Classification of Indian meteorological stations using cluster and fuzzy cluster analysis, and Kohonen artificial neural networks

K. Srinivasa Raju[1] and D. Nagesh Kumar[2]

[1]Department of Civil Engineering, Birla Institute of Technology and Science, Pilani 333 031, India
[2]Department of Civil Engineering, Indian Institute of Science, Bangalore 560 012, India

E-mail: *nagesh@civil.iisc.ernet.in*

Received 15 September 2005; accepted in revised form 5 January 2007

Abstract The present study deals with the application of cluster analysis, Fuzzy Cluster Analysis (FCA) and Kohonen Artificial Neural Networks (KANN) methods for classification of 159 meteorological stations in India into meteorologically homogeneous groups. Eight parameters, namely latitude, longitude, elevation, average temperature, humidity, wind speed, sunshine hours and solar radiation, are considered as the classification criteria for grouping. The optimal number of groups is determined as 14 based on the Davies–Bouldin index approach. It is observed that the FCA approach performed better than the other two methodologies for the present study.

Keywords Cluster analysis; Davies–Bouldin index; fuzzy cluster analysis; India; Kohonen artificial neural networks; meteorological stations

Nordic Hydrology **Vol 38 No 3 pp 303–314** © IWA Publishing 2007

Introduction

Classification of huge databases is useful in system modeling situations where a large number of datasets is reduced to a small number of groups. This is essential to produce a concise representation of a system's behavior. Such a classification can be performed using clustering algorithms (Zopounidis and Doumpos 2002; Ma 2004). In the perspective of water resources systems, classification of meteorological stations into hydrologically homogeneous groups for regionlisation purposes will be useful to evolve a different scale of measure suitable to each group. It will also be useful for prediction of various events such as floods and droughts and in the study of variability of rainfall over long time scales. Such classification may also facilitate the choice of strategies appropriate for each group in respect of agricultural practices and planning for utilisation of water resources of various regions (Gadgil and Iyengar 1980). McDonnell and Woods (2004) in their editorial discussed the necessity of catchment classification. They mentioned that classification would enable us to sort and group the considerable variability in space, time and process, which is present in natural hydrological systems across the globe. Rao and Srinivas (2006*a*,*b*) mentioned the necessity of systematic classification of watersheds for hydrological homogeneity as conventional regions are based on geographical, political, administrative or physiographic boundaries, and such delineated regions do not guarantee hydrological homogeneity.

It is difficult to manually classify datasets/watersheds/catchments/meteorological stations into groups for regionalisation purpose. There is also a threshold beyond which the difference between any two datasets is imperceptible to manual capabilities, compared to the machine. The method of clustering can be used to reduce the number of datasets to a more

doi: 10.2166/nh.2007.013

manageable subset (Morse 1980). Clustering analysis offers several advantages over a manual grouping process such as (1) the clustering program can apply a specified objective function criterion consistently to form the groups, avoiding the inconsistency due to human error and (2) the clustering algorithm can form the groups in a small fraction of the time that is required for manual grouping, particularly if a long list of criteria is associated with each data set (Jain and Dubes 1988).

Numerous authors used various types of classification methods. Jingyi and Hall (2004) applied a geographical approach (Residuals method), Ward's cluster method, fuzzy c-means method and Kohonen neural network to 86 sites in the Gan River Basin of Jiangxi Province and the Ming River Basin of Fujian Province in the southeast of China to delineate homogeneous regions based on site characteristics. It was concluded that the Kohonen methodology is the preferred approach. Rao and Srinivas (2006*a*) studied the applicability of three hybrid-clustering algorithms, which use a partitional clustering procedure, to identify groups of similar catchments in Indiana, USA. The hierarchical clustering algorithms used were single linkage, complete linkage and Ward's algorithms, while the partitional clustering algorithm used was the K-means algorithm. They also employed four cluster validity indices to determine their effectiveness in identifying optimal partitions provided by the clustering algorithms. Rao and Srinivas (2006*b*) applied fuzzy cluster analysis (FCA) to the above case study. They discussed the effectiveness of several fuzzy cluster validation measures in determining optimal partitions provided by the FCA. Similar studies were reported by Burn and Boorman (1993), Cunderlik and Ouarda (2006) and Lin and Chen (2006). In the present study, keeping the regionalisation in view, the practical applicability of three classification methods, namely fuzzy cluster analysis (FCA), Kohonen artificial neural networks (KANN) and cluster analysis (CA) methodologies, is explored for grouping 159 meteorological stations in India into meteorologically homogeneous groups.

Methodology

Fuzzy cluster analysis

Fuzzy cluster analysis is a clustering method in a fuzzy environment wherein each data set belongs to a cluster to some degree, specified by a membership grade. The algorithm is based on minimizing an objective function that represents the distance from any given data set to a cluster center, weighted by that data set's membership grade. In other words, the objective of the methodology is to represent the similarity a point shares with each cluster with a membership function, whose value lies between zero and one, and each sample has a membership in every cluster (Ross 1995) but degree of membership may vary from cluster to cluster (between zero to one). The sum of the membership values for each data set will be equal to 1. A brief methodology of FCA is given below:

1. Normalize the data.
2. Choose the number of clusters c ($2 \leq c \leq n$, where n is the number of data sets), number of iterations and termination criteria.
3. Formulate initial fuzzy partition matrix.
4. Compute cluster centers for each iteration and ascertain thereby Euclidean distance between the dataset and the cluster centers.
5. Update the fuzzy partition matrix for each iteration.

If the fuzzy partition matrix between two successive iterations is less than the specified termination criteria, the algorithm stops. Otherwise steps 2–5 are to be repeated until the above requirement is satisfied. A more detailed description of FCA is available in Ross (1995), Jingyi and Hall (2004) and Rao and Srinivas (2006*b*).

Kohonen artificial neural networks

The neural network based unsupervised classification algorithm, namely Kohonen artificial neural networks, consists of competitive layers that use the Kohonen learning rule to classify inputs. The neurons of the competitive layer learn to recognise groups of similar input vectors. It is a self-organising mapping technique with only two layers, input and output. Each layer is made up of neurons. The number of neurons in input layer, M, is equal to the dimensionality of the input vectors and the number of neurons in the output layer, N, is determined by the number of groups into which the input data will be partitioned. Each neuron in the output is interconnected with all those in the input layer by a set of weights or a weight vector, e.g. the jth output neuron has a weight vector connecting to input neurons, $w_j = \{w_{ji}\}$, $i = 1, 2, \ldots M$. The function of an input neuron is to transmit input data to the next layer, whereas an output neuron calculates the Euclidean distance between its weight vector w_j and input vector X' to measure their similarity. The main objective of the Kohonen network is to transform an incoming vector with arbitrary dimension into a one- or two-dimensional discrete map, and to perform this transformation adaptively in a topologically ordered fashion (Kohonen 1989; Liong *et al.* 2004; Raju *et al.* 2006). Important input parameters in KANN are learning rate, conscience rate and number of epochs.

Cluster analysis

The cluster analysis partitions data sets into relatively homogeneous groups. In clustering, datasets in a cluster are ensured to be more similar to each other than those in the other clusters. The K-means clustering algorithm (Jain and Dubes 1988) is used to minimise within cluster sums of squares of differences to obtain the final partitions. In this method, data sets are grouped so that each data set is assigned to one of the fixed number K of groups. The sum of the squared differences of each criterion from its assigned cluster mean is used as the objective function. Data sets are transferred from one cluster to another, so that, within a cluster, the sum of squared differences decreases. In a pass through the entire dataset, if no transfer occurs, the algorithm stops. The total square error value E_K for cluster group K is given by

$$E_K = \sum_{k=1}^{K} e_k^2 \tag{1}$$

where e_k = error value for each cluster group k.

Case study

The study area consists of 159 meteorological stations spread across India. Locational parameters (latitude, longitude and elevation) and observed variables (average temperature, humidity, wind speed, sunshine and solar radiation) were collected from the FAO website (http://www.fao.org./landandwater/aglw/cropwat.stm). These observed values were averaged over a year (to get the average of the 12 monthly values) to obtain a single value for each parameter for each station. These eight parameters were used as classification criteria.

Results and discussion

Data normalisation

Datasets were normalised for all the 8 criteria for the 159 meteorological stations to make the data dimensionless. Normalisation of criterion j for meteorological station i was defined as

$$y_{ij} = \frac{x_{ij} - x_{j\,\min}}{x_{j\ \max} - x_{j\,\min}} \tag{2}$$

where x_{ij} is the jth criterion for the ith meteorological station and $x_{j\ \max}$ and $x_{j\ \min}$ are maximum and minimum values of the jth criterion among the 159 meteorological stations. Normalised values thus obtained were then used for classifying the meteorological stations.

Software

Fuzzy cluster analysis algorithm was coded in the MATLAB environment using the function fcm. The input to this system was an Excel file containing two sheets of data. The first sheet contains the number of meteorological stations and criteria. The second sheet contains the normalised data corresponding to all eight parameters for the 159 meteorological stations. The other inputs include minimum and maximum number of clusters to find the optimum number, number of iterations and tolerance value. Output data were stored in different files (based on the number of cluster groups chosen) which already contain input information such as normalised values of the eight parameters for each meteorological station, membership grade of each meteorological station and the representative meteorological station for each group. The MATLAB based solution methodology was employed for KANN also (http://www.mathworks.com/access/helpdesk/help/toolbox/nnet).

Determination of optimal number of clusters

An effort was made to ascertain the optimal number of groups for the present problem. The fuzzy cluster analysis software was run for 3 to 20 clusters (total 18 in number) with 1 000 iterations and a tolerance value of 10^{-6}. The stopping criterion was set as the difference of the current fuzzy objective function value from the value in the previous iteration which is to be less than the tolerance value. The squared error values of each number of groups ranging from 3 to 20 are shown in Figure 1. It can be observed from this figure that the squared error value is decreasing from 28.1 to 14.1 and that there is a sharp decrease of error from cluster 3 to 4 (from 28.1 to 24.7).

The Davies–Bouldin index (Davies and Bouldin 1979) was used as the basis to find the optimum number of clusters for grouping the meteorological stations. The index is defined as

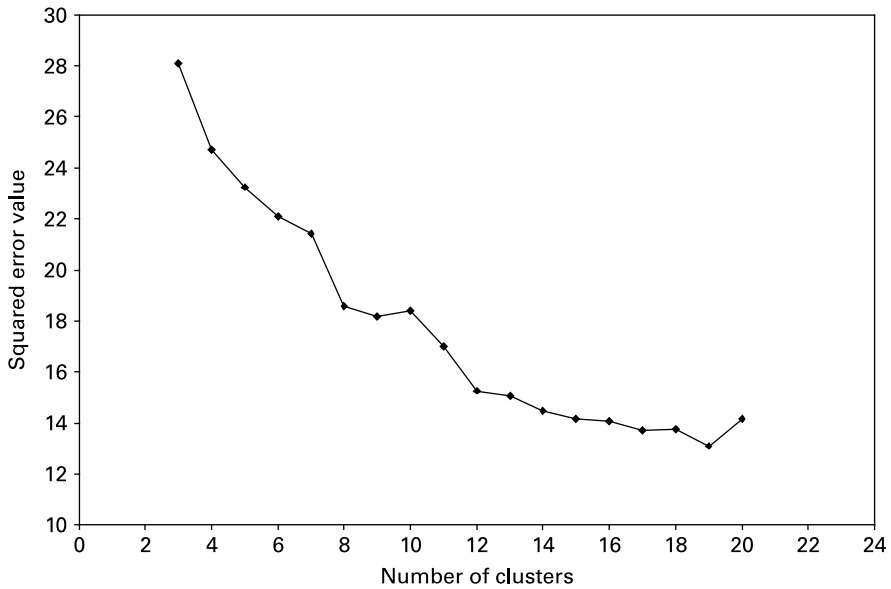

Figure 1 Squared error values for different number of clusters

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c} \max\left[\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)}\right] \quad i \neq j \tag{3}$$

where $\delta(X_i, X_j)$ defines the intercluster distance between clusters X_i and X_j; $\Delta(X_i)$ represents intracluster distance (diameter) of cluster X_i and c is the number of clusters of partition U. In this case, a small index value represents good clusters, i.e. clusters are compact and their centers are not far away from each other. Davies–Bouldin validation index values were computed for 3 to 20 clusters and presented in Figure 2. A minimum value of Davies–Bouldin index is desirable (Jain and Dubes 1988). It can be observed from Figure 2 that the Davies–Bouldin index for 3 clusters is 3.64 and for 20 clusters it is 2.69 with fluctuations of values in between. The minimum value of Davies–Bouldin index is 2.26, which occurs for 14 clusters. It was therefore inferred that the optimum number of clusters is 14 and the same was adopted for further analysis.

Application of FCA

The membership value in each group indicates the probability for the meteorological station to be clustered in that specific group. An extract of the membership values of each of the 159 meteorological stations under each of the 14 groups is presented in Table 1. The group which has the highest membership value among the 14 groups is the representative group for that meteorological station. For meteorological station 1, membership values for the 14 groups are 0.1867, 0.0204, 0.2839, 0.0847, 0.0425, 0.0150, 0.098, 0.0414, 0.0124, 0.0252, 0.1141, 0.0134, 0.0368 and 0.0255. The sum of these values should be equal to 1 (Ross 1995). The representative group for the meteorological station no. 1 is group 3 (having the maximum membership value of 0.2839). Similarly all other meteorological stations were analysed and grouped. Numbers of meteorological stations falling in cluster groups 1–14 are 12, 11, 14, 14, 10, 12, 10, 14, 10, 8, 9, 9, 12, and 14, respectively. The minimum number of stations per group is 8 and the maximum is 14 and the average number of stations per group is 11, which indicates a reasonable distribution. It is observed that none of the 14 groups is empty. This may be due to the advantage of FCA which allows each data set to have partial membership

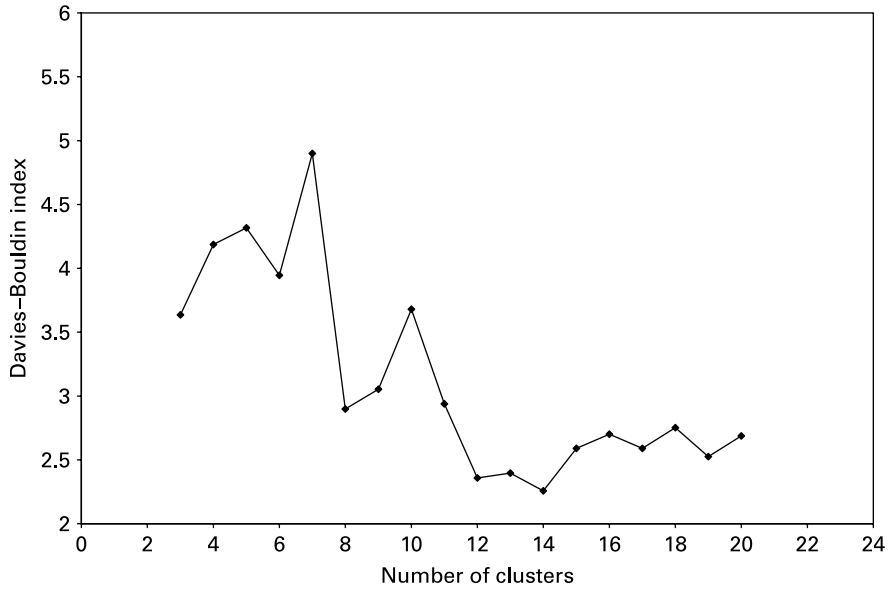

Figure 2 Davies–Bouldin index values for different numbers of clusters

Table 1 Membership values of the meteorological stations under each group showing the representative group of each station

Met. station	Membership values for group number														Representative group
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
1	0.1867	0.0204	0.2839	0.0847	0.0425	0.0150	0.0980	0.0414	0.0124	0.0252	0.1141	0.0134	0.0368	0.0255	3
2	0.1972	0.0353	0.0706	0.0426	0.0250	0.0234	0.1613	0.1771	0.0191	0.0566	0.0563	0.0091	0.1094	0.0170	1
3	0.1288	0.0608	0.0594	0.0465	0.0283	0.0343	0.1170	0.1474	0.0260	0.0967	0.0544	0.0113	0.1685	0.0206	13
4	0.2939	0.0270	0.0883	0.0506	0.0261	0.0180	0.1938	0.0981	0.0143	0.0379	0.0599	0.0103	0.0636	0.0182	1
5	0.1309	0.0339	0.0512	0.0519	0.0265	0.0186	0.2648	0.1518	0.0147	0.0555	0.0546	0.0078	0.1232	0.0146	7
6	0.0479	0.1105	0.0412	0.0378	0.0332	0.1785	0.0480	0.0743	0.1033	0.1311	0.0476	0.0161	0.0993	0.0311	6
7	0.3841	0.0164	0.2006	0.0423	0.0239	0.0133	0.0887	0.0445	0.0092	0.0224	0.0924	0.0082	0.0364	0.0175	1
8	0.0370	0.0108	0.0849	0.0595	0.0431	0.0095	0.0294	0.0162	0.0062	0.0133	0.6451	0.0067	0.0178	0.0205	11
	...	...	...	...	...	...		...	...	...	...	...	...	...	...
	...	...	...	...	...	...		...	...	...	...	...	...	...	...
155	0.0424	0.0216	0.0870	0.4358	0.1177	0.0138	0.0403	0.0209	0.0129	0.0215	0.1089	0.0166	0.0248	0.0359	4
156	0.0422	0.0115	0.1048	0.0467	0.0412	0.0109	0.0285	0.0170	0.0067	0.0141	0.6250	0.0077	0.0192	0.0247	11
157	0.0461	0.1614	0.0396	0.0483	0.0442	0.0853	0.0503	0.0611	0.0773	0.1655	0.0513	0.0151	0.1255	0.0289	10
158	0.0548	0.1013	0.0472	0.0482	0.0420	0.1305	0.0612	0.0908	0.0992	0.1140	0.0543	0.0231	0.0941	0.0392	6
159	0.0294	0.0705	0.0330	0.0367	0.0489	0.3719	0.0293	0.0337	0.1041	0.0607	0.0455	0.0294	0.0462	0.0608	6

in all clusters. It is relevant to note that the distribution of the stations among the 14 groups with the FCA approach is fairly even. For a country like India with largely varying hydro-meteorological features, the clustering into 14 distinct groups is quite appropriate, representing the various types of regions of the country.

The station with the highest membership value in a group is the representative station for that group. The representative meteorological stations for groups 1–14 are 65 (Guna), 119 (Mysore), 105 (Mainpuri), 128 (Pendra), 142 (Sambalpur), 86 (Kakinada), 27 (Bhopal), 78 (Jalgaon), 94 (Kozhikode), 57 (Gadag), 8 (Allahabad), 100 (Lumbding), 96 (Kurnool) and 106 (Malda) with the highest membership values of 0.6607, 0.5094, 0.6492, 0.7882, 0.6506, 0.5515, 0.8203, 0.5861, 0.8424, 0.5358, 0.6451, 0.9277, 0.2631 and 0.7251, respectively. A geographic map showing the location of all 159 meteorological stations is presented in Figure 4. Figure 5 shows the group number to which each station belongs obtained using FCA.

The representative meteorological stations for each group are also found by the squared error methodology. For this purpose the squared error values between the group mean and the value for each criterion for each meteorological station in that group are calculated. The sum of these squared error values for all criteria gives the total squared error value corresponding to each meteorological station in that group. For example, for group number 1, among the 12 meteorological stations in that group, the minimum squared error value of 0.00841 occurred for meteorological station 83 (Jhalawar). On the same basis, meteorological stations 68 (Hassan), 20 (Bareilly), 125 (Pachmarhi), 142 (Sambalpur), 86 (Kakinada), 27 (Bhopal), 78 (Jalgaon), 94 (Kozhikode), 141 (Salem), 156 (Varanasi), 52 (Dibrugarh), 30 (Bijapur) and 106 (Malda) represent the group numbers 2–14, with minimum squared error values of 0.008 36, 0.002 73, 0.034 28, 0.009 12, 0.040 53, 0.0044, 0.024 02, 0.006 79, 0.027 24, 0.002 88, 0.061 62, 0.0166 and 0.007 77, respectively. On comparison of the above two analyses, it may be noticed that the meteorological stations Sambalpur, Kakinada, Bhopal, Jalgaon, Kozhikode and Malda are common in both approaches using the membership values and the squared error values.

Application of KANN

The schematic diagram of KANN for the present classification problem is presented in Figure 3. In KANN, the input layer consists of eight criteria, namely latitude, longitude, elevation, average temperature, humidity, wind speed, sunshine hours and solar radiation and the output layer

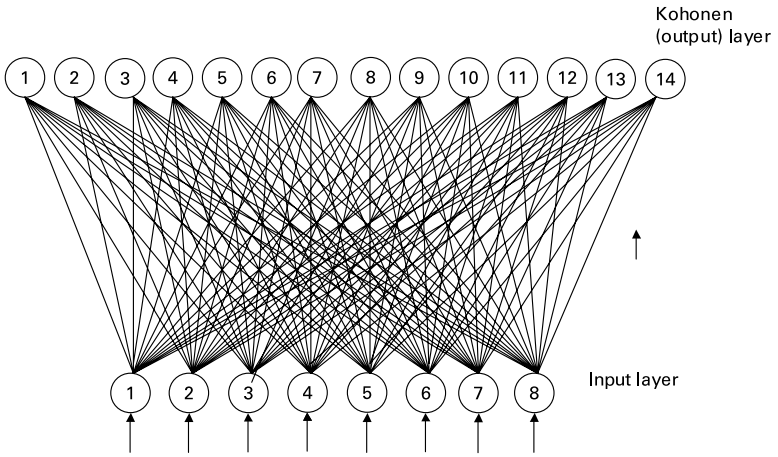

Figure 3 Schematic diagram of Kohonen artificial neural network for the clustering problem. Inputs 1–8 correspond to latitude, longitude, elevation, temperature, humidity, wind speed, sunshine and radiation (see text)

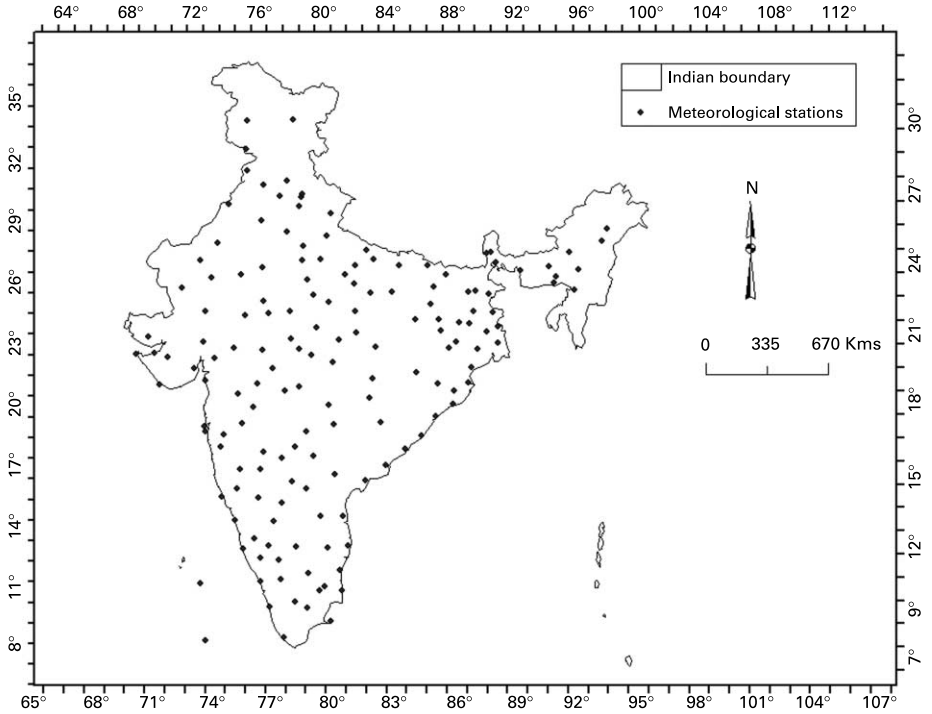

Figure 4 Location map of the 159 meteorological stations considered in the study

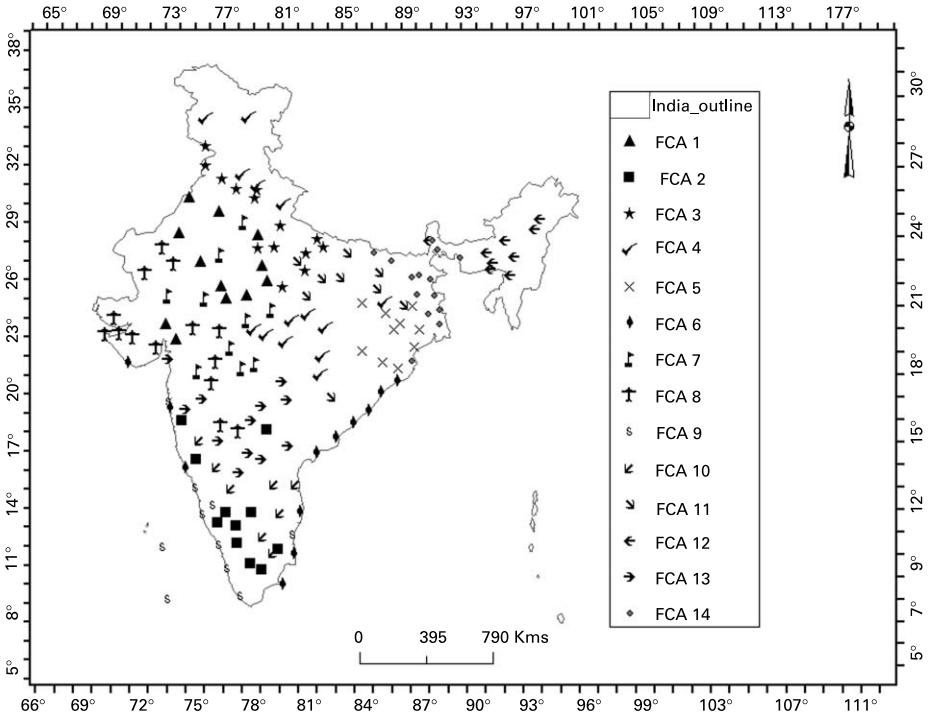

Figure 5 Map showing the group number to which each meteorological station belongs as obtained by fuzzy cluster analysis (FCA)

consists of the number of the group (1–14). The parameters used for training the algorithm are the number of groups as 14 (optimal number of clusters obtained from the Davies–Bouldin approach), learning rate 0.01, conscience rate 0.001 and number of epochs 1 000 (more information about input parameters for KANN is available at the web site http://www.mathworks.com/products/neuralnet/index.html). Out of the targeted 14 groups, all meteorological stations are bundled up into only 5 groups (numbers 2, 4, 8, 11 and 14), leaving the balance of 9 groups empty. The numbers of meteorological stations falling into these 5 groups are 38, 6, 77, 26 and 12, respectively. It may be noticed that this distribution of stations amongst the five groups is very uneven. The representative meteorological stations for these groups are 119 (Mysore), 118 (Mussorie), 78 (Jalgaon), 143 (Satna) and 106 (Malda), respectively. Thirty-six meteorological stations are commonly identified in the same groups by both FCA and KANN, working out to 23% of the 159 meteorological stations considered.

In KANN, the learning rate, for a given conscience rate and the number of epochs, plays a major role. It can be observed from Table 2 that for various learning rates (0.01, 0.1 to 0.9), the number of groups formed is less than the targeted 14. By increasing the learning rate to 0.50, the number of groups increased to 11. On further increasing the learning rate to 0.70, the number of groups increased to 13, which is very near to the 14 with the FCA approach. But it can be observed from Table 2 that the distribution of stations into 11 or 13 groups is very uneven, rendering 2 out of 11 and 4 out of 13 groups practically inactive. Against this, the distribution of stations into 14 groups with the FCA approach is fairly even. Thus in KANN with any value of the learning rate, some of the groups are empty, thus performing inferior to FCA in which no group was empty. Even though in some cases the number of groups is the same (5 groups for learning rates 0.01 and 0.1; 8 for learning rates 0.3 and 0.4; 11 for learning rates 0.5 and 0.9), the number of meteorological stations in each group is different. Accordingly the number of common meteorological stations will be different from those with FCA. It is also observed that, for learning rates 0.5, 0.7 and 0.9, there are some groups in which only one meteorological station is present. It is also observed that the effect of learning rate is significant on the squared error value but no fixed trend is observed. Extensive sensitivity analysis on various learning rates, conscience rates and number of epochs indicated that a careful selection of parameters is of utmost importance for obtaining meaningful results.

Application of CA

On adopting CA methodology, the numbers of meteorological stations falling in group numbers 1–14 are 18, 1, 7, 2, 9, 21, 1, 18, 26, 18, 10, 22, 2 and 4. It is observed that 2 groups have only one station each, 2 clusters have two stations each and another has only 4 stations. Thus it is observed that out of 14 groups, only 9 are active containing 149 stations (excluding the above 5 groups with 10 stations). The number of stations in each group with this approach is compared with those with the FCA approach in Table 3. It can be observed from this table that the distribution of stations amongst the groups is very much uneven with the CA approach, whereas the same is fairly even with the FCA approach. This indicates that the CA approach does not have the flexibility and advantage which the FCA has (Rao and Srinivas 2006*b*). Out of 159 stations, 83 meteorological stations are common in FCA and CA methodologies, i.e. 52%, and this is 45% in the case of CA and KANN. The representative meteorological stations for the 14 groups by the CA methodology are 128 (Pendra), 92 (Kodaikanal), 59 (Guwahati), 97 and 149 (Leh and Shimla), 68 (Hassan), 78 (Jalgaon), 117 (Mukteshwar), 106 (Malda), 83 (Jhalwar), 43 (Cuddalore), 86 (Kakinada), 20 (Bareilly), 118 and 150 (Mussorie and Srinagar) and 145 (Shillong). It may be noted that groups 4 and 12 have 2 representative stations each, viz. 97 and 149, and 118 and 150. This is due to their equal distance from the group mean.

Table 2 Number of stations under each group with different learning rates (Conscience rate = 0.001; number of epochs = 1000)

			Number of stations under each group in group number														
No.	**Learning rate**	**Squared error**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**Total**	**Number of groups**
1	0.01	32.9582		38		6				77			26			12	159	5
2	0.10	28.4065	53								46		4	25		31	159	5
3	0.20	28.1834		11		86	6			6			29			21	159	6
4	0.30	26.4128			27		13	6	3	4	29		21	56			159	8
5	0.40	36.3313			16	28		1		9	39			2	22	42	159	8
6	0.50	26.9944	31	18	4	13	9	1		16		27		18	9	13	159	11
7	0.60	36.9930		3	4		3	97			8	10	34				159	7
8	0.70	28.7811	5	5	32	9	42	1	1	17		31	6	6	3	1	159	13
9	0.80	35.9683		3	15	18	23			77		11	6	2	4		159	9
10	0.90	37.1077	28	6	3		1	4	39	38	16	1		2	21		159	11

Table 3 Number of stations in each group with FCA and CA approaches

	Group number													
Approach	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
FCA	12	11	14	14	10	12	10	14	10	8	9	9	12	14
CA	18	1	7	2	9	21	1	18	26	18	10	22	2	4

Sensitivity analysis

It may be noted that the percentage of common stations of 23% between FCA and KANN is out of 14 groups for FCA and 5 groups for KANN. Similarly the percentage of common stations of 52% between FCA and CA is out of 14 groups for FCA and 9 active groups for CA. Also, the percentage of common stations of 45% between CA and KANN is out of 9 active groups for CA and 5 groups for KANN.

Efforts are then made to group the stations into 5 clusters using all three methodologies (based on the minimum number of clusters obtained from KANN analysis). In this case, the number of common meteorological stations is found to be 106 out of 159, i.e. 67% in the case of FCA and CA; 94 (59%) in the case of FCA and KANN and 113 (71%) in the case of CA and KANN. It may be inferred from the above analysis that, as the number of groups decreases, the percentage of common meteorological stations increases. It can also be noticed that FCA is performing better as meteorological stations are occupying all the groups and the distribution of the stations amongst the groups is fairly even which will provide meaningful assessment for the purpose of regionalisation.

Potential of the proposed methodology

The above FCA methodology evolved and, integrated with the Davies–Bouldin index, can be used in many practical situations. The classification of an irrigated area into meteorologically homogeneous groups will greatly facilitate planning suitable irrigation policies to suit each particular condition of the group (Gadgil and Iyengar 1980). In the case of flood frequency studies, classification of the catchment area of a river basin into meteorologically homogeneous groups will facilitate segregation of the different frequencies for planning appropriate precautionary and remedial measures for each range of frequency (Jingyi and Hall 2004). In the case of analyzing droughts it would be necessary to adopt different scales of remedial measures to suit the different intensities of drought conditions in different parts of the region. Another important utility of the clustering technique is that, in the case of any missing data, which is not infrequent, the corresponding data of the representative station can be safely substituted, without any cognizable error.

Conclusions

Three classification methodologies, namely fuzzy cluster analysis, Kohonen artificial neural networks and cluster analysis, were employed to group 159 meteorological stations in India into meteorologically homogeneous groups. Eight parameters, namely latitude, longitude, elevation, average temperature, humidity, wind speed, sunshine hours and solar radiation, are used for the classification. Data were normalised for effective classification. The optimal number of groups is chosen, based on Davies–Bouldin index, as 14. The results of FCA, KANN and CA approaches are analysed and compared. The following inferences are drawn from the study:

1. The FCA methodology is suitable for the present planning problem as the stations are distributed more evenly, compared to CA and KANN, and also due to its advantage of assigning every station with partial membership in each group.

2. 23% of the meteorological stations are observed to be common to particular groups between FCA and KANN; 52% between FCA and CA; 45% in the case of CA and KANN. These are 59%, 67% and 71%, respectively, in case of the restricted 5 clusters. It is also observed that, as the number of specified groups decreases, the percentage of common meteorological stations increases.
3. The effect of learning rate on squared error and the number of groups formed is significant, indicating that a careful selection of parameters is of utmost importance in the case of the KANN approach.

The above results are based on an average of 12 months of normalised data and assumption of equal importance to all classification criteria. The study may be further refined adopting a monthly basis if sufficient data is available to get more representative results.

Acknowledgements

This work is partially supported by INCOH, Ministry of Water Resources, Govt. Of India, through project # 23/51/2006-R&D/203-14. The first author wishes to thank his colleague, Mr Amarendra Kumar Sandra, for introducing him to the concept of optimum number of clusters. The authors wish to thank Mrs. A. Anandhi, for help rendered in preparing Figures 4 and 5 using GIS.

References

Burn, D.H. and Boorman, D.B. (1993). Estimation of hydrological parameters at ungauged catchments. *J. Hydrol.*, **143**, 429–454.

Cunderlik, J.M. and Ouarda, T.B.M.J. (2006). Regional flood-duration-frequency modeling in the changing environment. *J. Hydrol.*, **318**, 276–291.

Davies, D.L. and Bouldin, D.W. (1979). A cluster separation measure. *IEEE Trans. Patt. Anal. Machine Intelligence*, **1**, 224–227.

Gadgil, S. and Iyengar, R.N. (1980). Cluster analysis of rainfall stations of the Indian peninsula. *Q.J.R. Meteorol. Soc.*, **106**, 873–886.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data*, Prentice-Hall, Englewood Cliffs, NJ.

Jingyi, Z. and Hall, M.J. (2004). Regional flood frequency analysis for the Gan-Ming river basin in China. *J. Hydrol.*, **296**, 98–117.

Kohonen, T. (1989). *Self Organization and Associative Memory*, Springer-Verlag, Berlin.

Lin, G.F. and Chen, L.H. (2006). Identification of homogeneous regions for regional frequency analysis using the self-organizing map. *J. Hydrol.*, **324**, 1–9.

Liong, S.Y., Tariq, A.A.F. and Lee, K.S. (2004). Application of evolutionary algorithm in reservoir operations. *J. Inst. Engrs (Singapore)*, **44**, 39–54.

Ma, Y. (2004). Fuzzy analysis on water resources of Heilongjiang state farms. *Nature Sci.*, **2**(1), 44–47.

McDonnell, J.J. and Woods, R. (2004). On the need for catchment classification. *J. Hydrol.*, **299**, 2–3.

Morse, J.N. (1980). Reducing the size of the nondominated set; pruning by clustering. *Comput. Oper. Res.*, **7**, 55–66.

Raju, K.S., Kumar, D.N. and Duckstein, L. (2006). Neural networks and multicriterion analysis for sustainable irrigation planning. *Comput. Oper. Res.*, **33**(4), 1138–1153.

Rao, A.R. and Srinivas, V.V. (2006a). Regionalization of watersheds by hybrid-cluster analysis. *J. Hydrol.*, **318**, 37–56.

Rao, A.R. and Srinivas, V.V. (2006b). Regionalization of watersheds by fuzzy cluster analysis. *J. Hydrol.*, **318**, 57–79.

Ross, T.J. (1995). *Fuzzy Logic with Engineering Applications*, McGraw-Hill, New York.

Zopounidis, C. and Doumpos, M. (2002). Multicriteria classification and sorting methods: a literature review. *Europ. J. Oper. Res.*, **138**(2), 229–246.