further subdivided to achieve greater homogeneity. In this study a primary data reduction was achieved by taking only the mean attribute values for the three Master horizons. Since soil depth levels tend to be correlated (Wickramagamage 1982), this is particularly desirable when Euclidean distance is used as the similaritymeasure, which requires the attribute vectors to be mutually independent (uncorrelated). This method does not discard too much information compared to the method which takes only the mean attribute values for the entire soil profile.

3. Results and Discussion

The classifications produced by the seven strategies are shown by dendrograms (Fig. 1—7). The dendrograms drawn for the classification produced by Ward's ESS method (Fig. 6) shows well defined clusters whereas the single linkage method (Fig. 1) shows no clusters. The other dendrograms fall in between those two extremes.

The cophenetic correlation coefficients (r_c) calculated for the seven strategies are listed together with inter-strategy correlation coefficients in Table 3. It can be seen that goodness of-fit of the strategies tends to vary considerably as indicated by r_c (0.38—0.76).

Table 3

Cophenetic correlation coefficients and inter-strategy correlation coeffcients

Strategy	r_c	Inter-strategy Correlation					
		1	2	3	4	5	6
1	0.76						
2	0.47	0.40					
3	0.74	0.90	0.55				
4	0.73	0.98	0.42	0.91			
5	0.38	0.61	0.55	0.60	0.64		
6	0.44	0.33	0.62	0.31	0.24	0.36	
7	0.41	0.71	0.60	0.75	0.74	0.80	0.46

on the basis of cophenetic correlation, the seven clustering strategies considered here can be divided into two distinct classes (Table 4). The strategies in A have less distortion than those in B.

Table 4

Classification of agglomerative strategies on the basis of cophenetic correlation

Class	Strategies	r_c
A	Single Linkage Method	0.76
	Average Linkage Method	0.74
	Centroid Method	0.73
B	Complete Linkage Method	0.47
	Ward's ESS Method	0.44
	Lance -Williams Method $\beta = 0{\cdot}00$	0.41
	Median Sort Method	0.38

It can be seen from Table 3 that the strategies in A are highly correlated to each other compared to those in B ($r \geqslant 0.90$). The clusters produced by all three methods in A are very similar to each other with respect to their composition. Therefore, it can be concluded that these three methods produce similar classifications and also they tend to preserve original similarities between individuals. It must be noted here that Centroid Method is likely to suffer from reversing (Webster 1977) due to tempoıary rise in similarity as the hierarchy develops making it unsuitable for classification.

The strategies in B are not very similar to each other as indicated by the inter-strategy correlation matrix (Table 3). However, certain strategies seem to have a higher correlation than the others. For example, Median Sort Method and Lance williams Method have a high correlation of 0.80 and the dendrograms produced by these two strategies have a considerable resemblance. They are the two strategies which had the lowest cophenetic correlation (Fig. 5 & 7).

The strategies in A have less distortion than those in B as has been indicated by cophenetic correlation. Generally these methods (A) tend to suffer from chaining and do not produce well defined clusters. However, Average Linkage Method is capable of producing tight clusters when the population is strongly structured. The methods in B tend to produce better clusters as can be seen from the dendrograms (Fig. 2, 5, 6 & 7) ; Ward's ESS Method is important in this respect. It is well known to produce clear clusters and the groups tend to have greater homogeneity (Wickramagamage 1982).

The results reported above show that goodness-of-fit may be maintained at the expense of the clarity of clusters. Single Linkage Method, which has the least distortion, has failed to show clear clusters. When a great emphasis is placed on the original relationships, a certain degree of chaining is unavoidable. Therefore, the strategies which have the least distortion may have a very limited use in classification of natural populations, perhaps as a test for misclassification. At lower levels of the hierarchy, in all dendrograms fusion of individuals occurs in the same manner, but they take diffierent courses when groups are fused. Therefore, the difference between strategies occur at higher levels of the hierarchy. Ward's ESS Method not only produce clear clusters but also have an added advantage of being the only agglomerative strategy which proceeds through fusion of individuals and groups by minimizing within-group variance. Therefore, the groups produced by this method tend to be more homogeneous than those produced by other methods considered here. It may be argued that it is really not necessary to preserve the original relationships unless the hierarchy is of interest. The relative similarity between individuals should be determined at the population level and not by taking pairs of individuals seperately. On the other hand, it is not possible to produce clearly defined clusters and preserve the original similarities at the same time. Therefore, it can be concluded that intensely clustering strategies are preferable to space preserving ones.

References

Lance, G. N. and Williams, W. T. (1967) "A general theory of classificatory sorting Strategies : I Hierarchical systems", *comp. Journal*, p. 373—380.

Sneath, P. H. A. and Sokal, R. R. (1973) *Numerical Taxonomy*, Freeman, San Francisco, 573pp.

Sokal, R. R. and Rohlf, F. J. (1962) "The comparison of dendrograms by objective methods" *Taxon*, vol. 11, p. 33.40

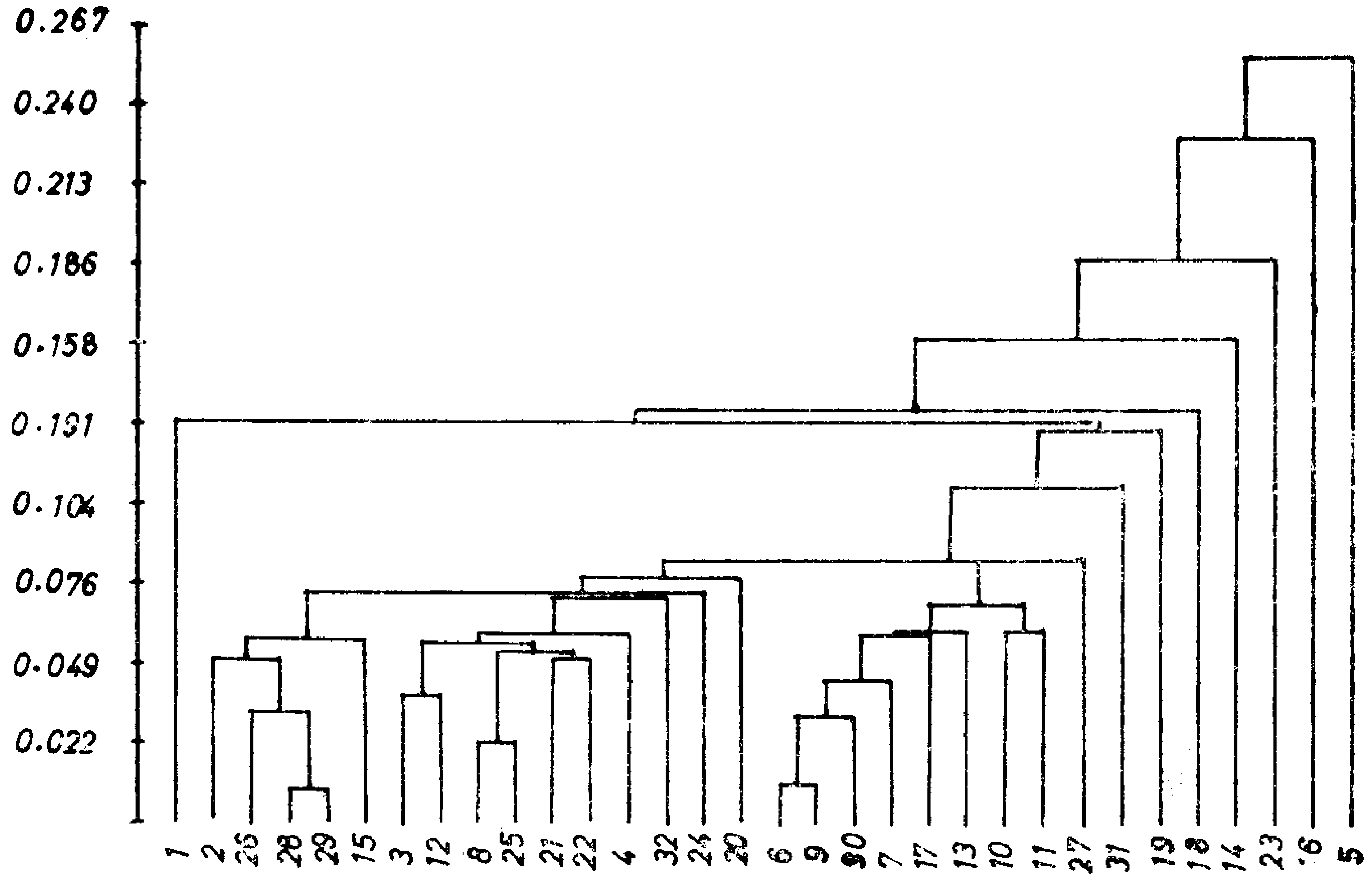
Webster, R. (1971) "Wilk's Criterion : A measure for comparing the value of general purpose soil classifications". *J. Soil Sci*, 22, p. 254—260.

——(1977) *Numerical and Qantitative Methods in Soil Classification*, Oxford Univ. Press, pp. 279.

Wickramagamage, P. (1982) *Studies in Numerical Taxonomy of Soils*, (Unpublished Ph.D. thesis, Dept. of Geography, University of London).

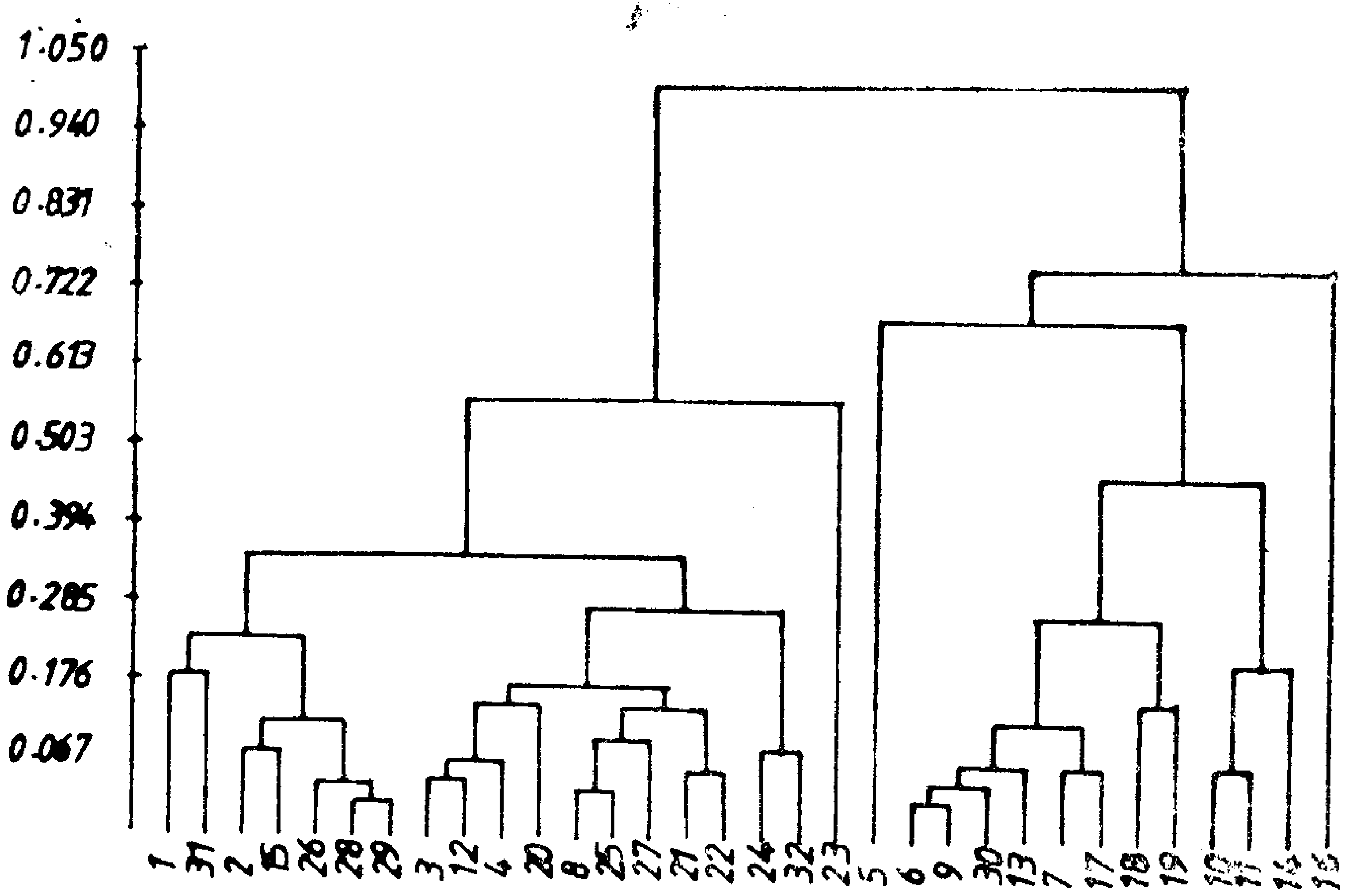
USDA (1975) *Soil Taxonomy : A basic System of Soil Classification for Making and Interpreting of Soil Survey*, Government Printing Office, Agriculture Handbook, 436 pp.

FIG : 1



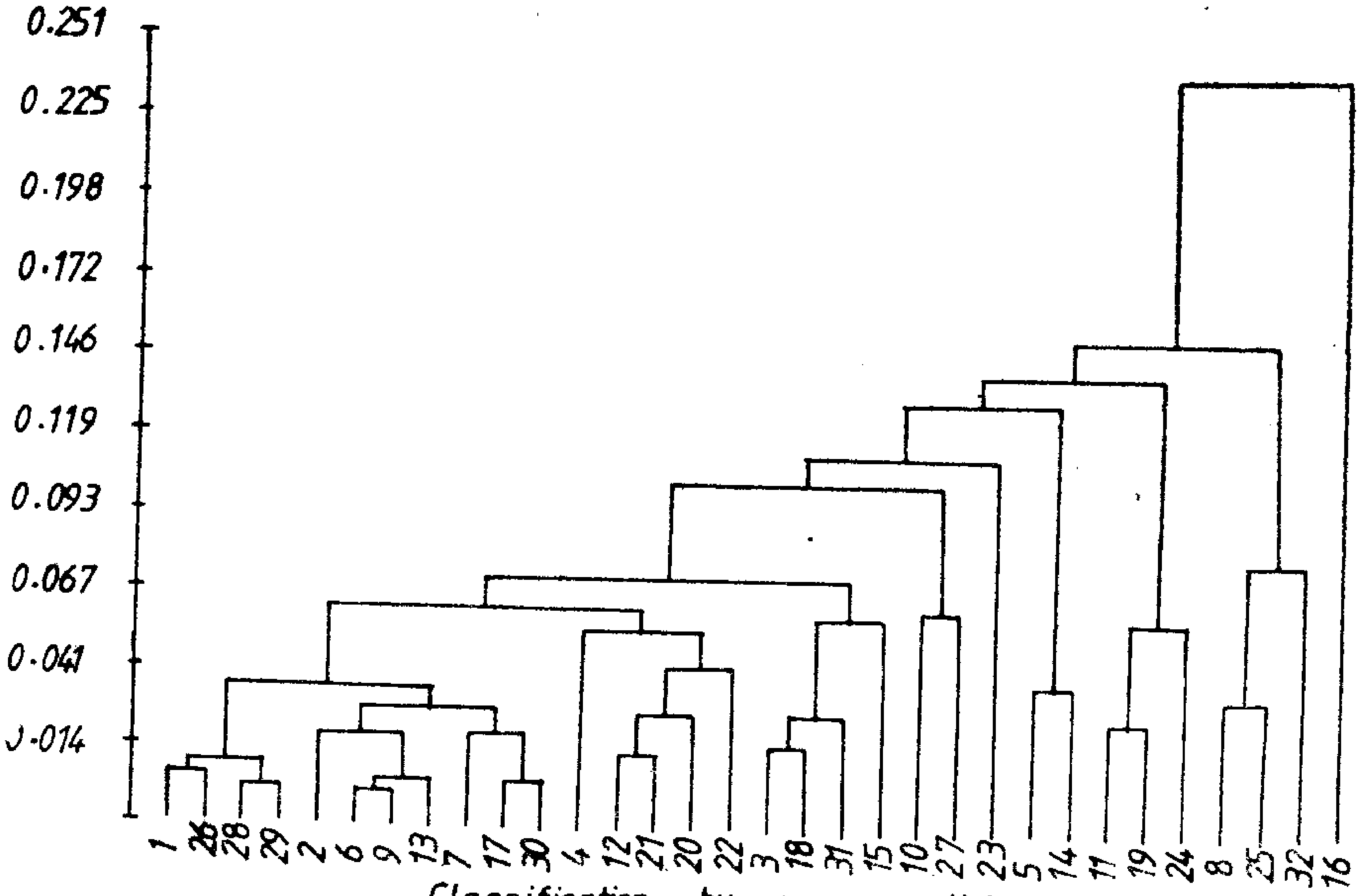
Classification by single linkage method

FIG 2



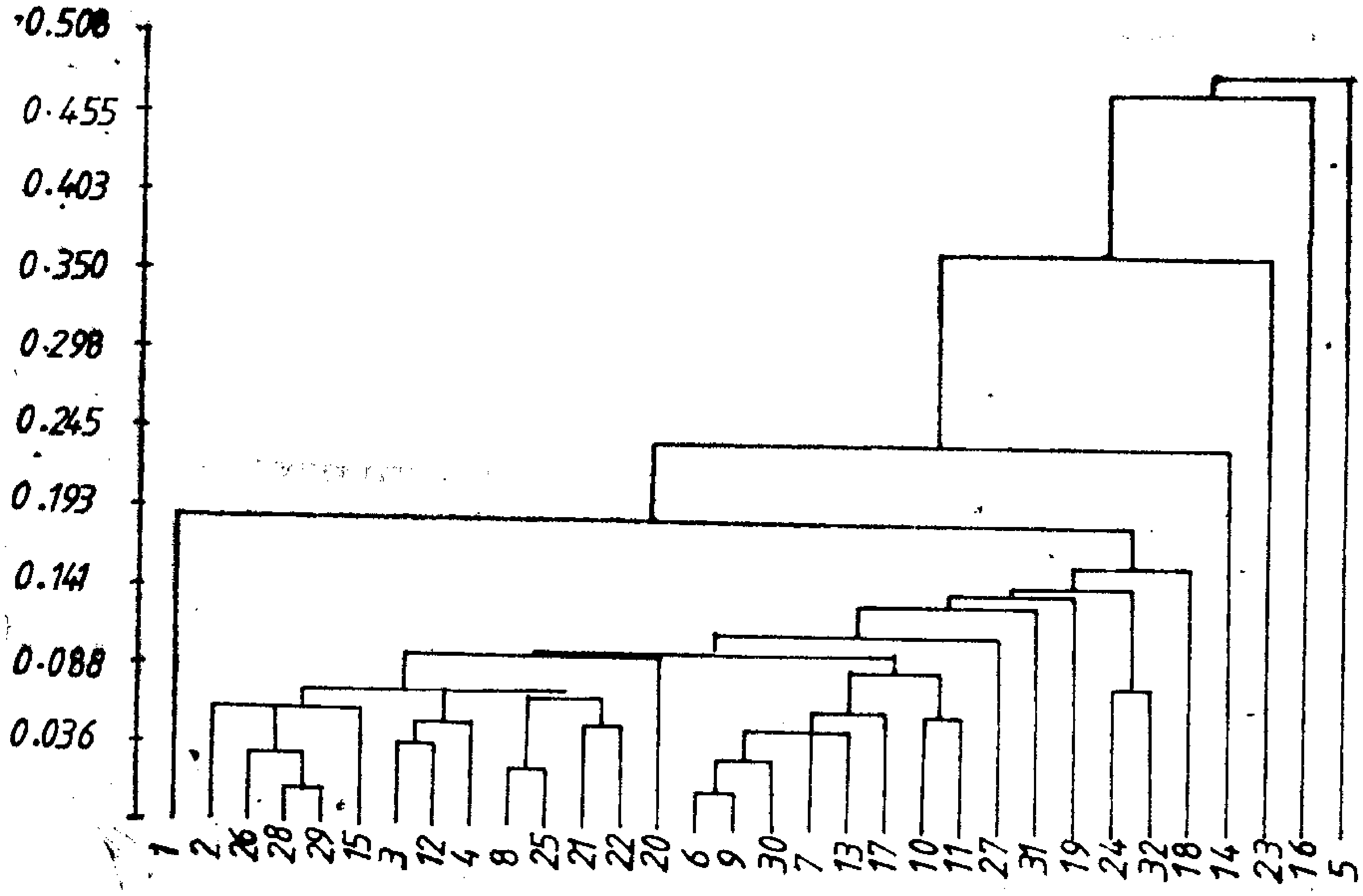
Classification by complete linkage method

FIG: 3



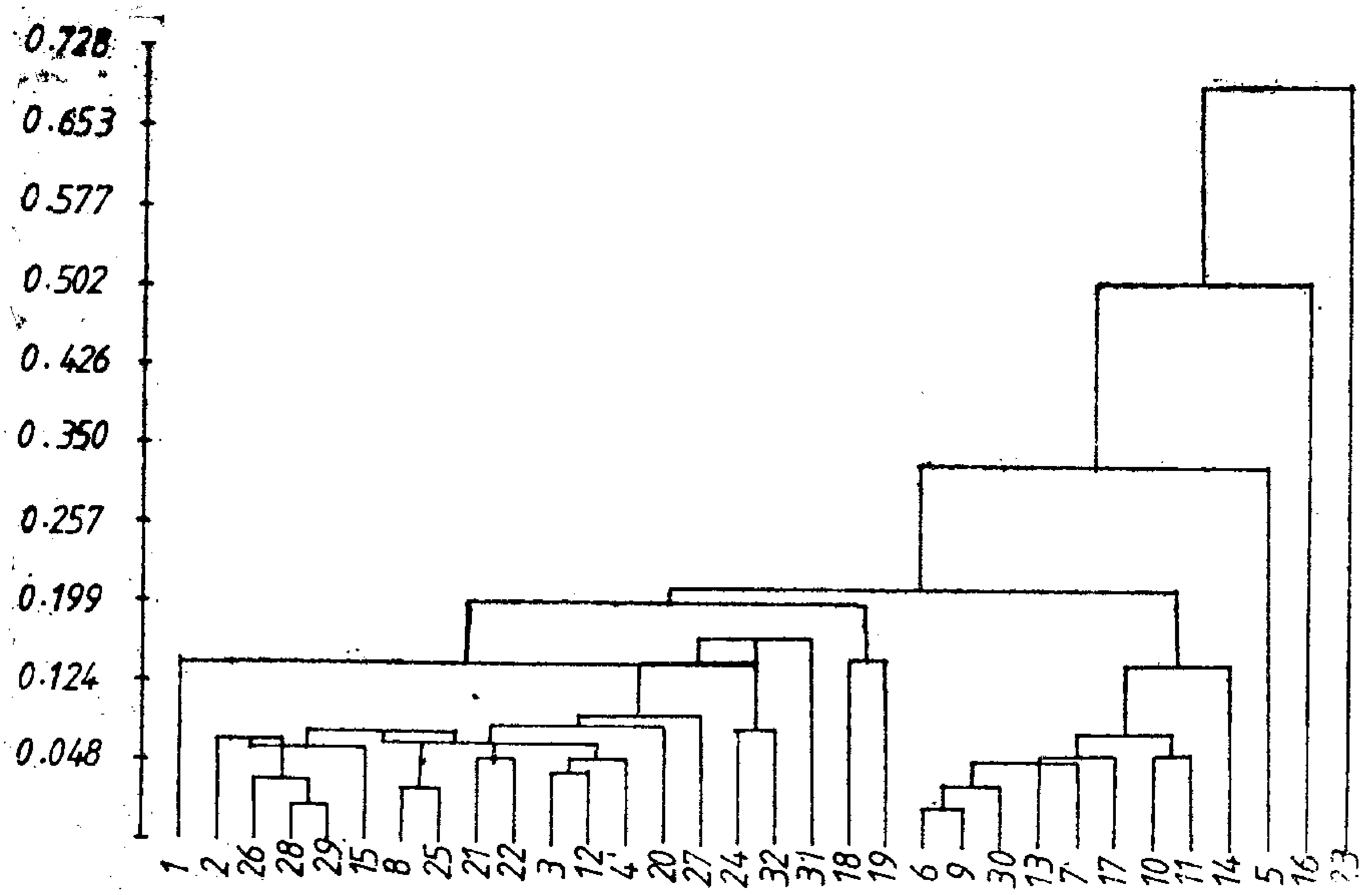
Classification by average linkage method

FIG: 4



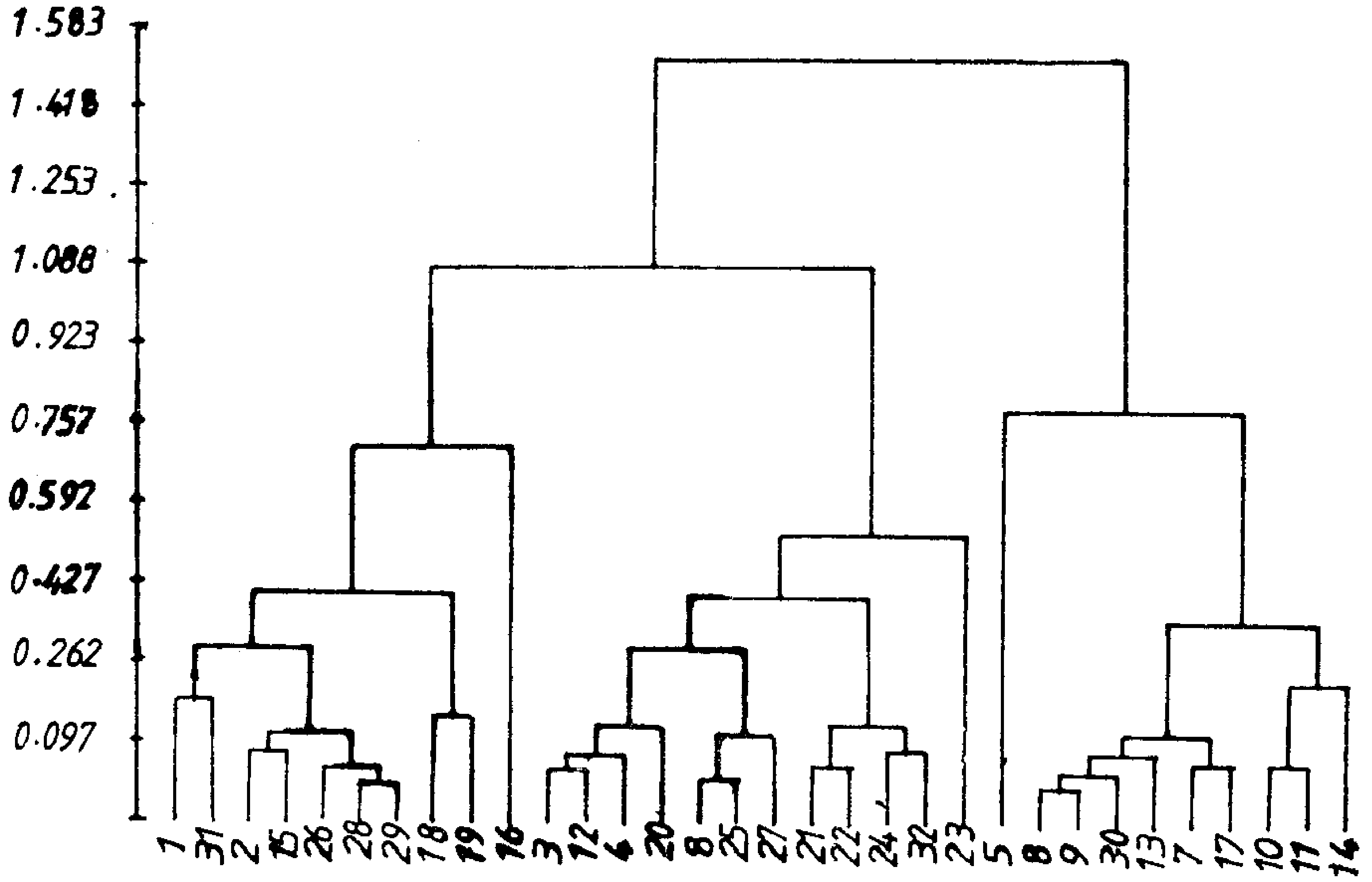
Classification by · centroid method

FIG:5



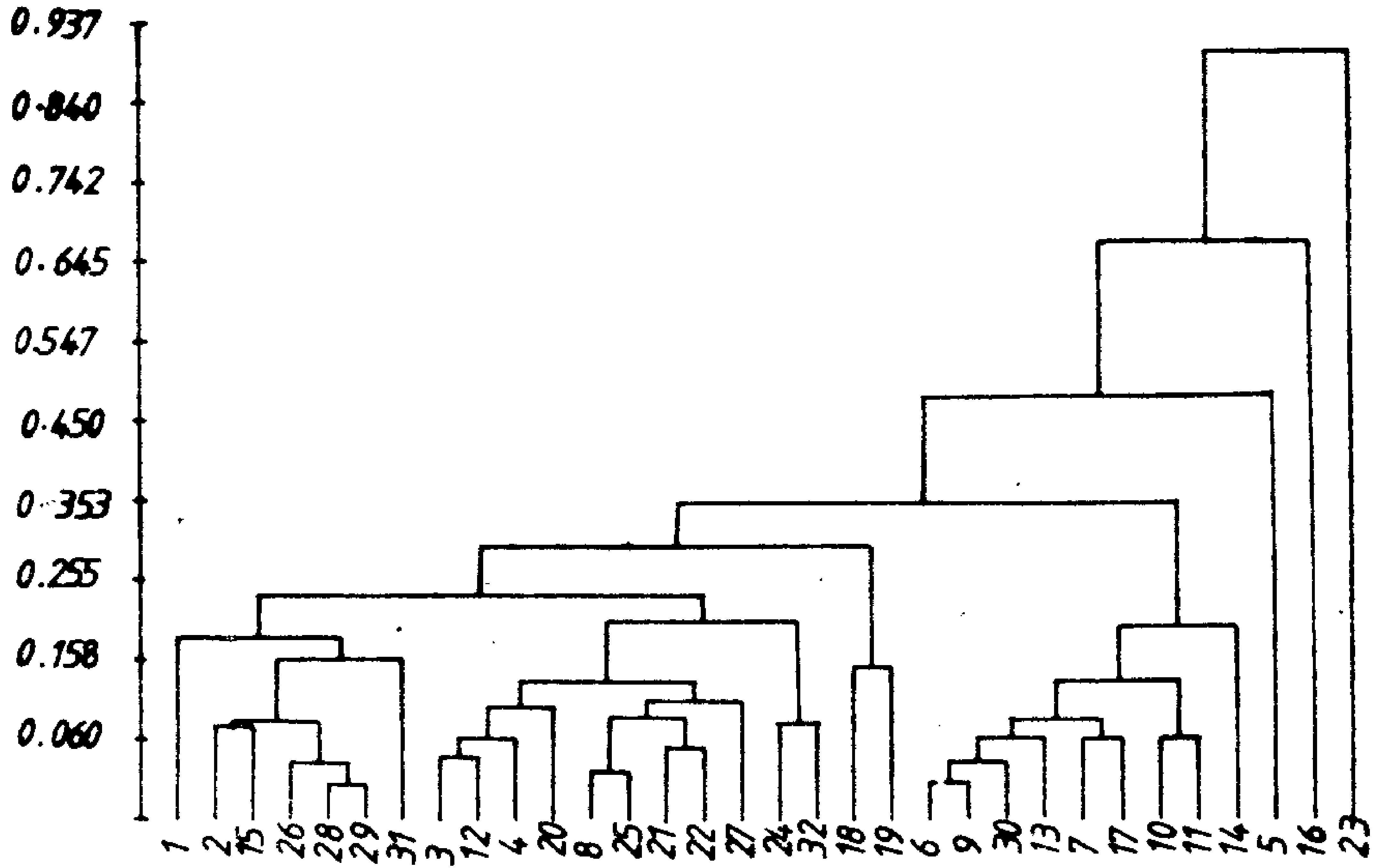
Classification by median sort

FIG: 6



Classification by wards method

FIG: 7



Classification by Lance Williams flexible sort (α = -0.25)