Indian Minerals, Vol. 43, No. 2 (April-June, 1989), pp. 128-135

CLUSTER ANALYSIS REVISITED: A CASE STUDY FROM BIHAR MICA BELT GRANITES, EASTERN INDIA

S. S. SARKAR

Geology Department, Presidency College, Calcutta-700 073

A. CHATTERJEE

Department of Mathematics, Burdwan University, Burdwan

AND

S. NANDY

Computer Science Unit, Indian Statistical Institute, Calcutta

ABSTRACT

In the present paper a review on various aspects of cluster analysis has been made. The variables used are five mineralogical attributes [viz., quartz (vol. %), colour index, anorthite content of plagioclase, N_Z of biotite and feldspar ratio] of a large number of samples selected from eleven intrusive granite plutons (each pluton was treated as a group) of Bihar Mica Belt, eastern India. It has been demonstrated using the data of Bihar Mica Belt granites that the pattern of dendrograms depends to a large extent on (*i*) the 'similarity measure' and (*ii*) the methods of clustering, used in the problem concerned and thus different dendrograms may indicate in some cases mutually contradictory inferences, not in tally with the geological observations on the same granite groups. Therefore, for deriving a comprehensive and convincing inference, different clustering methods employing different distance measurements are to be attempted. However for a cursory investigation average linkage clustering method on the basis of Mahalanobis D^2 statistic is recommended.

INTRODUCTION

In any classification scheme for geological data, the choice of a few discriminating variables that retain major information regarding different groups under study is a great problem and is often not free from subjective bias. To get rid of this problem, some sort of summarising method is required which takes into account all the variables under consideration at a time and in doing so should be capable of rendering a magnified view of the differences (however small) amongst the groups. Due to its over-simplicity and capability of visual display of the ultimate inference, cluster analysis technique has found immense application in geological problems.

The present work reviews some of the important aspects of cluster analysis technique with special reference to categorising the intrusive granites of Bihar Mica Belt, eastern India (henceforth called as BMB granites). A detailed account on the petrochemistry and evolution of these granites is given in Saha *et al* (1987). A total of 395 samples from eleven BMB granite plutons are chosen for the present study; five mineralogical attributes, viz., quartz (vol. %), colour index, feldspar ratio [i.e. K-feldspar$\times$100/(K-feldspar+plagioclase)], anorthite content

of plagioclase and N_Z of biotite are used as variables, whose means and standard deviations (for individual plutons) are depicted in Table 1.

TABLE 1

MEAN AND STANDARD DEVIATION OF THE MINERALOGICAL ATTRIBUTES FOR DIFFERENT BIHAR MICA BELT GRANITE PLUTONS

Group No.	Name of pluton	N		Quartz (Vol. %)	Colour Index	R. I. of Biotite (N_Z)	An % of plagioclase (mole. %)	Feldspar ratio
1	Tisri	91	x	34.33	7.46	1.643	15.46	65.25
			s.d.	4.46	2.83	0.009	6.43	5.14
2	Chauki	32	x	32.76	9.39	1.655	20.75	66.09
			s.d.	5.05	2.42	0.013	6.13	5.50
3	Manihari	48	x	34.38	8.10	1.652	16.94	65.67
			s.d	4.28	2.55	0.013	6.02	4.12
4	Banresar	60	x	30.82	8.29	1.656	18.98	64.72
			s.d.	4.83	2.25	0.012	4.50	4.72
5	Gawan	25	x	32.71	11.15	1.663	21.44	59.15
			s.d.	4.40	3.01	0.009	6.96	9.68
6	Barmi	25	x	31.14	11.98	1.645	14.68	71.55
			s.d.	5.12	3.12	0.008	3.04	3.42
7	Debaur	21	x	28.03	10.43	1.648	20.76	70.55
			s.d.	4.26	2.93	0.011	5.58	8.82
8	Khobarwa	23	x	30.11	6.58	1.648	16.30	55.33
			s.d.	5.49	3.31	0.011	6.03	8.30
9	Bandapahar	31	x	31.97	10.07	1.656	21.97	77.89
			s.d.	2.27	4.10	0.010	8.35	5.79
10	Kalapahar	23	x	32.56	6.13	1.661	26.48	57.82
			s.d.	9.74	2.73	0.010	4.68	12.13
11	Simratari	16	x	34.90	8.84	1.644	15.25	63.66
			s.d.	5.97	5.06	0.009	3.17	13.25
	All groups	395	x	32.53	8.63	1.651	18.41	65.50
			s.d.	5.03	2.98	0.011	5.89	6.83

N=No. of samples; x=Mean; s.d.=Standard deviation.
Data source: Group 1—Mitra (1984); 2, 3 & 4—Sarkar, S. S. (Unpublished data); 5 & 10—Mukhopadhyay (1974); 6, 7, 8 & 11—Mukhopadhyay (1981); 9—Ray (1985).

METHODOLOGY

A sequence of classification in which larger clusters are obtained through merger of smaller ones is a 'nested' or 'hierarchical' classification. Two basic prerequisites for such classification schemes are: (*i*) Choice of *similarity measures* to characterise the relationships among groups to be clustered and (*ii*) the *method of linkage* to be used.

SIMILARITY MEASURES

In order to cluster the variables some numerical similarity measures for characterising the relationships among variables are required. The conventional approach to this requirement is to compute a measure of association for every pairwise combinations of the variables.

A basic assumption of all cluster analysis methods is that these numerical measures of association are all comparable to each other (although each measure reflects association in only a particular sense and thus in a particular case it is needed to choose a measure, appropriate to the problem concerned and its context).

For a given data set of n-variables and m-data units, a common device for displaying the measured values is the data matrix of n-rows and m-columns. The *i*th row of the matrix contains all scores pertaining to *i*th variable and the *j*th column contains all scores for *j*th data unit. Within this setting a *row vector* of scores is the collective response of all data units to a single variable and consequently all scores are comparable to each other. On the other hand, a *column vector* of scores for a single data unit cuts across all the variables (rendering it more versatile with respect to the former, from the stand point of inductive generalisation). There may be quite a variety of measurement units and variable types. This heterogeneity makes it especially difficult to define meaningful measures of association between data units within the context of a given set of variables. The usual way of manoeuvring this difficulty is to introduce the concept of distance and the familiar euclidean distance is given by

$$D_2 (x_j , x_k) = [\sum_{i=1}^{n} (x_{ij} - x_{ik})^2]^{1/2}$$

where, x_{ij} be the score achieved by the *j*th data unit on *i*th variable and the vectors for *j*th data unit is

$$x_{ji} = (x/_j \ldots\ldots\ldots\ldots x_{nj}).$$ [cf. Anderberg (1973); Le Maitre (1982)]

However, some authors recommend to use a squared version of the above expression to avoid complication in calculation.

In this connection it is noteworthy that euclidean distance is scale dependent and hence will be weighted in favour of variables with large numerical values. Very often in petrologic problems it is found that a variable occurring in a relatively minor proportion may turn out to be a very good discriminant between two groups, compared to others, occurring in considerably large proportions. [e.g. the proportion of TiO_2 (occurring in $<2\%$) is a very much trusted discriminant in categorising orogenic andesites from its anorogenic counterparts, cf. Gill (1981)]. As a result it is strongly recommended that each score should be standardised to have zero mean and unit variance, before use.

Another method of distance measurement between two sets of multivariate data is Mahalanobis D^2 statistic, as described below: Let, $\bar{x}_\alpha^{(i)}$ and $\bar{x}_\alpha^{(j)}$ be the score vectors corresponding to *i*th and *j*th element, where,

$$\bar{x}_\alpha^{(i)} = (x_{1\alpha}^{(i)} \quad \ldots\ldots\ldots x_{p\alpha}^{(j)}), \alpha = 1 \text{ (i) } n_i$$
$$\text{and } \bar{x}_\alpha^{(j)} = (x_{1\alpha}^{(j)} \quad \ldots\ldots\ldots x_{p\alpha}^{(j)}), \alpha = 1 \text{ (i) } n_j$$

$$\text{then } D_{ij}^2 = \frac{n_i \; n_j}{n_i + n_j} \; [\bar{x}^{(i)} - \bar{x}^{(j)}]' \, S^{-1} \, [\bar{x}^{(i)} - \bar{x}^{(j)}]$$

where, $\bar{x}^{(j)} = i/n_i \, \Sigma_\alpha^{(i)}$ & $x^{(j)} = 1/n_j \, \Sigma \, x_\alpha^{(j)}$

$$S = i/(n_i + n_j - 2) \left[\sum_{\alpha=}^{n\,i} [\bar{x}_\alpha^{(i)} - \bar{x}^{(i)}] [\bar{x}_\alpha^{(i)} - \bar{x}^{(i)}]' \times \sum_{\alpha=1}^{n} [\bar{x}_\alpha^{(j)} - \bar{x}^{(j)}] [\bar{x}_\alpha^{(j)} - \bar{x}^{(j)}] \right]$$

It is noteworthy that the change of base and scale as is needed for euclidean distance is not required for Mahalanobis D^2, since it is invariant under change of base and scale, a very much desirable property for any distance function.

The similarity measures described in the preceding sections may be used to construct a similarity matrix describing the strength of all pairwise relationships among the entities (variables or data units) in the data set. The methods of hierarchical cluster analysis operate on this similarity matrix to construct a 'tree' depicting specific relationship among the different entities. Extreme branches represent the individuality while the root represents the entire collection of entities. Hierarchical clustering methods (which build a tree from branches to the root) are often called as 'agglomerative methods'—the principle of such method is listed in the following flow chart: Let s_{ij} be the similarity between entities i and j as defined by one of the similarity measures. Define $S=[s_{ij}]$ [It is to note that $s_{ij}=s_{ji}$, $i, j=1$ (i) n.].

1. Begin with n nos. of clusters, each consisting of exactly one entity. Let the clusters be labelled with the numbers 1 through n.

2. Search the similarity matrix for the most similar pair of clusters. Let the chosen cluster be labelled as p and q and let their associated similarity be s_{pq}, $p>q$.

3. Reduce the number of cluster by 1 through merger of clusters p and q. Label the product of the merger q and update the similarity matrix entries in order to repeat the raised similarities between cluster q and all other existing clusters. Delete the row and column of S pertaining to cluster p.

4. Perform steps 2 and 3 a total of (n—1) times [at the point where all entities will be in one cluster]. At each stage record the identity of the clusters which are merged and the value of similarity between them in order to have a complete record of results.

DIFFERENT LINKAGE METHODS

In this section different methods of hierarchical clustering are described as follows:

Single Linkage Method

At each stage, clusters p and q have been merged, the similarity between the new cluster (labelled t) and some other cluster r is determined as follows:

If s_{ij} is a distance like measure (e.g. euclidean distance) $s_{tr}=\min(s_{pr}, s_{qr})$

If clusters t and r were to be merged then for any entity in the resulting cluster the distance to its nearest neighbour would be at most s_{tr} [cf. Le Maitre (1982, p. 166) for detailed discussion].

Single linkage clustering is invariant to any transformation which leaves the ordering of the similarities unchanged.

COMPLETE LINKAGE METHOD

At each stage after cluster p and q have been merged the similarity between the new cluster (labelled t) and some other cluster r is as follows:

If s_{ij} is a distance like measure: $s_{tr} = \max (s_{pr}, s_{qr})$

If clusters t and r were to be merged then every entity in the resulting cluster would be no farther than s_{tr} from other entities in the cluster.

Apart from these, two other linkage methods, viz., *unweighted average method* and *weighted pair group average method* were also used, a detailed discussion of which can be found in Le Maitre (1982, pp. 167-168).

COMPUTATION AND OBSERVATION

To trace out the bearing of similarity measures and method of clustering on the disposition of the dendrogram, we have in the first step computed euclidean (standardised) distance matrix and Mahalanobis D^2 matrix (vide Tables 2 & 3 respectively) for the eleven BMB granite plutons (here each pluton constitutes a group), using the aforesaid five mineralogical attributes as variables. In the second step, for each of the two above mentioned distance matrices, four dendrograms were constructed [Fig. 1(a-d) using euclidean distance and Fig. 2(a-d) using Mahalanobis D^2 statistics]. The clustering methods, used were single linkage (Figs. 1a & 2a), complete linkage (Figs. 1b & 2b), unweighted average linkage (Figs. 1c & 2c), weighted pair group average linkage (Figs. 1d & 2d). It is noted that groups 1 (Tisri pluton) and 11 (Simratari), 2 and 4 (Chauki and Banresar respectively) are in all cases clustered together. Group 3 (Manihari) is clustered in some dendrograms with groups 1 and 11 [Fig. 1(a-d)]

TABLE 2

EUCLIDEAN DISTANCE (STANDARDISED) MATRIX

	1	2	3	4	5	6	7	8	9	10	11
1	0.000										
2	2.686	0.000									
3	1.455	1.607	0.000								
4	2.803	1.295	1.984	0.000							
5	3.983	1.895	2.978	2.439	0.000						
6	3.217	2.940	3.142	3.066	3.900	0.000					
7	4.087	2.797	3.810	2.445	3.764	2.548	0.000				
8	2.828	3.174	2.952	2.291	3.970	4.092	3.676	0.000			
9	3.826	2.016	2.995	2.572	3.287	3.004	2.633	4.310	0.000		
10	4.435	2.653	3.581	2.955	3.197	5.746	4.677	3.742	4.221	0.000	
11	0.876	2.587	1.467	2.948	3.807	2.920	5.151	3.143	3.840	4.629	0.000

S. Nos. 1 to 11 represent the different groups, as explained in Table 1.

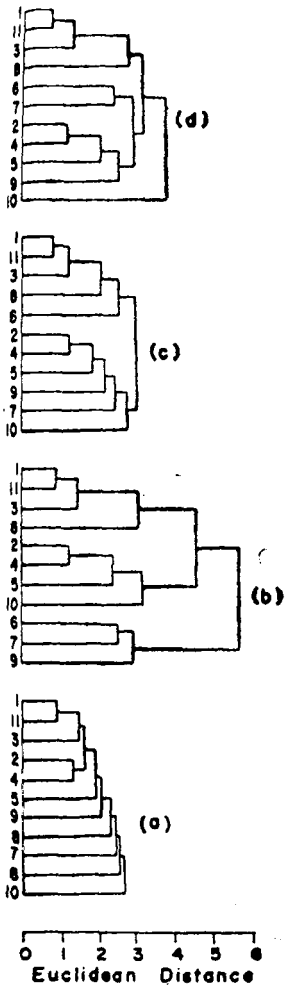

Fig. 1 (*a-d*) Dendrograms constructed on the basis of Euclidean distance, using (*a*) single linkage, (*b*) complete linkage, (*c*) unweighted average linkage method, and (*d*) weighted pair group average method respectively. Serial numbers representing different granite groups are same as in Table 1.

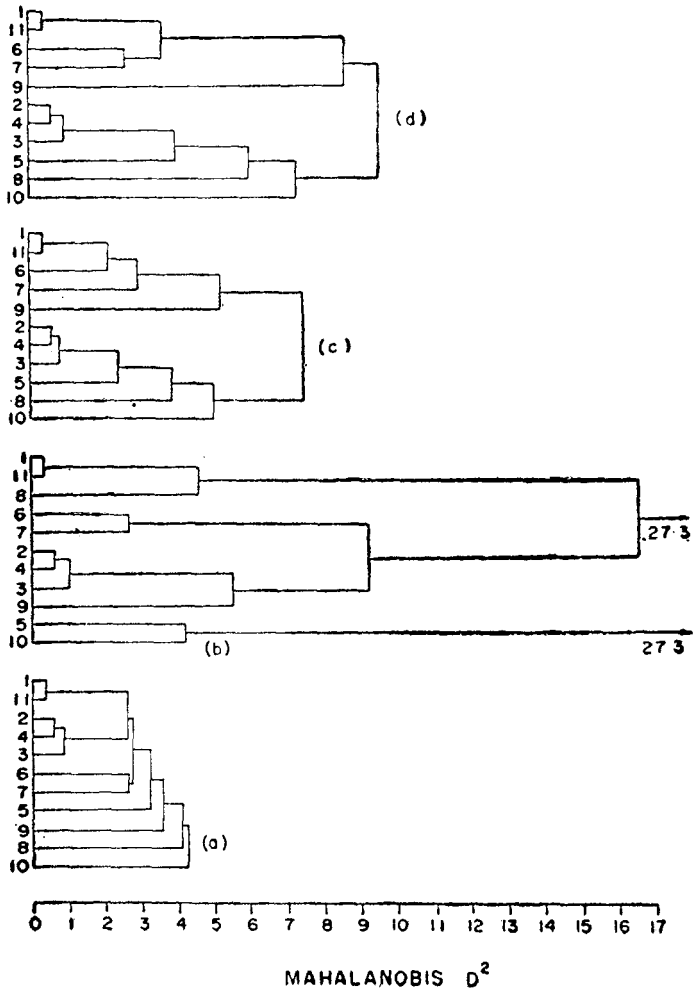

Fig. 2 (*a-d*). Dendrograms constructed on the basis of Mahalanobis D^2 statistic using (*a*) single, (*b*) complete, (*c*) unweighted average, and (*d*) weighted pair group average linkage methods respectively. Group serial numbers are same as in Table 1.

while in others with 2 and 4 [Fig. 2(a-d)]. In 7 out of 8 dendrograms groups 6 and 7 (Barmi and Debaur respectively) are either merged with each other to form an althogether separate branch or, are located on a branch just following/preceding the earlier/later one [cf. Figs. 1(a-b) & 2(a-d)]; only in the dendrogram for average linkage within new groups using euclidean distance (Fig. 1c) such relation is absent. For the rest of groups, viz., 5 (Gawan), 8 (Khobarwa), 9 (Bandapahar) and 10 (Kalapahar) no definite clustering criteria can be deciphered from the aforesaid dendrograms.

TABLE 3

MAHALANOBIS—D^2 MATRIX

	1	2	3	4	5	6	7	8	9	10	11
1	0.000										
2	5.285	0.000									
3	2.572	0.863	0.000								
4	5.438	0.586	0.974	0.000							
5	13.747	3.170	5.287	3.500	0.000						
6	3.225	7.734	6.237	9.185	17.152	0.000					
7	3.925	3.781	4.218	4.514	12.174	2.621	0.000				
8	4.634	6.180	4.079	4.167	9.546	10.936	8.406	0.000			
9	10.098	3.519	5.480	5.255	10.737	8.813	4.147	16.494	0.000		
10	18.520	6.321	8.787	5.353	4.219	27.337	17.093	11.599	13.739	0.000	
11	0.372	5.508	2.907	6.107	12.787	2.714	4.527	4.516	11.252	19.605	0.000

S. Nos. 1 to 11 represent the different groups (i.e. granite plutons) as explained in Table 1.

It is noteworthy that each of these observations are only in partial tally with the geological inferences on the same granite groups (cf. Saha *et al*, 1987) and thus complete parity between these two lines of observations has not been found in any single instance.

DISCUSSION

From the preceding sections it has been noted that unless there is a very strong sense of 'nearness' amongst the groups (e.g. groups I & II or, 2 & 4) the pattern of dendrograms (even for a single set of data) varies distinctly for different similarity measures and/or methods of clustering. Therefore it is imperative to carry out cluster analysis (for any set of data) on the basis of at least two different similarity measures using as many linkage methods as possible. However for a preliminary investigation (in case of a cursory survey) Mahalanobis D^2 statistic using average linkage clustering (of both types) methods can be tested for the reasons discussed below.

Mahalanobis D^2 statistic originates from testing the equality of location of two multivariate normal populations, therefore it possesses some distinct optimum properties. Besides this, it

can be evaluated without difficulty when the number of samples in two groups are unequal. Whereas for the euclidean distance we have to apply the measure on some scaled central value (e.g. mean) of the variables, instead of the original scaled variables.

While using single linkage method on distance like measure, the minimum of two 's_{ij}' values preserve the triangular inequality of the metric (in contrast to any other linkage methods) which can be claimed to be a unique property of this method. Nevertheless the average linkage methods (of both the types) take into account the concept of a sort of compromise between the procedures of single linkage method on one hand and complete linkage on the other and thus furnish a more realistic and comprehensive picture to user.

If it is conceived that the efficacy of forming cluster of a particular method is best measured when maximum number of clustering is achieved at the closest proximity of a perfect similarity (e.g. closest to zero, for distance like measures), then both single linkage and average linkage methods usually turn out to be equivalent to each other.

ACKNOWLEDGEMENT

The authors are indebted to Prof. A. K. Saha, Department of Geology, Presidency College, Calcutta, for his supervision and to Dr. S. L. Ray, B.E. College, Howrah, for reviewing the manuscript. Research grants from the U.G.C. programme "Financial Assistance for Minor Research Projects" and the University of Burdwan, to S.S.S. and A.C. respectively, are thankfully acknowledged. Thanks are due to the Head, Computer Science Unit, Indian Statistical Institute, Calcutta, for providing computational facilities.

REFERENCES

ANDERBERG, M. R., 1973. Cluster Analysis for Application. Academic Press, New York, 359 p.

GILL, J. B., 1981. Orogenic Andesites and Plate Tectonics, Springer-Verlag, Berlin, 390 p.

LE MAITRE, R. W., 1982. Numerical Petrology, Elsevier, Amsterdam, 281 p.

MITRA, M., 1984. Geology of the Area Around Tisri, Giridih District, Bihar, With Special Reference to the Granite Emplacement, Unpublished M.Sc. Thesis, University of Cleutta, 210 p.

MUKHOPADHYAY, B., 1981. Geology of the Area Around Pihira, Giridih and Hazaribagh Districts, Bihar, With Special Reference to the Lithium Bearing Pegmatites. Unpublished M.Sc. Thesis, University of Calcutta, 114 p.

MUKHOPADHYAY, P., 1974. Geology of the Area in and Around Singho, Giridih District, Bihar, With Special Reference to the Structural and Petrological Evolution of the Metasediments. Unpublished M.Sc. Thesis, University of Calcutta, 147 p.

RAY, A., 1985. A Structural-Geochemical-Geomathematical Study of the Genesis of Bandapahar Granite, Bihar Mica Belt. Unpublished M.Sc. Thesis, University of Calcutta, 86 p.

SAHA, A. K., SARKAR, S. S. AND REJ, S. S., 1987. Petrochemical Evolution of the Bihar Mica Belt Granites, Eastern India. *Ind. J. Earth Sci.*, Vol. 14, No. 1, pp. 22-45.