MAUSAM, **60**, 2 (April 2009), 185-196

551.577.2 : 551.553.21

Precipitation regionalization of southwest monsoon by hierarchical cluster analysis

A. B. MAZUMDAR

India Meteorological Department, Pune, India

(*Received 3 December 2003, Modified 3 April 2008*)

e mail : abmazumdar@gmail.com

सार – इस शोध पत्र में पहले से चले आ रहे परम्परागत समूहों का आकलन करते हुए भारतीय क्षेत्र में दक्षिण पश्चिम मानसून वर्षा की स्थितियाँ उत्पन्न करने वाले क्षेत्रों की पहचान करने का प्रयास किया गया है। विभिन्न समेकन पद्धतियों द्वारा तैयार किए गए वृक्ष आरेखों की जाँच से मौसम विज्ञानिक उपखंडों के 13 मूल समूहों (न्यूकिल्आई कल्स्टर्स) की उपस्थिति का पता चला है। इस मूल समूहों की बनावट को उनके औसत प्रमुख घटक (पी. सी.) स्कोरो और पी. सी. से संबद्ध सिनॉप्टिक लक्षणों द्वारा स्पष्ट किया जा सकता है। उपखंडों के विभिन्न प्रकारों के समूह तैयार करने के लिए विभिन्न समेकन पद्धतियों में उच्च स्तर के अंतः मूल समूह इंटर– न्यूकिल्आई जॉयनिंग्स बने।

मौसम विज्ञानिक उपखंडों के सुवितरित समूह उपलब्ध कराने वाली लचीली पद्धति उपयुक्त पाई गई है। इस पद्धति से भारत में वर्षा के छह समरूपी क्षेत्रों की पहचान की गई है। मानसून की स्थितियाँ उत्पन्न करने वाले क्षेत्रों में मौसम विज्ञानिक उपखंड समान रूप से वितरित पाए गए हैं। इस पद्धति से प्राप्त किए गए परिणाम तर्कसंगत और अधिकांशतः व्याख्या करने के योग्य रहे है।

ABSTRACT. An attempt has been made to identify coherent zones of southwest monsoon rainfall over the Indian region by employing hierarchical cluster analysis. Examination of dendrograms produced by different fusion strategies revealed the presence of 13 nuclei clusters of meteorological subdivisions. Formation of these nuclei clusters could be interpreted by their average principal component (PC) scores and associated synoptic features of PCs. Higher level inter-nuclei joinings have occurred in various fusion strategies to produce different types of clusters of subdivisions.

A flexible strategy providing well separated groups of meteorological sub-divisions has been found to be suitable. The method has identified six homogeneous regions of rainfall over India. The meteorological subdivisions have been found to be evenly distributed in these coherent zones. The clustering obtained by this method has been reasonable and largely interpretable.

Key words – Precipitation regionalization, Coherent monsoon rainfall zones, Hierarchical cluster analysis.

1. Introduction

Considerable attention has been directed at investigating the application of various statistical and numerical methods for the purpose of defining spatial patterns of weather elements, permit its regionalization or demonstrate its spatial organization. The approaches used for these purposes centre either on traditional spatial correlation analysis, or the eigentechnique, or principal component analysis (PCA) or PCA along with cluster analysis.

Over the last two decades multivariate analyses using eigentechniques, like, PCA, have been in extensive use for obtaining precipitation regionalization over various locations of the globe (for example, Ehrendorfer (1987), Pandzic (1988), Ogallo (1989), Eklundh and Pilesjo (1990), Murata (1990), Opoku-Ankomah and

Cordery (1993), Drosdowsky (1993), etc.). All these studies have utilized various types of rotations of principal components (PC) to arrive at regionalizations in their respective regions. In this method of regionalization the PCs are rotated to obtain simple structure (Richman, 1986), when the objects lie close to the PC axes.

In a number of studies the regionalization have been arrived by the application of PCA followed by cluster analysis to group areas according to PC attributes (for example, Willmott, 1978; Anyadike, 1987; Spackman and Singleton, 1982; Singleton and Spackman, 1984, Sumner *et al*., 1993; Drosdowsky 1993). An advantage of cluster analysis over other methods of regionalization is that the population structure can be viewed in complete multidimensional space when there are several attributes (PCs), unlike the approximate distances suggested in two dimensional space, when comparing two components at a time.

A limited number of attempts have been made to identify homogeneous zones by the application of eigentechniques utilizing Indian rainfall data. Gadgil and Iyengar (1980) applied empirical orthogonal function (EOF) analysis on rainfall data collected at 53 stations of peninsular India to identify homogeneous regions using two dominant EOFs. The resulting clusters contained stations that were geographically contiguous and the rainfall patterns were logically interpretable. In a related study for India, Gadgil and Joshi (1984) used similar technique to develop climatic regions, using mean monthly values of precipitation, minimum temperature and a moisture index for 119 stations. Prasad and Singh (1988) obtained precipitation regionalization utilizing seasonal monsoon rainfall of 31 sub-divisions of India for 80 years to classify Indian region into 7 homogeneous region. They also used two dominant EOFs to arrive at the regionalization. Iyengar (1991) also obtained coherent zones in Karnataka by using two dominant EOFs. In all these studies the groupings and their separations were obtained in a two dimensional space defined by the first two EOFs. Shukla (1987) identified 7 homogeneous zones based on EOFs and correlation analysis of monsoon rainfall. Kulkarni and Reddy (1994) classified districts of Andhra Pradesh on the basis of seasonal and annual rainfall by using a clustering technique.

In the present study a regionalization in multidimensional space of PCs has been attempted to identify coherent zones during the southwest monsoon period by employing Hierarchical Cluster Analysis (HCA). The broad-scale synoptic features that could have influenced the groupings have also been discussed with the help of earlier studies by Mazumdar (1998b) and De and Mazumdar (1999), bringing out possible association between PCs and broad-scale weather patterns.

The identification of coherent sub-regions of India, as obtained in this study by HCA would be useful for agricultural planning, precipitation network design, water resources management and in medium range prediction of rainfall during the monsoon season.

2. Hierarchical Cluster Analysis (HCA)

An outline of the hierarchical cluster analysis employed in the study is provided in this section following Mather (1976) and Mazumdar (1995).

If there are '*n*' objects to be classified, the HCA assumes that each group at level '*i*' is part of a larger group at level '*i* +1' and all groups are submerged into a universal cluster at level '*n*-1'.

The grouping in the HCA are achieved by utilizing a measure of similarity (*e.g*., correlation coefficient) or dissimilarity (*e.g*., Euclidean distance coefficient, d_{ij}). The later being the most commonly used measure for the HCA.

Given two objects '*i*' and '*j*', the coefficient 'd*ij*' is defined in a multidimensional space with 'p' orthogonal axes as:

$$d_{ij} = \frac{\left[\sum_{k=1}^{p}\left(X_{ik} - X_{jk}\right)^2\right]^{1/2}}{p} \qquad (2.1)$$

Where 'X_{ik}' and 'X_{jk}' are the coordinates of the objects '*i*' and '*j*' corresponding to axis 'k'.

A large number of HCA methods are available. The general strategy underlying each of the clustering methods is similar and can be summarised as follows:

Step I : A distance matrix D containing Euclidean distances d_{ij} is constructed.

Step II : The smallest value of d_{ij} in D is identified.

Step III : Objects '*i*' and '*j*' are fused into a group '*k*'.

Step IV : New distances d_{km} (where 'm' represents each of the remaining points) are computed. Distances 'd_{im}' and 'd_{jm}' in D are replaced by the new distances.

Step V : Steps II to IV are repeated for (*n*-1) cycles.

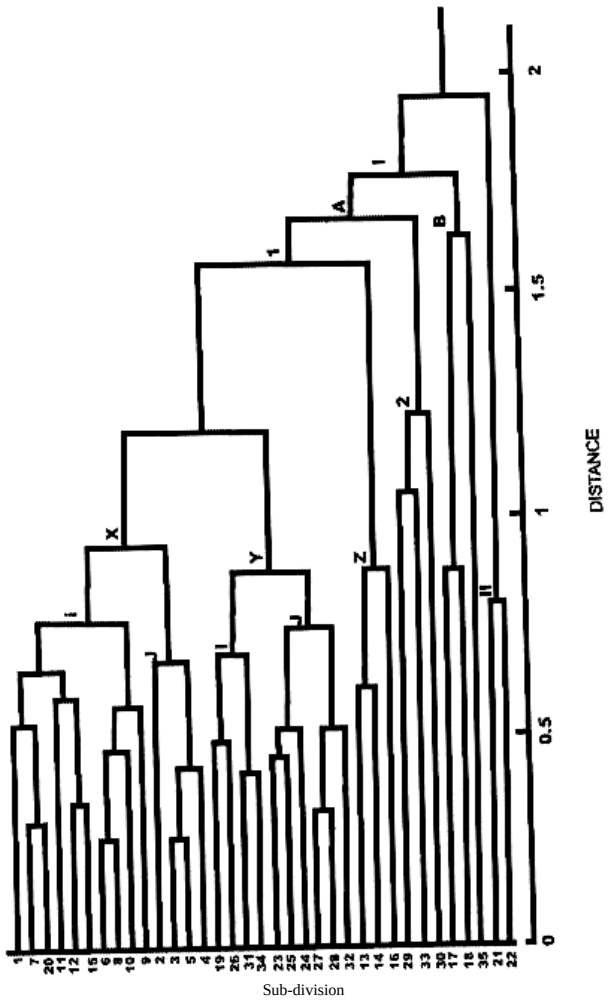


Fig. 1. Dendrogram of Group Averages method. The numbers of meteorological sub-divisions are shown in the abscissa. The locations of these sub-divisions have been indicated in Fig. 4

Various methods can be adopted to combine two objects or groups. When two groups of objects '*i*' and '*j*' with 'n_i' and 'n_j' members, respectively, at a smallest distance 'd_{ij} ' are fused to form new group 'k' with $n_k = n_i + n_j$ members, the new distances 'd_{km}' can be calculated by :

$$d_{km} = \alpha_i d_{im} + \alpha_j d_{jm} + \beta d_{ij} + \gamma | d_{im} - d_{jm} | \qquad (2.2)$$

Clustering algorithms are identified by the values of αi, αj, β and γ used. Different fusion techniques lead to space-distortion, because the initial matrix D defines a space containing all the objects and as the groups from, the updated D matrix does not define a space with the original properties. Sometime the effect leads to space contraction when the groups seem to move closer to some or all the remaining objects. Thus, the chance that an object will join an existing group rather than act as the nucleus of a new group is increased. Other strategies may produce space-dilating effects, when groups appear to recede on formation.

In the nearest neighbour algorithm, the distance between two group '*i*' and '*j*' is defined as the distance between their two closest members. The coefficients of equation (2.2) being $\alpha_i = \alpha_j = 0.5$, $\beta = 0$ and $\gamma = -0.5$. The algorithm has space-contracting properties.

The farthest neighbour method is converse of the nearest neighbour method. The distance between pairs of groups is defined as the distance between their farthest members. The coefficients of equation (2.2) are $\alpha_i = \alpha_j = \gamma = 0.5$, and $\beta = 0$. The resulting clusters are well separated and it is a space-dilating technique.

In the centroid method, the inter-group distance is defined as the distance between the centroids (multivariate means) of the two groups. In this method the coefficients are $\alpha_i = n_i / n_k$, $\alpha_j = n_j/n_k$, $\beta = \alpha_i\alpha_j$ and $\gamma = 0$. The method is space conserving.

The coefficients in the median method are selected to have the centroid of a new group to lie at the mid-point of the shortest side of the triangle joining the centroid of '*i*', '*j*' and any other group or point 'm'. The coefficients in this procedure are: $\alpha_i = \alpha_j = 0.5$, $\beta = -0.25$, and $\gamma = 0$.

In the group average method the distance between two groups is defined as the arithmetic average of distances between pairs of members of '*i*' and '*j*', *i.e.*, as $1/n_i n_j \Sigma_i \Sigma_j d_{ij}$. The coefficients in this scheme are $\alpha_i = n_i/n_k$, $\alpha_j = n_j/n_k$, and $\beta = \gamma = 0$. The group average method is space-conserving.

The simple average method gives equal weight to two group to be joined irrespective of the values of ni and nj, with the values of the coefficients being $\alpha_i = \alpha_j = 0.5$ and $\beta = \gamma = 0$. The method leads to space dilation.

In the Ward method, the two groups to be combined at any given level are those whose fusion produces the least increase in the within-group sum of squares of distances. The coefficients of the scheme are: $\alpha_i = (n_m+n_i)/(n_m + n_k)$, $\alpha_j = (n_m + n_j)/(n_m+n_k)$,

$\beta = -n_m/(n_m + n_k)$ and $\gamma = 0$, where 'n_m' is the number of objects in any other group. The method is space conserving.

It is possible to select the values of the coefficients in the flexible strategy provided that :

$$\alpha_i + \alpha_j + \beta = 1$$

$$\alpha_i = \alpha_j$$

$$|\beta| \leq 1$$

$$\gamma = 0 \qquad (2.3)$$

By choosing values of β in the range from –1 to +1, the method can be made to range from space-dilating ($\beta = -1$) to space contracting ($\beta = 1$).

Considerable care is necessary when using the method of HCA because there is divergence of views concerning the choice of similarity measure and, subsequently, the most appropriate fusion strategy of hierarchical techniques (Williams, 1971, 1976; Everitt, 1980). For meteorological classification studies, two methods are in common use: centroid (Morgan, 1971) and Ward method (Willimott, 1978; Anyadike, 1987; Stone, 1989). As several workers have emphasized (Moore *et al*., 1972; Webster, 1975, 1977; Everitt, 1980) more than one technique should be used and if similar groups are produced then they are worth further interpretation. The clusters obtained by any of the method could be selected, provided the groupings are realistic and interpretable.

The results of the HCA can be presented in a tabular form known as the linkage order or diagrammatically in the form of a linkage tree or dendrogram (Fig. 1).

2.1. *Selection of significant PCs*

One of the important issues involved in the HCA is the selection of significant PCs to be used for such analysis. Two types of criteria have been developed to address the problem of determining the number of significant component. The first type assesses the significance of the eigenvalue against a theoretical or simulated value. The simplest of these is the Kaiser-Guttman test (Kaiser, 1958), which is based on the assumption that eigenvectors which have eigenvalues less than 1 for a correlation matrix explain less variance than uncorrelated white noise. The second type of test examines the sequence of eigenvalues. The Scree test (Cattel, 1966) looks for a cut-off in the difference between successive eigenvalues or a break in slope in eigenvalue sequence, with the eigenvalues representing noise

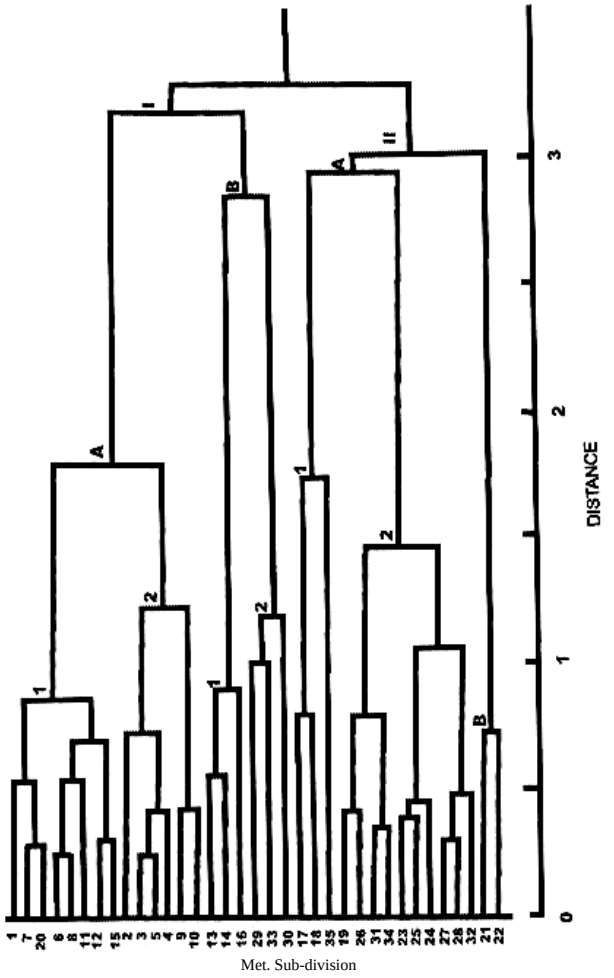


Fig. 2. Dendrogram of ward method. The numbers of meteorological sub-divisions are shown in the abscissa. The locations of these sub-divisions have been indicated in Fig. 4

decreasing in geometric progression. Craddock and Flood (1969) suggest the break is more distinct in a plot of log eigenvalue (LEV) against eigenvector number. North *et al*. (1982) used sampling theory to establish error limits for the eigenvalues, and suggested that eigenvectors whose eignevalues overlap, form degenerate multiplets and should not be split when truncating the eigenvector sequence.

3. Data

The input data used for the HCA have been the elements of principal component (PC) score obtained from the PC analysis in *T*-mode of a special data set of weekly rainfall anomalies of the 35 meteorological sub-divisions of India. The sub-divisional weekly rainfall anomalies pertain to a period of ten years from 1977 to 1986, when the entire country was under the spell of southwest

monsoon from June to September each year. The first and the last weekly rainfall anomaly values in any year correspond to the week of complete establishment of SW monsoon over the entire country and the week corresponding to the commencement of withdrawal from the country, respectively. More details regarding preparation of data set are available in Mazumdar (1995, 1998a).

4. Methodology

A PC analysis has been done on a data set of weekly rainfall anomalies in temporal domain. Details of PCA are given in earlier studies by Mazumdar (1995, 1998b). The score matrix, obtained as one of the outputs of the PC analysis in *T*-mode, providing the corrdinates of the meteorological subdivisions of India in a multidimensional space defined by the significant orthogonal PCs, has been used to compute the distance matrix. Different clustering methods have been applied using the distance matrix as the input and a regionalization has been obtained as per interpretability and suitability of resulting clusters.

The sampling error test (North *et al*., 1982) the Scree test (Cattel, 1966) and log-eigenvalue (LEV) test have been applied to decide the number of significant PCs to be retained for the HCA.

5. Result and discussion

As per the sampling error test (North *et al*., 1982), first four PCs have been found to be significant (Mazumdar 1998b), whereas, the Scree and LEV plots suggest first six PCs to be significant. Considering this, average of these two has been utilized for the clustering.

The results of PC analysis and the synoptic features associated with first four PCs in temporal domain have been presented in earlier studies by Mazumdar (1998b) and De and Mazumdar (1999). Important results of these studies have been:

(*i*) The PC I has association with the most active phase of the monsoon with well defined east-west oriented monsoon trough systems, westerly/west-north-westerly movement of intense low pressure areas from the Bay of Bengal, systems in westerlies affecting western and northwestern parts of the country.

(*ii*) The PC II has association with north-westerly movements of intense cyclonic systems originating from Bay of Bengal, active eastern half of monsoon trough systems and systems in westerlies affecting northern/northwestern parts of the country;

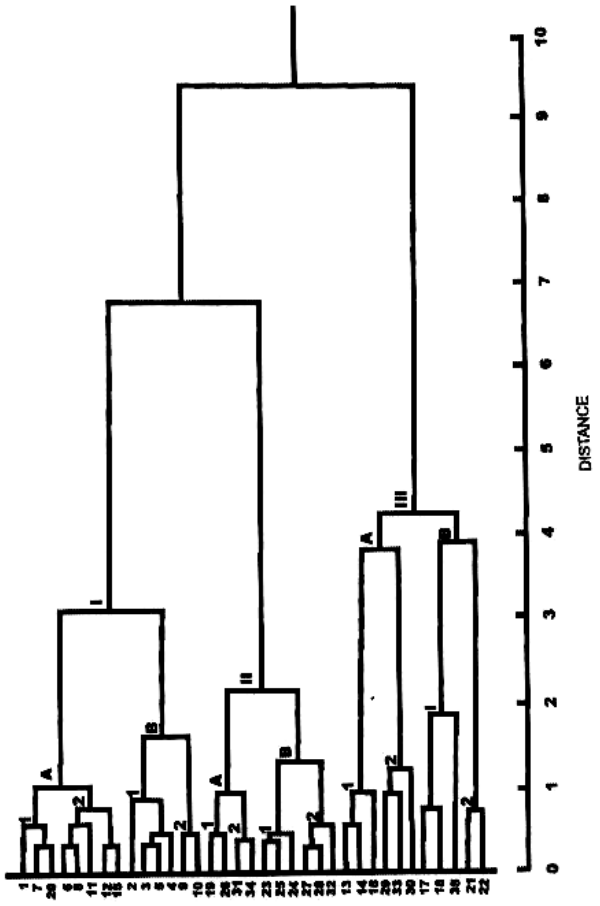


Fig. 3. Dendrogram of flexible (= -0.5) method. The numbers of meteorological sub-divisions are shown in the abscissa. The locations of these sub-divisions have been indicated in Fig 4.

(*iii*) The PC III has association with *in-situ* formation of low pressure systems over the land areas of Uttar Pradesh (UP) and Madhya Pradesh (MP), active eastern half of the monsoon trough and off-west coast trough;

(*iv*) The PC IV has association with higher than normal southward tilt with height of short-lived cyclonic systems originating from Bay of Bengal & monsoon trough system, systems in westerlies affecting northern parts and systems in easterlies affecting southern parts of the country.

A total of 15 clustering runs have been undertaken using all the fusion strategies outlined in section 2. The total includes eight computations of flexible strategies with different values of β.

Scrutiny of the dendrograms obtained from all the fusion strategies indicates presence of 13 cluster nuclei as

TABLE 1

The listing of nuclei clusters and the average PC scores of these nuclei clusters

Cluster nucleus number	Numbers of Met. sub-divisions *	Average scores			
		PC I	PC II	PC III	PC IV
1	1, 7, 20	-0.003	0.124	0.746	-0.011
2	6, 8, 11	-0.247	1.069	0.732	-0.162
3	12, 15	0.079	0.899	-0.168	0.090
4	2, 3, 4, 5	-0.016	0.238	0.067	-1.349
5	9, 10	-1.135	1.022	1.186	-0.913
6	13, 14, 16	-.210	1.564	-0.717	1.707
7	29, 30, 33	-1.534	-.0986	-1.906	0.082
8	17, 18, 35	1.197	0.322	-1.906	0.082
9	19, 26	1.157	-0.266	0.679	0.136
10	31, 34	0.274	-0.784	0.640	0.701
11	23, 24, 25	0.401	-1.551	0.810	0.101
12	27, 28, 32	-0.324	-1.207	0.348	1.105
13	21, 22	2,108	0.842	0.792	-1.781

* Names and locations of Met. sub-divisions are indicated in Fig. 4

TABLE 2

The clusters of Group Average and average scores of the clusters

Cluster	Numbers of Met. sub-divisions *	Average scores			
		PC I	PC II	PC III	PC IV
IA1X*i*	1, 7, 20, 11, 12, 15, 6, 8, 10, 9	-0.379	0.783	0.694	-0.29
IA1X*j*	2,3,4,5	-1.016	0.238	0.067	-1.349
IA1Y*i*	19, 26, 31, 34	0.716	-0.525	0.659	0.418
IA1Y*j*	23, 24, 25, 27, 28, 32	0.038	-1.379	0.629	0.603
IA1Z	13, 14, 16	0.210	1.564	-0.717	1.707
IA2	29, 33, 30	-1.534	-0.986	-1.906	0.082
IB	17, 18, 35	1.197	0.322	-1.132	0.149
II	21, 22	2.108	0.842	0.792	-1.781

* Names and locations of Met. sub-divisions are indicated in Fig. 4

tabulated in Table 1. The final structure of almost all the clustering strategies could be explained by different combinations of these core nuclii clusters. The difference in the various dendrograms have been found to occur as a result of inter-nucleus joinings at different distances (Figs.1, 2 and 3) in various strategies.

The core clusters of Table 1 are interpretable to a great extend with the help of average values of PC scores and the background knowledge of associations of PCs with broad scale synoptic patterns as brought out in earlier studies by Mazumdar (1998b) and De and Mazumdar (1999) and as mentioned briefly above.

The first nucleus of Andaman and Nicobar Islands, Orissa and east MP & Chattisgarh appears to have formed as a result of formation and movement of cyclonic systems from Andaman sea to the northern parts of east coast during the course of one week. The high score of this group on PC III may have occurred because of the movement of the cyclonic systems close to these areas before their stagnation over UP and MP.

The second nucleus of Gangetic West Bengal (WB), Jharkhand and west Uttar Pradesh and the fifth nucleus of Bihar and east UP have formed, perhaps, because of northwesterly movement of intense cyclonic systems originating from the North Bay and *in-situ* formation of low pressure systems over UP. That is why, both these core clusters have high scores over PC II and III. These two nuclei can be differentiated from each other by their scores of PC I and IV, which are very low for nucleus 5, as westerly tracks of low pressure areas from the Bay of Bengal and high southward tilt of monsoon trough and low pressure system cause decreased rainfall over these areas.

The nucleus 3 consisting of Uttaranchal and Himachal Pradesh has high score on PC II, suggesting their joining due to northwesterly course of low pressure systems originating from the Bay of Bengal and recurvature and increased precipitation due to the presence of systems in mid-latitude westerlies and low level circulation/ system in westerlies.

The north-eastern states of Nagaland, Meghalaya, Arunachal Pradesh, Assam, Manipur, Mizoram, Tripura, Sikkim and Sub-Himalayan WB have joined together in nucleus 4 because of their very low scores on PC I and IV. It is well known that rainfall in these areas decreases when intense cyclonic systems from the Bay of Bengal have westerly courses and the cyclonic systems have large tilt to south. The cluster get good rainfall during epochs of weak/break monsoon.

The Jammu and Kashmir, Punjab and Haryana have clustered together in nucleus 6, because of their high scores on PC II and PC IV and low scores on PC III. These may have occurred mainly because of the effect of mid-latitude westerly systems, northwesterly course of cyclonic systems from the Bay and their recurvature and decreased activity in these areas when almost stationary, *in-situ* formation of low pressure areas occur over UP and MP.

The nucleus 7 consisting of Tamil Nadu, Rayalasleema and south-Interior Karnataka have very low scores on the first three PCs because of decrease/increase in rainfall activity during strong/weak monsoon situations represented by these three PCs. Relatively weaker monsoon situations represented by PC IV cause slightly above normal rainfall in these areas.

The clustering of Rajasthan with Lakshadweep Island in nucleus 8 is difficult to appreciate. It may be due to some distant effect. Perhaps, this may be the effect of successive northward propagation of monsoon pulses, when one pulse reaches north-western part of the country a fresh one appears over the south-western parts. The nucleus has high scores on PC I and low scores on PC III, which could be accounted for Rajasthan. The relationship needs further examination from synoptic angle.

West MP has joined Vidarbha to form the nucleus 9. Both have high scores on PC I and PC III which can be readily explained. Explicitly, westward moving systems having long tracks accentuates the monsoon currents from the Arabian sea and produce good rainfall in these sub-divisions.

Kerala and coastal Karnataka form the nucleus 10 as a result of their high scores on PC III and PC IV, and low score on PC II. Northward moving zonal cloud bands (Sikka and Gadgil, 1980) and shallow off shore troughs along the west coast could be the synoptic reasons for producing this cluster.

Konkan and Goa, Madhya Maharashtra and Marathwada forming the nucleus 11, have low scores on PC II, high scores on PC III and moderately high scores on PC I; these are interpretable considering the features of these PCs. Decreased rainfall activity due to northward shift of the monsoon trough and recurvature of low pressure systems contribute towards formation of this cluster.

The nucleus 12 consisting of coastal Andhra Pradesh, Telengana and North Interior Karnataka has high score on PC IV and low score on PC II which are interpretable. This group arises as a result of rainfall activity in association with systems moving in low latitudes symbolic of the weak monsoon situation. These regions also get low rainfall when low pressure areas move in the normal westerly/west-northwesterly directions and the monsoon trough is in normal or slightly north of the normal position.

Saurashtra & Kutch and Gujarat region makes up the last nucleus which also have interpretable high scores on PC I, moderately high scores on PC II and III and low score on PC IV. These areas get good rainfall from the Arabian sea branch of the monsoon when generally westward moving systems reach the north-western India.

TABLE 3

The clusters of Ward method and average scores of the clusters

Cluster	Numbers of Met. subdivisions*	Average scores			
		PC I	PC II	PC III	PC IV
IA1	1, 7, 20, 6, 8,11, 12, 15,	-0.043	0.554	0.514	-0.023
IA2	2,3,4,5, 9, 10	-1.076	0.63	0.627	-1.132
IB1	13, 14, 16	0.210	1.564	-0.717	1.707
IB2	29, 33, 30	-1.534	-0.986	-1.906	0.082
IIA1	17, 18, 35	1.197	0.322	-1.132	0.149
IIA2	19, 25, 31, 34, 23, 25, 24, 27, 28, 32	0.378	-0.952	0.619	0.511
IIB	21, 22	2.108	0.842	0.792	-1.781

* Names and locations of Met. sub-divisions are indicated in Fig. 4

These areas also get good rainfall due to mid-tropospheric cyclones (MTC).

Because of the different initial assumptions made, and also differences in deducting similarity or dissimilarity, various clustering approaches may yield different groupings for the same data. However, there seems to be general agreement that the one yielding the most logical or interpretable groups should be utilised. Everitt (1980) argues that the median method is inappropriate or incompatible with measures of correlation, and that the centroid method has the disadvantage that it produces groups of different sizes, the larger of which, when fused, tend to obliterate the smaller. The Ward method is more popular in climatology, primarily because it is based on mutually exclusive subsets, and does not assume normality, although it is primarily considered for use with population in excess of 100 (Cox. 1957).

Considering above and by visual scrutiny, three dendrograms, *viz*., group average, Ward and flexible (β = -0.5) have been selected for detailed study. The type of groupings obtained by applying these strategies are shown in Figs. 1, 2 and 3 and listed in Tables 2, 3 and 4.

The clusters are labeled by following multiple stages of branching. The first major branching are labeled as I, II, III, etc., further stages of branching by A, B, C, etc., followed by 1, 2, 3, etc., X, Y, Z, etc., and *i*, *j*, *k*, etc.

For the flexible (β = -0.5) and Ward methods three stages have been sufficient, but the group average required multiple stages of branching. The clusters of group average, Ward and flexible (β = -0.5) strategies are listed in Tables 2, 3 and 4, respectively. The average values of the PC scores of these clusters for the first four PCs are also shown in these tables. With the help of the feature of the 13 nuclei clusters (as discussed above), the features of the PCs and the average scores of each sub-cluster, it is possible, to a large extent, to appreciate probable influences which could have led to the formation of the sub-groups in each strategies.

The group average method has produced multiple stage clustering (Fig. 1). The number of members present in each group is also not distributed uniformly. Ten subdivisions have been clustered in one group in the cluster IA1X*i*. When a large number of sub-divisions are clustered together, the scores get evened-out and it becomes difficult to study the influences responsible for the clustering (Table 2). But, the interpretability is not lost completely, *e.g*., the largest cluster containing 10 sub-divisions (IA1X*i*) has moderately high scores on PC II and III and the resulting clustering could be due to the formation of low pressure areas over the Bay of Bengal and their subsequent movements in a north-westerly direction. The multiple stages of clustering and wide variations in the number of members present in each group make the clustering obtained by group average unsuitable.

TABLE 4

The clusters of flexible strategy (β = -0.5) and the average scores of the clusters

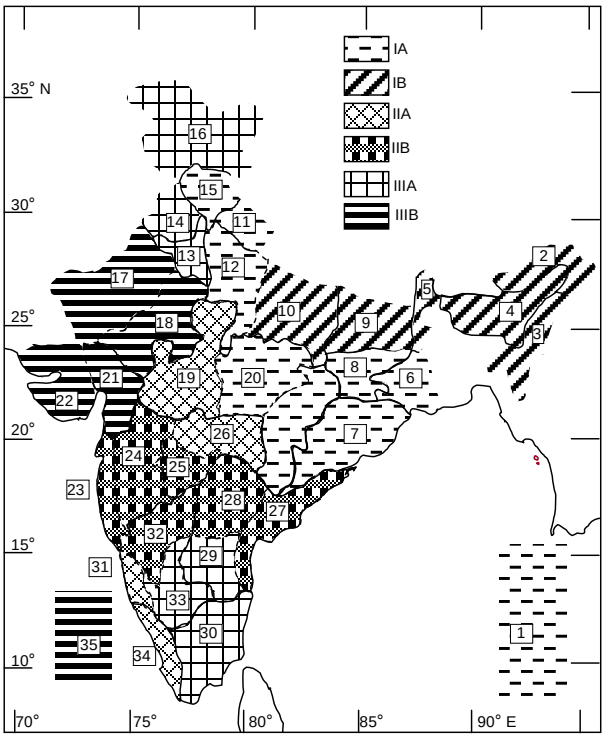
Cluster nucleus number	Numbers of Met. sub-divisions*	Average scores			
		PC I	PC II	PC III	PC IV
IA1	1, 7, 20	-0.003	0.124	0.746	-0.011
IA2	6, 8, 11, 12, 15	-0.03	1.033	0.107	-0.042
	Average of IA	-0.017	0.579	0.427	-0.027
IB1	2, 3, 4, 5	-0.016	0.238	0.067	-1.349
IB2	9, 10	-1.135	1.022	1.186	-0.913
	Average of IB	-1.076	0.630	0.627	-1.131
IIA1	19, 26	1.157	-0.266	0.679	0.136
IIA2	31, 34	0.274	-0.784	0.640	0.701
	Average of IIA	0.716	-0.525	0.660	0.419
IIB1	23, 24, 25	0.401	-1.551	0.810	0.101
IIB2	27, 28, 32	-0.324	-1.207	0.348	1.105
	Average of IIB	0.039	-1.379	0.579	0.603
IIIA1	13, 14, 16	0.210	1.564	-0.717	1.707
IIIA2	29, 30, 33	-1.534	-.0986	-1.906	0.082
	Average of IIIA	-0.662	0.289	-1.312	0.895
IIIB1	17, 18, 35	1.197	0.322	-1.906	0.082
IIIB2	21, 22	2.108	0.842	0.792	-1.781
	Average of IIIB	1.653	0.582	-0.17	-0.818

* Names and locations of Met. sub-divisions are indicated in Fig. 4

The clusters obtained by Ward method are better than that obtained by the group average method, as a three stage clustering is sufficient to obtain seven reasonable clusters (Fig. 2). With the help of the average scores (Table 3), it is possible to explain the major influences that could have produced these clusters. But the number of members present in each group varies considerably, which makes it unacceptable.

The flexible strategy with β = -0.5 produces six groups after two stage clustering, which can be further divided in twelve sub-groups by increasing one more stage of clustering (Fig. 3). The sub-divisions are distributed uniformly in groups and sub-groups. The groupings are interpretable by the utilization of average scores (Table 4) and the features of the nuclei groups presented earlier in this paper.

The first group formed as a result of amalgamation of cluster nuclei numbers 1, 2 and 3. The group has moderately high scores on PC II and III and low scores on PC I and IV. The joining could be attributed to long travel



1.	Andaman & Nicobar Islands	8.	Jharkhand	15.	Himachal Pradesh	22.	Saurashtra & Kutch	29.	Rayalaseema
2.	Arunachal Pradesh	9.	Bihar	16.	Jammu & Kashmir	23.	Konkan & Goa	30.	Tamil Nadu & Puducherry
3.	Assam & Meghalaya	10.	East Uttar Pradesh	17.	West Rajasthan	24.	Madhya Maharashtra	31.	Coastal Karnataka
4.	Nagaland-Manipur-Mizoram-Tripura	11.	West Uttar Pradesh	18.	East Rajasthan	25.	Marathwada	32.	North interior Karnataka
5.	Sub-Himalayan West Bengal & Sikkim	12.	Uttarakhand	19.	West Madhya Pradesh	26.	Vidarbha	33.	South interior Karnataka
6.	Gangetic West Bengal	13.	Haryana, Chandigarh & Delhi	20.	East Madhya Pradesh & Chattisgarh	27.	Coastal Andhra Pradesh	34.	Kerala
7.	Orissa	14.	Punjab	21.	Gujarat region, Daman, Dadra & Nagar Haveli	28.	Telangana	35.	Lakshadweep

Fig. 4. Coherent zones obtained by flexible strategy with $\beta = -0.5$

of intense low pressure systems originating from the Bay of Bengal and following a northwesterly course affecting up to western parts of UP.

The second group has formed due to joining of nuclei number 4 and 5, which have low scores on PC I and IV. This is also interpretable, as northeast India, Bihar and UP receive low rainfall when the cyclonic systems originating from the Bay have westerly courses and the rainfall is distributed mainly in the southern and central parts of India when the troughs and cyclonic systems have large southward tilt.

The third group has been due to the joining of cluster nuclei 9 and 10, which have moderately low scores on PC II and moderately high scores on PC I, III and IV. The fourth cluster resulted due to joining of cluster nuclei numbers 11 and 12. This group has low score on PC II and moderately high scores on PC III and IV. These clustering can also be interpreted in the similar fashion.

The fifth group has formed due to joining of cluster nuclei numbers 6 and 7. The group has low scores on PC III and I and high score on PC IV. This combination could have the influence of weak monsoon situations, when these two regions receive above normal rainfall.

The last group has formed by joining of cluster nuclei numbers 8 and 13. The group has high score on PC I and low score on PC IV. This clustering also can be interpreted with the help of the features of the PCs.

The well separated six groups that are obtained after two stages of clustering by the flexible strategy, as discussed above and presented in Fig. 4, appears to be the best among all the clustering obtained after application of various strategies, even though a large amount of space dilation has occurred in this method.

6. Conclusion

It is possible to arrive at a reasonable broad-scale precipitation regionalization employing HCA.

The scrutiny of the dendrograms obtained by various fusion strategies reveals the presence of 13 nuclei clusters. The difference in the final grouping in various fusion methods arises mainly due to joining of these nuclei clusters at larger distances.

The formation of these nuclei is largely interpretable by utilizing the average PC scores and the synoptic features that have been associated with each P.C., as described in earlier studies by Mazumdar (1998b) and De and Mazumdar (1999).

A flexible strategy is found to produce reasonable, interpretable and well separated grouping. The meteorological subdivisions are found to be evenly distributed in each group. As per this, the Indian region could be divided into six coherent zones. The broad scale synoptic features that control the rainfall in these zones, in medium range scale can also be identified which will give more homogenous and useful for specific applications.

Acknowledgements

The author is grateful to Dr. U. S. De, Retd. Additional Director of Meteorology (Research), India Meteorological Department, for his guidance and thankful

to Mrs. B. S. Sabade and Mr. A. V. Pawar for their assistance in the preparation of the manuscript of the paper.

References

Anyadike, R.N.C., 1987, "A multivarite classification and regionalization of West African climates", *J. Climatol*., **7**, 157-164.

Cattel, R. B., 1966, "The scree test for the number of factors", *Mult. Behav. Res*., **1**, 245-276.

Cox, D. P., 1957, "Note on grouping", *J. Am. Stat. Assoc*., **52**, 543-547.

Craddock, J. M. and Flood, C. R., 1969, "Eigenevectors for representing the 500 mb geopotential surface over the Northern Hemisphere", *Quart. J. R. Met. Soc*., **95**, 576-593.

De, U. S. and Mazumdar, A. B., 1999, "Principal components analysis of rainfall and associated synoptic models of the southwest monsoon over India", *Theor. Appl. Climatol*., **64**, 213-228.

Drosdowsky, W., 1993, "An analysis of Australian seasonal rainfall anomalies: 1950-1987 I: Spatial Patterns", *Int. J. Climatol*., **13**, 1-30.

Ehrendorfer, M., 1987, "Regionalization of Austria's precipittion climate using principal component analysis", *J. Climatol*., **7**, 71-89.

Eklundh, L. and Pilesjo, P., 1990, "Regionalization and spatial estimation of Ethiopian mean annual rainfall", *Int. J. Climatol*., **10**, 473-494.

Everitt, B., 1980, "Cluster Analysis", 2nd Edn., Halsted Heinemann, London, p136.

Gadgil, S. and Iyengar, R. N., 1980, "Cluster Analysis of rainfall stations of the Indian Peninsula", *Quart. J. R. Met. Soc*., **106**, 873-886.

Gadgil, S. and Joshi, N. V., 1984, "Climatic clusters of the Indian region", *J. Climatol*., **3**, 47-63.

Iyengar, R. N., 1991, "Application of principal component analysis to understand variability of rainfall", Proc. Of Indian Acad. of Sci. (Earth and Planetary Sci.), **100-2**, 105-126.

Kaiser, H. F., 1958, "The varimax criterion for analytic rotation in factor analysis", *Psychometrika*, **23**, 187-200.

Kulkarni, B. S. and Reddy, D. D., 1994, "The cluster analysis approach for classification of Andhra Pradesh on the basis of rainfall", *Mausam*, **45**, 325-332.

Mather, P. M., 1976, "Computational methods of multivariate analysis in Physical Geography", John Wiley & Sons, N.Y., p524.

Mazumdar, A. B., 1995, "A study of some circulation features and intraseasonal variation of southwest monsoon over India, Ph.D.Thesis, Banaras Hindu University, Varanasi, p152.

Mazumdar, A. B., 1998a, "Southwest monsoon rainfall in India: Part I – Spatial variability", *Mausam*, **49**, 71-78.

Mazumdar, A. B., 1998b, "Southwest monsoon rainfall in India: Part II – Principal components in temporal domain", *Mausam*, **49**, 301-308.

Moore, A. W., Russell, J. S. and Ward, W. T., 1972, "Numerical analysis of soils; a comparison of three soil profile models with field classification", *J. Soil Sci*., **23**, 193-209.

Morgan, R. P. C., 1971, "Rainfall of Western Malaysia– a preliminary regionalisation using principal components analysis", *Area*, **3**, 222-227.

Murata, A., 1990, "Regionality and periodicity observed in rainfall variations of the Baiu Season over Japan", *Int. J. Climatol*., **10**, 627-646.

North, G. E., Bell, T. A., Cahalan, R. F. and Moeng, F. J., 1982, "Sampling errors in the estimation of empirical orthogonal functions", *Mon. Wea. Rev*., **110**, 699-706.

Ogallo, L. J., 1989, "The spatial and temporal patterns of the East African seasonal rainfall derived from principal component analysis", *Int. J. Climatol*., **9**, 145-167.

Opoku-Ankomah, Y. and Cordery, I., 1993, "Temporal variation of relations between New South Wales rainfall and the southern oscillations", *Int. J. Climatol*., **13**, 51-64.

Pandzic, K., 1988, "Principal component analysis of precipitation in the Adriatic – Pannonian area of Yugoslavia", *Int. J. Climatol*., **8**, 357-370.

Prasad, K. D. and Singh, S. V., 1988, "Large-scale features of the Indian summer monsoon rainfall and their association with some oceanic and atmospheric variables", *Adv. Atmos. Sci*., **5**, 499-513.

Richman, M. B., 1986, "Rotation of principal components", *J. Climatol*., **6**, 293-335.

Shukla, J., 1987, "Interannual variability of monsoon", Monsoons, J. S. Fein and P. L. Stephens (eds.), Wiley, NY., 399-464.

Sikka, D. R. and Gadgil, S., 1980, "On the maximum cloud zone and the ITCZ over Indian longitudes during the southwest monsoon", *Mon. Wea. Rev*., **108**, 1840-1853.

Singleton, F. and Spackman, E. A., 1984, "Climatological network design", *Met., Mag*., **113**, 77-89.

Spackman, E. A. and Singleton, F., 1982, "Recent developments in the quality control of climatological data", *Met. Mag*., **111**, 301-311.

Stone, R. C., 1989, "Weather types at Brisbane, Queensland: an example of the use of principal components and cluster analysis", *Int. J. Climatol*., **9**, 3-32.

Sumner G., Ramis, C. and Guijarro, J. A., 1993, "The spatial organization of daily rainfall over Mallorca, Spain", *Int. J. Climatol*., **13**, 89-109.

Webster, R., 1975, "Intuition and rational choice in the application of mathematics to soil systematics", *Soil Sci*., **110**, 394-404.

Webster, R., 1977, "Quantitative and numerical methods in soil classification and Survey", Clarendon, Oxford, p269.

Williams, W. T., 1971, "Principles of clustering", *Ann. Rev. Ecol. Syst*., **2**, 303-326.

Williams, W. T., 1976, "Pattern analysis in Agricultural science", CSIRO and Elsevier, Melbourne.

Willmott, C. J., 1978, "P-mode principal components analysis, grouping and precipitation regions in California", *Arch. Meteorol. Geophys. Bioklimatol*., Ser B., 26, p277.