SIMILARITY AND TYPICALITY OF TAXA*

E. V. KRISHNAMURTHY

Indian Institute of Science, Bangalore

Received on July 15, 1975

ABSTRACT

Discovering a structure among a set of items of information (or subjects) is of primary importance in every branch of science. This survey article outlines the principles and practice of numerical taxonomy, the science of grouping of items of information on the basis of their attributes. The role of digital computer in establishing the resemblance or similarity among the items from a quantitative description of their attributes, as well as identifying typical members is described. Important application areas are mentioned and a relevant bibliography is provided.

Keywords: Numerical Taxonomy; Description space; Nominal-ordinal-interval-ratio scales; Taxon; Similarity; Resemblance; Correlation; Distance measure; Typicality; *k*-means; Graphs; Clusters and Factor analysis; Phylogenetic classification; Nosology; Physiognomy.

INTRODUCTION

One of the most primitive and common activities of man consists of sorting (like) things into categories or trying to discover a structure among the items of information presented to him. The persons, objects, events or other items of information encountered over a small period of time are too numerous for mental processing as unique entities. Therefore each stimulus is described primarily in terms of category membership. Ideally one strives to find the minimum number of choices that would identify the given item unambiguously and uniquely. Accordingly, in any situation one tries atmost to enumerate a set of possible characteristics of objects or items of which any adequate subselection constitute the specified object. In practice, however, some of these characteristics may only partially participate and accordingly while trying to assign the category membership the following three situations arise:

1. The category membership is very well defined by a set of attributes.

---

*Talk delivered at the Bandipur meeting on 'Mathematical Models in Genetics and Ecology' *held between June 17 and 21, 1975* by Centre for Theoretical Studies, Indian Institute of Science, Bangalore.

2. The category structure is not well known, it varies from nearly complete description of categories to knowing merely the number of categories.

3. Little or nothing is known about the category structure. All that is available is a collection of observations whose category memberships are unknown.

When the category membership is well-defined, we call the process of allocation or assignment of additional unidentified objects to the correct class, as identification (some people call this as classification; this is not a correct usage).

When the category structure is not too well-known the problem is one of discrimination.

When the category structure is completely unknown the operational objective is to discover a category structure which fits the observation. The problem is then to find the 'natural groups' in such a way that the degree of 'natural association' is high among the members of the same group and low among members of different groups. Most of the problems we face in genetics, ecology, agriculture, biology and other sciences belong to situation 3. (We call this as classification into groups.) (Of course complications arise even in the case of situations 1 and 2 due to imperfect class definitions, overlapping categories and random variations in observations). The essence of Cluster analysis, Factor analysis, Principal component analysis and other topics of Multivariate statistical analysis is to assign meanings to natural groups and natural association. In our discussion we will only be concerned with situation 3.

In numerical taxonomy each item of information is called a taxonmic unit and the categories as taxa. Numerical taxonomy† is the science of grouping of taxonomic units on the basis of their attributes. These methods require the conversion of information about taxonomic entities into numerical quantities and the application of numerical/statistical computational techniques for grouping them.

In practice, the analysis is carried out in the following sequence :

Step 1.—Organisms and characters (called data units) are chosen and recorded.

---

† This term was coined by biologists. In information retrieval, it is called 'clumping'. In geography it is called 'regionalization'. Anthropologists call it 'seriation'. Botanists/Ecologists call it as 'typology'. However, in all cases, the methodology remains essentially the same.

Step 2.—The resemblances between every pair of organisms are evaluated by using appropriately defined similarity (by using angular measures) or dissimilarity (by using distance measures) coefficients.

Step 3.—'Natural groups' based on these resemblances are formed by using clustering algorithms.

All these three steps demand enormous attention and with each step is associated a number of problems.

I. Choice of Data Units

While choosing the data units two different situations arise:

1. The sample is the complete object of analysis. The purpose is to discover a classification scheme for the given set of data units. It is not intended that the results should be applied to any additional data units outside the sample. In such a case the principal consideration is to make sure that no important data units are omitted.

2. The sample is a portion of a much larger population which is the true object of interest. We can then apply the principles of radom * and independent selection.

The data units must now be consistently described in terms of their characteristics, attributes, class membership, etc. Collectively these descriptors are the variables of the problem.

These characteristics** may be morphological, physiological, ethnological distributional, etc. One should guard against introducing bias into the choice of characteristics. For instance

1. Meaningless characters should be eliminated; which are not really attributes, *e.g.*, the number of leaves in a tree.

---

* Randomization means all data units are equally likely as far as selection of a sample is concerned (unbiassed). Under random selection any groups that exist in the data will tend to be represented in the sample in proportion to their relative size in population. The size of the sample must be chosen so that small or rare groups are not lost. Independence means the choice of each data unit is not influenced by the choice of any other. But if selection of some data units promotes the candidacy of others, the effect should be exploited for the evidence of association rather than neutralized in deference to independence.

** Usually the number of characteristics that are chosen are around 60 so as to be easily handled by a computer.

2. Logically correlated characters should be eliminated. Any redundant property that is a logical consequence should be avoided.

3. Partially logically correlated characters should be carefully tackled (see next section).

4. Invariant characteristics are to be eliminated.

5. Empirically correlated characters should be eliminated.

II. Description of Obects, Description Space and Representation of Data

Thus we conceptually visualize every object by making representation or symbolism (*e.g.*, the name is a symbol). By a representation it is meant any structure whether abstract or concrete of which the features purport to symbolize or correspond in some sense with the given set of objects.

Using these we try to study the resemblance or distinguish the objects. An object can then have associated with it a descriptive statement which locates it as a point in the n-dimensional description space Ω. All the dimensions that we can distinguish present in Ω and discriminations along any one dimension are assumed to be as fine as can be made. Normally these discriminations along any one dimension are called scores. While forming these scores we usually use different scales of measurement.

1. A *nominal scale* merely distinguishes between objects or classes. That is, with respect to A and B one can only say

$$x_A = x_B \quad \text{or} \quad x_A \neq x_B$$

e.g., $A = \text{crow}, \quad B = \text{coal}, \quad C = \text{Rose}$

$x = \text{Black (Property)}$

$x_A = x_B; \quad x_A \neq x_C$

2. An *ordinal scale* induces an ordering of the objects. In addition to distinguishing between $x_A = x_B$ and $x_A \neq x_B$ the case of inequality is further refined to distinghish

$$x_A > x_B \quad \text{or} \quad x_A < x_C$$

namely the comparative degree.

e.g., Coal is darker than crow.

3. An *interval scale* assigns a meaningul measure of the difference between the two objects.

One may say not only $x_A > x_B$ but also $(x_A - x_B)$ units of difference. *e.g.*, density or grey levels.

4. A *ratio scale* is an interval scale with a meaninful zero point.

If $x_A > x_B$ then one may say that A is x_A/x_B times superior to *B*. *e.g.*, Specific gravity.

These scale definitions are ordered hierarchically from nominal up to ratio scale. Each scale embodies all the properties of all the scales below in ordering. Therefore, by giving up information one may reduce a scale to any lower order scale. Frequently variables on nominal and ordinal scales are referred to as (categorical) qualitative variables often with ambiguity as to whether any order relation exists. For contrast, variables on interval or ratio scales are referred to as quantitative variables.

The n quantitative characteristics or attributes of t specimens are tabulated as an $n \times t$ data matrix (called score matrix) thus*

Characters) Attributes	Operational Taxonomic Unit (OTU) Specimens 1 2 · · · · t
1	
2	
·	Score X_{ij}
·	
·	
n	

Here the t columns represent the t individuals to be grouped on the basis of resemblances and whose n rows are the n unit characters. Each X_{ij} (0 or 1 or multivalued) is the score of the individual j for character i.

The standard score (zero mean and unit variance) Z_{ij} is defined thus

$$Z_{ij} = \frac{X_{ij} - \bar{X}_i}{S_i} \; ; \; \bar{X}_i = \sum_{j=1}^{t} X_{ij}/t = \text{univariate mean of } i\text{th character}$$

* By the use of principal components method (and extra computation) it is possible to construct a set of fewer than n composite variables which are linear combinations of the original variables and which account for the variance of the original data. But another way, the axes representing the original variables may be rotated individually to be orthogonal with each other and in the process it may be found that fewer than n orthogonal axes will span the space. The principal components method helps to define such an orthogonal set with maximum variance properties. (*See* Bibliography). To find the principal components we have to find the eigenvalues and eigenvectors of the sample variance-Covariance matrix.

and

$$S_i = \sqrt{\frac{\sum_{j=1}^{t} (X_{ij} - \overline{X}_i)^2}{t}} = \text{univariate standard deviation.}$$

III. Resemblance or similarity between OTU and typicality

Estimation of resemblance† or similarity is the most important and fundamental step in numerical taxonomy.

For visualizing the idea of degree of similarity consider the illustration in Fig. 1. Imagine the smaller form inflated by an internal pressure

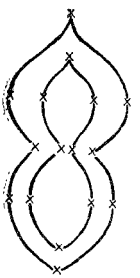


FIG. 1

(growth, in effect) so that the difference in volume between the two forms is reduced. The growth pattern may be of any kind, and the two forms become identical when the set of marked points coincide.

Formally, for describing this, we can set up an n-dimensional set of Cartesian coordinates in which the axes are the directions of displacement of the marked points. Each of the two forms can then be located within the framework of reference by noting the amount of the displacementt of each marked point along its axis of variation. In order to measure their

† The association of pairs of characters (rows) can be examined over all OTU (columns). This is called R-technique. The converse, namely, the association of pairs of OTU's (columns) over all characters (rows) is called the Q-technique. Main emphasis in numerical taxonomy is the Q-technique. The main mathematical steps are formally the same and an R-study can be made by transposing the data matrix so that the characters (rows) become the individuals comparable to the former OTU's and the actual OTU's or taxa (columns) become the attributes over which the association is computed.

similarity one can use the Euclidean distance between the two forms by using the n-dimensional Pythagoras equation

$$d^2 = \sum_{1}^{n} X_i X_j$$

This distance (measured in this Euclidean space) will be meaningul only if the movement of any one of the marked point is *independent* of the movement of other points there being no correlation between the measured characters. In general, substantial correlations exist and this bias our measure of distance.

To eliminate the effect of correlated characters, we set the angles between the axes of our chart so that the cosine of the angle between any two axes equals the coefficient of correlation between the characters whose displacement they represent. To act in this way would be very cumbersome. Therefore, we use a method by which the distances can be computed whilst taking into account the correlations between characters. This method consists in inserting into the calculation a metric tensor (the fundamental tensor descriptive of a space) in which the correlations between the characters are removed by distorting the space to a calculated extent. We now have

$$D^2 = \Sigma\, g^{ij} X_i X_j$$

where D^2 is the generalized distance‡ between the two forms, adjusted for any correlation that may exist between the measured characters and g_{ij} is the metric tensor which represents the inverse of the dispersion matrix (covariance or variance-covariance matrix). $X_i X_j$ is the vector of differences between the characters on smaller form and those on the larger one.

If $g^{ii} = 1$, $g^{ij} = \theta\ (i \neq j)$, then we go to the Pythagorean equation. In fact g^{ij} describes the extent to which the Riemannian hyperspace has to be distorted to accommodate the interrelationship existing between the characters when measured in Euclidean space.

The generalized distance has the character of a geodesic, the line of shortest path between the two forms in a curved space; it reduces to a straight line in Euclidean space.

---

‡ Mahalanobis is the originator of this very important and fundamental concept; hence this distance measure is known as Mahalanobis distance. This is widely used after the advent of high speed computers.

Most of the measures for resemblance are either based on distance or angular measures. The distance measure s, in particular, satisfy the metric properties.*

The following list gives some of the important distance, measures used. (Note that all are Q-techniques)

Euclidean : $\triangle_{jk} = \left[\sum_{i=1}^{n} (X_{ij} - X_{ik})^2\right]^{\frac{1}{2}}$

Minkowski : $d_r(j, k) = \left[\sum_{i=1}^{n} | X_{ij} - X_{ik} |^r\right]^{1/r}$

Manhattan or city block metric:

$$d_1(j, k) = \left[\sum_{i=1}^{n} | X_{ij} - X_{ik} |\right]$$

Canberra metric : $d_{can}(j, k) = \sum_{i=1}^{n} \frac{| X_{ij} - X_{ik} |}{X_{ij} + X_{ik}}$

Mahalanobis metric† $D^2_{jk} = \delta^T_{jk} S^{-2} \delta_{jk}$

where S^{-1} is the inverse of variance-covariance matrix and δ_{ik} = vector difference between means of samples j and k for all characters.

Among the angular measures, the Pearson poduct-moment correlation coefficient is the one most widely used. This coefficient computed between OTU j and k is

$$r_{jk} = \frac{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)}{\sqrt{\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{n} (X_{ik} - \bar{X}_k)^2}}$$

---

* $\phi(a, b) \geqslant 0$; $\phi(a, a) = 0$
$\phi(a, b) = \phi(b, a)$ (symmetry)
$\phi(a, c) \leqslant \phi(a, b) + \phi(b, c)$ (triangle inequality)
If, however, $\phi(a, c) \leqslant$ max. $[\phi(a, b), \phi(b, c)]$, we call it ultrametric.

† $S = \frac{1}{n} XX^T = \frac{1}{n}\left[\sum_{i=1}^{n} (X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k)\right] = \{S_{ik}\}$

where

$$X = \{X_{ij} - \bar{X}_j\}, \qquad \bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} = \text{mean of all state values of OTU } j$$

The S-matrix is sums of square and cross products of deviation scores matrix divided by the number of objects. It is also called the dispersion matrix.

where

$\bar{X}_{ij}$ = character state value of character i in OTU j

$\bar{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ = mean of all state values for $OTU\,j$

n = number of characters sampled.

Other measures for agreement are by the use of association coefficients; usually these are computed in practical problems with two state characters.

When character states are compared over pairs of columns in a conventional data matrix the outcome can be summarized in a conventional 2×2 frequency table such as

		$OTU\,j$ 1	$OTU\,j$ 0	
$OTU\,k$	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d=n$

In the left upper hand corner, we place the number of characters coded 1 in both *OTU*, while in the right hand lower corner, we write the number of characters coded 0 in both. The other quadrants register the number of characters in which the two *OTU*'s disagree being coded 1 for j and 0 for k (or converse)

n = sum of frequencies

$m = a + d$ (number of matches)

$u = b + c$ (number of mismatches)

$m + u = n$.

Then the following coefficients can be defined:

1. Jaccard coefficient : $a/(a+u) = a/(a+b+c)$
2. Simple matching : $m/(m+u)$
3. Yule coefficient : $(ad - bc)/(ad + bc)$.

IV. Taxonomic Structure (typicality and cluster)

Once the resemblance between any pair of taxonomic units is established we can form a $t \times t$ matrix R with elements S_{ij} denoting the similarity

coefficient. Since symmetric distance measures are usually used the matrix R is symmetric with $r_{ii} = 1$.

		OTU		
		1 2 ··· t		
	1	r_{11}	···	r_{1t}
	2	··	···	··
OTU	⋮	··	···	··
	t	r_{t1}		r_{tt}

The taxonomic structure is now to be detected using this matrix. For this we need to group or luster *OTU* hich have a high degree of association thereby partitioning the given set of t elements.

While clustering some important problems arise:

1. What is a cluster centre? How to define this?

It can be viewed in two different ways, *viz.*, as a point representing an actual organism or as a point representing hypothetical organism such as the average man who has 0·8 wife and 2·3 children).

The average organism or centroid $\bar{X}$ is given by the point in the description space whose coordinates are the mean values of each character over the given cluster of *OTU*'s. It is also the centre of gravity of the cluster.

For (0, 1) data ‡ another construct commonly used in microbiology is the hypothetical median organism. It is that organism which possesses the commonest state for each character (called typical).

The most usual measure of an actual *OTU* is the centrotype. It is the *OTU* with the highest mean resemblance to all other *OTU*'s of the cluster. It is the *OTU* nearest to the centroid (in Euclidean distance models; not necessarily in other models).

‡ For example, if the data matrix is

Characters	OTU						
	1	2	3	4	5	6	7
1	0	0	0	1	1	1	1
2	0	0	0	1	1	1	1
3	1	1	1	1	0	0	0
4	1	1	1	1	0	0	0

We say OTU 4 is the typical ' one.

Clustering and Classification

Ideally what we desire is to classify the objects and to take typical objects as representatives of the whole body of the objects. This helps to encode the original information as efficiently as possible. For example, when thinking of the human population, we may divide it into nations and bear only typical representatives of them in our mind.

Rank Correlations and Clustering

The formation of resemblance matrix considering $\frac{1}{2}t(t-1)$ pairs of individuals over the n-variables is a large computational problem (psychologists call this as Q-technqiue). In addition, the different variables may be measured in different units and correlations of a pair of individuals over n-values of non-comparable units, do not, in general make sense. This difficulty is not overcome even by standardizing (Reducing to zero mean and unit variance). For this purpose it is better to use rank correlation procedures.

Here corresponding to each property the objects are ranked. The variance of a set of t-ranks in $\frac{1}{12}(t^2-1)$; when ties are present this result needs modification, *viz.*,

$$\text{Var}(X) = \frac{1}{12t}[(t^3 - t) - (a^3 - a)]$$

where summation extends over all ties of extent a.

For each pair j, k, we calculate

$$S = \sum_{i=1}^{n} \frac{(X_j - X_k)^2}{\text{Var } X_j}$$

where X_j and X_k are the values of the ranks for each i, for the pair in question. Note S is also a kind of distance. Given all these distances the clustering is done as follows:

Pick the pair which are closest. Then add that member which increases their average distance the least; then add a fourth member which increases the mean distance the least; and so on until a point is reached at which the addition of a new member adds *too* much to the mean distance. The amount which is to be considered '*too much*' is an arbitrary figure If this procedure does not exhaust the set, proceed to the nearest unused pairs and repeat the procedure.

Hierarchical Clustering

There are several other procedures in which this measure of nearness is defined suitably. These can be roughly classified under two types.

Type 1.—Methods which start with an assumed number r of clusters and modify the value of r as the clustering algorith proceeds.

Type 2.—Methods which try to find out a fixed number r of clusters iteratively.

The former class is known as hierarchical and is divided into :

(i) Agglomerative hierarchical clustering where r decreases as the procedure continues.

(ii) Divisive hierarchical clustering where r increases as the procedure continues.

Usually type 2 methods are more widely used. For this purpose some criterion function is to be minimized.

Criterion Function

Suppose we have a set x of t samples $x_1 \cdots x_t$ and it is desired to find the disjoint clusters $X_1 \cdots X_r$ in such a way as to minimize a criterion function. Let X_k have m_k samples so that

$$t = \sum_{k=1}^{r} m_k$$

(i) Mean value of kth cluster

$$\mu_k = \frac{1}{m_k} \sum_{x \in X_k} x$$

(ii) Dispersion matrix of kth cluster

$$S_k = \frac{1}{m_k} \sum_{x \in X_k} (x - \mu_k)(x - \mu_k)^T$$

(iii) Mean vector of the entire data

$$\mu = \frac{1}{t} \sum_{x \in X_k} x$$

(iv) Scatter matrix for X

$$S_T = \frac{1}{t} \sum_{x \in X_k} (x - \mu)(x - \mu)^T$$

(v) Within cluster scatter matrix

$$S_W = \frac{1}{t} \sum_{k=1}^{r} m_k S_k$$

(vi) Between cluster scatter matrix

$$S_B = \frac{1}{t} \sum_{k=1}^{r} m_k (\mu_k - \mu)(\mu_k - \mu)^T$$

S_T, S_W and S_B obey the identity

$$S_T = S_W + S_B$$

(vii) Minimum variance criterion with Euclidean distance measure

$$D = \sum_{k=1}^{r} \sum_{x \in X_k} \| X - \mu_k \|^2$$

This is a measure of the deviation of the sample in cluster X_k from its centre μ_k. D attains minimum when μ_k is the centre of X_k. This criterion is most suited to data sets with widely separated, compact, ellipsoidal clusters.

For multivariate normal distribution, it is preferable to use the Mahalanobis distance measure with a covariance matrix corresponding to within cluster variance, *viz.*,

$$d(x, y) = (x - y)^T S_W^{-1} (x - y)$$

where S_W = within cluster scattering matrix.

In such a case, each cluster determines its own metric, *viz.*, the Mahalanobis distance for the cluster. However, in using this distance we pay a higher cost for computation. Using this concept, a procedure known as k-means method hasebeen developed.

k-means procedure

The k-means procedure consists of simply starting with k-groups each of which consists of a single random point and thereafter adding each new

point to the group whose mean, the new point is nearest. After a point is added to a group the mean of that group is adjusted in order to take account of the new point. Thus at each stage the k-means are in fact the means of the groups they represent.

There is another important algorithm due to Ward for hierarchical clustering.

Step 1.—Here we begin with t groups (t = number of individuals) each consisting of one observation. At this stage $D = 0$.

Step 2.—At each stage reduce the number of groups by one through merger of those groups whose combination gives the least possible increase in D.

Step 3.—Continue for a total of $(t - 1)$ merges until there is one group.

This technique tends to give minimum D partitions for each number of groups from t to 1.

Graph theory also plays a very important role in developing algorithms for clustering (see bibliography).

V. Concluding Remarks

Numerical taxonomy has wide applications in various fields ranging from biology to earth sciences. The development of high speed computers with large memory has made it possible to realize many of the algorithms for finding the similarity and typicality of taxa with ease. The following are some of the typical application areas in the context of our country's needs :

1. Drug design based on chemo taxonomy; classification of medicinal plants in terms of their chemism.
2. Microbiology
3. Protein taxonomy—Phylogenetic tree construction.
4. Nosology—Classification of diseases from Symptoms.
5. Forensic Science—Physiognomy—Human face recognition.

BIBLIOGRAPHY

Numerical Taxonomy

1. SNEATH, P. H. A. AND SOKAL, R. R., *Numerical Taxonomy*, W. H. Freeman and Company, San Fransisco, 1973.
2. CROWSON, R. A., *Classification and Biology*, Heinemann Educational Books Ltd., London, 1969.
3. COLE, A. J. (Ed), *Numerical Taxonomy*, Proc. Colloq. Num. Tax. held at Univ. St. Andrews, June 1968, Academic Press, London, 1969.
4. HAWKES, J. G. (Ed.), *Chemotaxonomy and Serotaxonomy*, Proc. Symp. Botany Dept, Birmingham Univ., 1967, Academic Press, London, 1968.
5. LOCKHART, W. R. AND LISTON, J. (Ed.), *Methods for Numerical Taxonomy*, American Society for Microbiology, 1970.
6. SESHACHAR, B. R. (Ed), *Symposium on Newer Ttrends in Taxonomy*, National Institute of Sciences, New Delhi, 1967.

Topics related to classification

1. MACKAY, D. M., *Information, Mechanism and Meaning*, M. I. T. Press, Cambridge, Mass., 1969.
2. WITTGENSTEIN, LUDWIG, *Tractatus Logico, Philosophicus* (Translation from German text by Pears, D. F. and McGuinnen, B. F.), Routledge and Kegan Paul, New York, 1963.
3. HUNT, E. B., *Concept Learning, An Information Processing Problem*, John Wiley, New York, 1962.
4. SALTON, G., *Automatic Information Organization and Retrieval*, McGraw- Hill Book Co., New York, 1968.

Cluster Analysis/Factor Analysis

1. ANDERBERG, M. R., *Cluster Analysis for Applications*, Academic Press, New York, 1973.
2. HORST, PAUL, *Factor Analysis of Data Matrices*, Holt, Rinehart and Winston, New York, 1965.
3. HARMAN, H. H., *Modern Factor Analysis*, The University of Chicago Press, 1960.
4. LAWLEY, D. N. AND MAXWELL, A. E., *Factor Analysis as a Statistical Method*, Butterworths, London, 1963.
5. FRIEDMAN, R. AND RUBIN, J., On some invariant criteria for grouping data, *J. Amer. Stat. Assoc.*, 62, Dec. 1967.
6. FUKUNAGA, K. AND KOONTZ, W. L. G., A criterion and an algorithm for grouping data, *IEEE Trans. Computers*, C-19, pp. 917-23, October 1970.

7. GITMAN, I. AND LEVINE, M. O., An algorithm for detecting unimodal fuzzy sets and its application as a clustering technique, *IEEE Trans. Comp.*, July 1970. C-19, 583–93,

8. HARALICK, R. M. AND DINSTEIN, I., An iterative clustering procedure, *IEEE Trans. Sys. Man and Cyb.*, 1971, SMC-1, 275–89.

9. JARVIS, R. A. AND PATRICK, E. A., Clustering using a similarity measure based on shared nearest neighbours, *IEEE Trans. Comp.*, 1973, C-22, 1025–34.

10. JENSEN, R. E., A dynamic programming algorithm for cluster analysis, *Oper. Research*, 1970, **18**, 1034–57

11. KENDALL, M. G., *Cluster Analysis*, in *Frontiers of Pattern Recognition*, S. Watanbe (Ed.), Academic Press, New York, 1972.

12. MUCCIARDI, A. N. AND GHOSE, E. E., An automatic clustering algorithm and its properties in higher dimensional spaces, *IEEE Trans. Sys. Man and Cyb.*, April 1972. SMC–2, 247–54.

13. RUSPINI, E. H., A new approach to clustering, *Information and Control*, 1969, **15**, 22-32.

14. HUBERT, L., Min. and Max. hierarchical clustering using asymetric similarity measures, *Psychometrika*, 1973, **38**, 63–72.

15. HUBERT, L., Some extensions of Johnson's hierarchical clustering algorithms, *Psychometrika*, 1972, **37**, 261–274.

16. HUBERT, L. J. AND BAKER, F. B., Hierarchical clustering and the concept of power, *J. Amer. Stas. Assoc.*, (1975) to appear.

17. CHERNOFF, H., Metric considerations in cluster analysis, *Proc. Sixth Berkeley Symp. on Mathem. Statistics and Probability*, Vol. 1, Ed. L. M. Le Cam., J. Neyman, and E. L. Scott, University California Press, Berkeley, 1972.

18. MACQUEEN, J., Some methods for classification and analysis of Multivariate observations, *Proc. Fifth Berkeley Symp. on Mathem. Stat. and Probability*, Vol. 1, Ed. L. M. Le Cam. and J. Neyman, Univ. California Press, Berkeley, 1967.

19. BRENNAN, R. L., Measuring agreement when two observers classify people into categories not defined in advance, *Br. J. Math. Psych.* 1974, **27** 154–163.

20. MAXWELL, A. E., Tests of association in terms of matrix algebra, *Br. J. Math. Psych.*, 1973, **26**, 155–166.

Graph Theory and Clustering

1. ZAHN, C. J., Graph theoretical methods for detecting and describing gestalt clusters, *IEEE Trans. Comp.*, 1971, C–20, 68–86.

2. HUBERT, L. J., Some applications of graph theory to clustering, *Psychometrika*, 1974, **39**, 283–309.

3. HUBERT, L. J., S me applications of graph theory and related non-metric techniques to problems of approximate deviation, *Brit J. Math. Statis. Psychology*, 1974, **27**, Part 2, 133–153.

4. HUBERT, L., Spanning trees and aspects of clustering, *Brit. J, Math. Statis. Psychology*, 1974, **28**, 14–28.

5. HUBERT, L., Approximate evaluation techniques for the single link and complete link hierarchical clustering procedures, *J. Am. Stat. Assoc.*, 1974, **69**, 698–704.

6. HUBERT, L. J. AND SCHULTZ, J. V., Data analysis and the connectivity of random graphs, *J. Math. Psych.*, 1973, **10**, 421–428.

7. HUBERT, L., Problems of seriation usingia subject by item response matrix, *Psychol. Bull.*, 1974, **81**, 976–983.

8. HUBERT, L., Monotone invariant clustering procedures, *Psychometrica*, 1973, **38**, 47–62.

Phylogenetic Trees—Estimation of Parameters

1. DAYHOFF, MARGARET OAKLEY, Computer Analysis of Protein Evolution, *Scientific American*, 1969, **221**, 87–95.

2. FLORKIN, M., *A Molecular Approach to Phylogeny*, Elsevier, Amsterdam, 1966.

3. FITCH, W. M. AND MARGOLIASH, E., Construction of Phylogenetic trees, *Science*, 1967, **155**, 279–284.

4. KASHYAP, R. L. AND SUBAS, S., Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process, *J. Theor. Biol.*, 1974, **47**, 75–101.

Multivariate Analysis

1. MORRISON, D. F., *Multivariate Statistical Methods*, McGraw-Hill, New York, 1967.

2. ROY, S. N., *Multivariate Statistical Analysis for Biologists*, Methuen and Co. Ltd., London, 1969.

3. KENDALL, M. G., *Discrimination and Classification in Multivariate Analysis*, (Ed. P. R. Krishnaiah), Academic Press, New York, 1966, pp. 165–186.

4. BLACKITH, R. E. AND REYMENT, R. A., *Multivariate Morphometrics*, Academic Press, New York, 1971.

5. KENDALL, M. G., *Rank Correlation Methods*, Charles Griffin, London, 1958.

6. COOLEY, W. W. AND LOHNES, P. R., *Multivariate Data Analysis*, John Wiley, New York, 1971.

Chemical Taxa

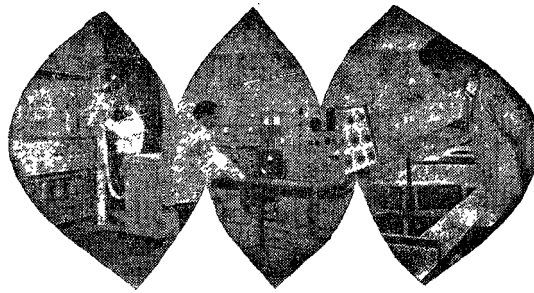
1. TETENYI, PETER, Infraspecific chemical taxa of medicinal plants, *Akademiai Kiado, Budapest*, 1970.

Calendar of events: Conferences/Symposia at the Indian Institute of Science Campus

Sl. No.	Name of the School	Period	Sponsoring Department of the Institute
1.	Material Science Sympoisum on 'Phase Tramsformations and Phase Equilibria'	October 1975	B.A.R.C., Bombay
2.	Intensive Course on Fluid Engineering	20 October to 2 November 1975	School of Automation
3.	Lecture Course on Cavitation	November to December 1975	Chemical Engineering
4.	Crystal Chemistry for College Teachers	December 1975	Inorganic and Physical Chemistry

On the basis of the information received by the Editorial Office on 15th October 1975.

In electronics our product list is long, our credentials strong...



And our market: world-wide

Recently we have received a sizeable order for radar systems from the renowned firm CONTRAVES of Switzerland. They are one of the foremost manufacturers of radars themselves !

Then why place the order on us ?

For two very practical reasons. They know that we have dependable production capability and technical expertise to deliver the goods —on schedule and custom-built to their designs and specifications. Secondly, the costs would be quite attractive.

Reasons sound enough for any business organisation.

It is not by accident that our exports touched Rs. 1.9 crores during 1973-74. Some of the orders came from countries highly advanced in the field of electronics such as U.K., West Germany, U.S.A., Canada, Australia, Japan, Hong Kong and Singapore.

You too could profitably utilise our capacity and technical skills. Spell out the equipment or system and we will make it for you.

Our Bangalore Complex, comprising six manufacturing Divisions, occupies a land area of 70 hectares and employs over 13,000 workers. Our Ghaziabad Unit near New Delhi is laid out over an area of more than 84 hectares and employs 1200 people.

Each Division is equipped with modern machinery and facilities. The sophistication and complexity of the machinery are matched by the skills and dexterity of our workers.

PRODUCT RANGE : A whole range of communication equipment from transreceivers to high power transmitters, radio and video broadcast transmitters and studio equipment, UHF radio relay systems, radar systems for surveillance, weapon control and meteorology, components like transistors, Integrated Circuits, receiving and transmitting tubes, X-ray and TV picture tubes occupy the present production spectrum.

PRODUCTION AND PROCESS TECHNIQUES AND CONTROL : Up to date techniques are employed using the latest machinery inspection equipment and the best available materials. A high level of standardisation has been achieved in all the activities right from the component choice to final assembly.

RELIABILITY. We are fully equipped to design, produce and inspect equipment to rigid international standards like DEF, BSS, MIL, JSS etc.

A series of special tests and checks including extensive environmental tests ensure high standards of quality.

DEVELOPMENT AND ENGINEERING. The highly qualified team of over 200 engineers in our Development and Engineering Division have to their credit the designing of several modern equipment and systems.

In fact our systems capability has reached such an advanced stage that we are setting up a multichannel UHF radio relay link over a 1200 km pipeline for the Indian Oil Corporation and Microwave Communication system for the Tamilnadu police.

Your enquiries will be processed by our team of dedicated engineers who will endeavour to give you satisfaction in every way.

BHARAT ELECTRONICS LTD.

BANGALORE ● GHAZIABAD

(A Govt of India Enterprise)