Journal of Spatial Hydrology
Vol.6, No.2 Fall 2006

Classification of Spatio -Temporal Pattern of Rainfall in Iran Using A Hierarchical and Divisive Cluster Analysis

Saeed Soltani[*] *and Reza Modarres*[#]

Abstract
The identification of spatial rainfall pattern is an essential task for hydrologists, climatologists as well as regional and local planners and managers. This is due to the variability of both the temporal spatial distribution of rainfall. In this study, a hierarchical and divisive cluster analysis was used to categorize these patterns of rainfall in Iran. According to results obtained, there are eight main spatial groups of annual rainfall over Iran. These groups can be classified into 3 main seasonal rainfall regimes namely, winter, winter-spring and fall regimes. The results also show that the elevation and sea neighborhood affect rainfall pattern of Iran. Moreover, a comparison between the Ward and average methods of hierarchical cluster analysis indicates that the Ward method performs the spatial pattern better.

Key Words: Rainfall pattern, Cluster Analysis, Ward Method, Iran

Introduction

Hydrologists have always been trying to classify hydrologic events in order to simplify hydrologic convolution and therefore to reduce massive body of information, observations and variables. The reason for this is to decrease time and save budget. Several methods have been, and currently, in use. Most of the methods are used to regionalize hydrologic phenomena such as rainfall, streamflow.

Multivariate techniques have been underlined as suitable and powerful tool to find hydrologically homogeneous region or to classify meteorological data such as rainfall. Principle component analysis, factor analysis and different cluster techniques have been used to classify daily rainfall patterns and their relationship to atmospheric circulation (Romero et al., 1999); to classify rainfall spatio-temporal pattern in Iran, to classify flood and drought years (Singh, 1999), to classify streamflow drought (Stahl and Demuth, 1999). The use of cluster analysis for regionalization involves grouping of various observations and variables into clusters, so that each cluster is composed of observations or variables with similar characteristics such as geographical, physical, statistical or stochastic behavior. Mosely (1981) used hierarchical cluster analysis on rivers in NewZealand and Tasker (1982) compared methods of defining homogeneous regions including cluster analysis with a complete linkage algorithm. Acerman (1985) and Acerman and Sinclair (1986) concluded that clustering has some intrinsic worth to explain the observed variation in data. Gottschalk (1985) applied cluster and principal component analysis to the territory of Sweden and found that cluster analysis is an appropriate method to use on a national scale with heterogeneous hydrological regimes. Nathan and McMahon (1990) performed hierarchical cluster analysis for the prediction of low flow characteristics in southeastern Australia. They found that Ward method with a similarity measure based on the squared Euclidean distance is the best method for cluster analysis.

*Assist. Professor, Isfahan university of Technology * Corresponding Author: Email: ssoltani@cc.iut.ac.ir
Isfahan university of Technology

The climate of Iran is characterized by complex pattern of spatial and temporal variability, with wide unpredictable rainfall fluctuations from year to year and region to region. Therefore, it is difficult to know the regional variation of rainfall. The identification of the rainfall temporal and regional variation requires long term series, which are not always available. The World Meteorological Organization (WMO) suggests using 30-year periods. However, when variations over time are considered, shorter periods such as 10 or 20 years may also be applied (Moron, 1997; Salinger and Mullan, 1999; Ramos, 2001). The aim of this study is to classify rainfall spatial time distribution pattern and also to evaluate annual and seasonal temporal pattern over Iran through cluster analysis.

Material and methods

Rainfall Data

The data set used in this study includes annual and seasonal rainfall of 28 capitals of their provinces of Iran as the candidates of rainfall spatial pattern over Iran. The variation of the mean rainfall of Iran is 224 mm to 300 mm which is widely distributed over Iran. The highest rainfall occurs in the north of Iran. The candidate city in this region is Rasht, which is the capital city of Guilan province. The spatial distribution of selected stations is presented in Fig. 1.

Fig. 1. Distribution of selected stations over Iran

Cluster Analysis

The purpose of cluster analysis is to place objects into groups so that the objects in each group have the highest similarity to each other while the objects in different clusters have the maximum dissimilarity. In cluster analysis, the multivariate data matrix **X** ($n \times p$) consists of observations obtained from the measurement of n subjects or objects with respect to p features or characteristics. The p columns of **X** are usually referred to as variables whereas the n rows are commonly called the profiles or pattern of the observational units. A profile is simply a vector of measurements whose elements are to be compared. Here, the profile are the n ($1 \times p$) vectors that constitute **X**. In this study, we have 28 columns of capital cities with p years record of data ranging from 15 to 32 years. A proximity matrix is an ($n \times n$) matrix that summarizes the degree of similarity and dissimilarity among all possible pairs of profiles in **X**. A commonly used dissimilarity measure is Euclidean distance which is written as follows:

$$d_{rs}^2 = \sum_{j=1}^{p} (x_{rj} - x_{sj})^2 \qquad (1)$$

where rth and sth rows of the data matrix **X** is denoted by (x_{r1}, x_{r2}, …,x_{rp}) and (x_{s1}, x_{s2}, …, x_{sp}), respectively. These tow rows correspond to the observations on two objects for all *P* variables. The quantity d_{rs}^2 will be refered to as the squared Euclidean distance (Jobson, 1992). The Euclidean distance of dissimilarity is then used in cluster analysis.

There are two main types of cluster techniques: Divisive and Hierarchical (Kaufman and Rousseuw, 1990).

Divisive cluster analysis is a common method, based on K-means algorithm, which measures the proximity between groups using Euclidean distance between group centeroids (Jobson, 1992). The K-means algorithm begins from a given partition specified by the assignment vector ***P*** (K, M) with the *M* observations allocated into *K* clusters. The centeroid of each of the *K* clusters is computed and the similarity among observations in the clusters is measured by an error function *e* given by:

$$e[P(M, K)] = \sum_{K} D[I, L(I)]^2 \qquad (2)$$

where *L(I)* is the cluster containing the *I*th case. $D[I, L(I)]^2$ represents the sum of the squared absolute deviations from the cluster centeroids (means of the variables) over all the observations in the cluster and over variables (Ramos, 2001).The initial assignment of the observation in *K* clusters is usually carried out by a random partion. The number of clusters to be kept is chosen by the minimum error function for an increasing number of clusters. Another suitable method to select the number of clusters is to calculate pseudo *F* and t^2 statistics (SAS,1999). The pseudo *F* statistics for a given level is:

$$\text{pseudo F} = \frac{\frac{T - P_G}{G - 1}}{\frac{P_G}{n - G}} \quad (3)$$

where T is $\sum_{i=1}^{n} \left\| x_i - \bar{x} \right\|^2$, n is the number of observations, $P_G = \sum W_J$, is the summation over *G* clusters at the *G*th level of the hierarchy. The pseudo t^2 statistics for joining C_K and C_L is

$$\text{pseudo } t^2 = \frac{B_{KL}}{\frac{W_K + W_L}{N_K + N_L - 2}} \quad (4)$$

where W_K is $\sum_{i \in C_K} \left\| x_i - \bar{x}_k \right\|^2$, W_L is $\sum_{i \in C_L} \left\| x_i - \bar{x}_L \right\|^2$, B_{KL} is $W_M - W_K - W_L$ if $C_M = C_K \cup C_L$, N_K and N_L are the number of observation in C_K and C_L , respectively.

There are several techniques in hierarchical cluster analysis such as Single, Average, Complete linkage and Ward's minimum variance method among which Average linkage and Ward methods are widely used (Milligan, 1980; Jackson and Weinand, 1995; Ramos, 2001).

If we have two clusters, C_K and C_L, to merge and produce another cluster, C_M, the distance between C_M and another cluster, C_J, is written as follows:

$$d_{J.M} = \frac{n_k d_{jk} + n_L d_{JM}}{n_M} \quad (5)$$

where n_L, n_K and n_M are the number of objects previously grouped together in clusters *L*, *K* and *M*, respectively, and d_{JK} and d_{JL} are distances between clusters *J* and *K* and between *J* and *L* clusters respectively.

The Ward's method calculates the distance between two clusters as the sum of squares between two clusters added up over all variables. At each cluster generation the sum of squares is minimized. If C_k and C_L are two clusters that merged to form cluster C_M, the distance between the new cluster and another cluster C_J is:

$$d_{J.M} = \frac{((n_J + n_K))d_{jk} + (n_J + n_L)d_{JL} - n_J d_{KL})}{n_J + n_M} \quad (6)$$

where n_J, n_K, n_L and n_M are the number of objects in clusters *J*, *K*, *L* and *M*, respectively, and d_{JK}, d_{JL} and d_{KL} represent the distances between the observations in clusters *J* and *K*, *J* and *L*, and *K* and *L*, respectively.

Results and Disscusions

The first step in statistical analysis is to investigate descriptive characteristics of the data. Descriptive analysis can help the investigators to have a preliminary judgment of the data and to decide for further analysis. The most important descriptive statistics are mean, standard deviation and coefficient of variation, (*Cv*), calculated as standard deviation divided by mean. In hydrology, however, there are two other important moments namely coefficient of skewness (*Cs*) as the measure of symmetry and coefficient of kurtosis (*Ck*) as the measure of shape of frequency function. The descriptive statistics of the annual rainfall are given in Table 1.

Table 1. Descriptive statistics of selected stations

Number	***Station Name***	***MEAN***	***STDEV***	***Cv***	***Cs***	***Ck***
1	*Ahwaz*	213.30	86.30	0.40	0.32	0.66
2	*Arak*	345.00	92.78	0.27	-0.28	-0.89
3	*Ardabil*	309.00	88.02	0.28	0.78	0.29
4	*Bandar Abbas*	192.00	121.80	0.63	0.75	0.12
5	*Bushehr*	275.60	118.81	0.43	1.41	3.38
6	*Ghaemshahr*	752.30	116.68	0.16	0.43	-0.59
7	*Gorgan*	612.10	102.83	0.17	0.02	-0.49
8	*Ghazvin*	315.90	89.48	0.28	0.56	-0.48
9	*Hamedan*	316.20	76.57	0.24	0.54	-0.53
10	*Isfahan*	121.40	40.10	0.33	-0.25	-0.81
11	*Ilam*	627.90	170.69	0.27	0.42	0.47
12	*Oroomieh*	349.30	98.40	0.28	0.95	0.36
13	*Ghom*	149.00	47.10	0.32	-0.08	-1.50
14	*Zahedan*	94.80	40.14	0.42	0.91	-0.23
15	*Zanjan*	317.60	72.60	0.23	-0.22	-0.12
16	*Yazd*	62.10	27.88	0.45	0.23	-1.17
17	*Yasuj*	822.90	183.02	0.22	0.05	-0.60
18	*Tehran*	229.20	63.92	0.28	0.20	-0.69
19	*Tabriz*	293.30	68.05	0.23	0.22	0.22
20	*Shiraz*	344.70	99.69	0.29	-0.21	-0.21
21	*Shahrecord*	319.00	86.57	0.27	0.24	0.24
22	*Semnan*	139.90	54.22	0.39	0.60	0.60
23	*Sanandaj*	471.00	118.78	0.25	-0.16	-0.16
24	*Rasht*	1353.00	279.35	0.21	0.56	0.56
25	*Mashhad*	257.50	77.41	0.30	0.27	-0.84
26	*Khoramabad*	515.10	125.61	0.24	-0.33	0.09
27	*Kermanshah*	450.80	120.40	0.27	0.10	-0.52
28	*Kerman*	158.90	50.18	0.32	0.35	0.23

STDEV: Standard Deviation: **Cv**: Coefficient of Variation;
Cs: Coefficent of skewness; **Ck**: Coefficient of Kurtosis

K-Means Algorithm

The first step in *K*-means cluster analysis is to find the suitable number of clusters based on error function (equation 2). To apply the least square error in this function we apply analysis of variance (ANOVA). Based on ANOVA for 2 to 13 clusters, we calculated error functions of equation 2. As shown in Fig. 2 the error functions decreases to 12 clusters and then increases in 13th cluster. There are also three other sharp drops for 3 and 6 and 8 clusters. The number of suitable cluster seems to be 3, 6 or 8 clusters in this case. These values show the possibility of the existence of 3, 6 or 8 regions based on rainfall regimes but we cannot determine the regions by *K*-means algorithm. We use hierarchical methods to find the suitable number of the clusters.

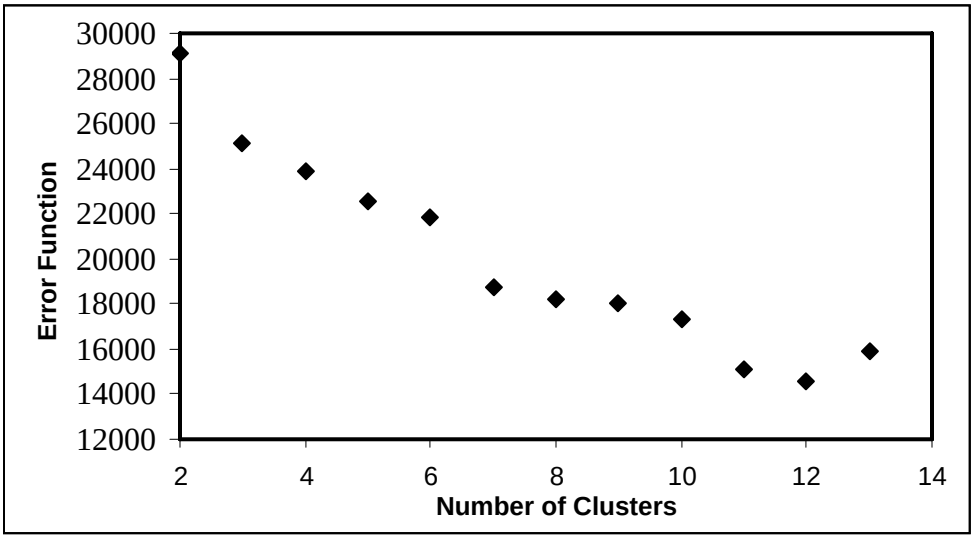

Fig. 2. Error function obtained with *K*-means algorithm for different number of clusters of annual series

Hierarchical Analysis

Based on Euclidean distance, two method of hierarchical clustering, average method and Ward method are used to classify annual rainfall into 12 similar clusters. Using equations 3 and 4, pseudo *F* and t^2 were calculated for 13 clusters (Fig. 3). In this figure, both pseudo *F* and t^2 have significant change at 3, 6 and 8 clusters which indicates potential number of clusters for classifications. To choose the proper number of clusters one can refer to the total spatial variance for each number of clusters. The R-Squared, squared multiple correlation or the decrease in the proportion of variance accounted for due to joining two clusters to form the current cluster (SAS/STAT, 1999), is the key to select the number of proper clusters. Fig. 4 shows R-squared against the number of clusters.

In this figure the coefficient of determination (R^2) increases with the number of clusters. The values of R^2 of two, three, six and eight clusters are 0.632, 0.804, 0.930 and 0.96, respectively. In this case, we accept 8 clusters which cover more than 95% of rainfall variance over Iran. The result of this analysis is usually shown in a graphical illustration called “Dendrogram” (not shown here). From this, we can find a group of arid and semi arid stations in the center of Iran including Yazd, Zahedan, Isfahan, Semnan, Kerman, Ghom,Tehran. However, the existence of Ghazvin and Mashhad station is suspicious which are located in high mountains.

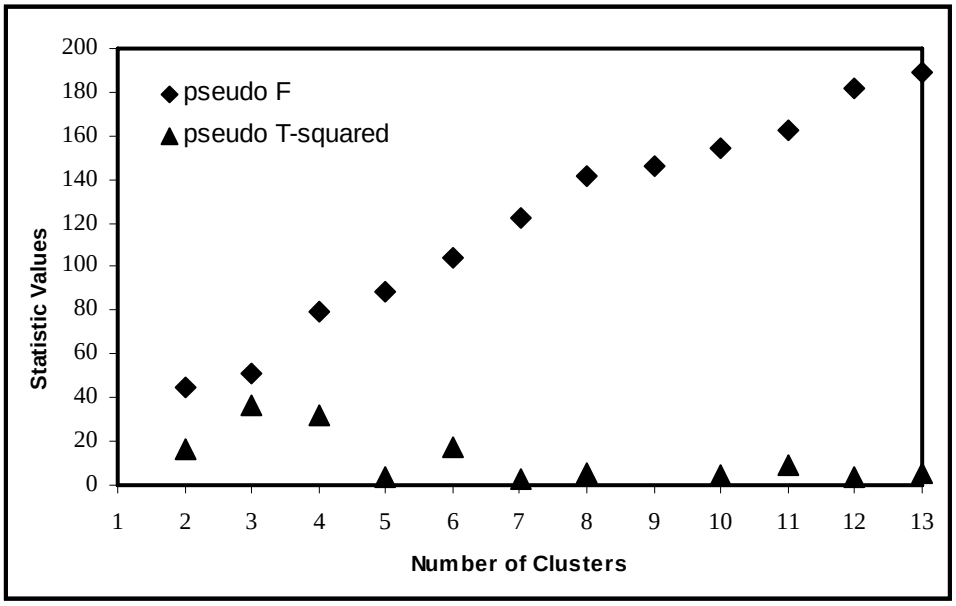

Fig. 3. pseudo *F* and t^2 statistics against number of clusters

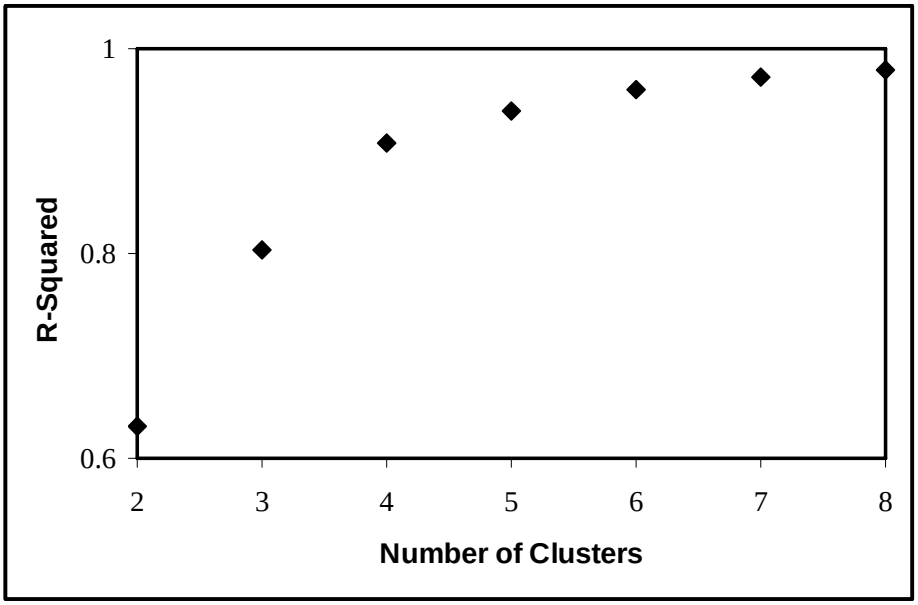

Fig. 4. R-Squared against number of clusters

Another group consists northwestern cold stations, Oroomieh and Tabriz, which was called " Azari" group by Masoodian (1998). However, Ardabil station is classified in another group with Arak and Shahrecord. This classification is also under suspicious. We can find another misleading in the clusters derived from average method such as Bandarrabbas station in western region (Sanandaj, Kermanshah and Khoramabad). It seems that average method is not a suitable method to classify rainfall regime (Ramos, 2001).

Based on the dendrogram of Ward method, the previous mistakes are not observed. The following eight clusters can be seen based on most similarity:

1- Yazd, Zahedan, Isfahan, Semnan, Kerman and Ghom.
2- Arak, Shahrecord, Ghazvin, Tehran, Mashhad and Hamedan.
3- Oroomieh, Ardabil, Zanjan and Tabriz.
4- Ahwaz, Bushehr, Shiraz, and with a small difference, Bandarabbas.
5- Kermanshah, Sanandaj and Khoramabad.
6- Ghaemshahr and Gorgan.
7- Ilam and Yasuj.
8- Rasht

In order to illustrate these clusters, we apply Canonical Discriminant Analysis (SAS/STAT, 1999). Canonical discriminant analysis is a dimension-reduction technique related to principal component analysis and canonical correlation. In a canonical discriminant analysis, we find linear combinations of the quantitative variables that provide maximal separation between the classes or groups. Two output data sets can be produced: one contains the canonical coefficients and another contains scored canonical variables. The scored canonical variables output data set can be used to plot pairs of canonical variables to aid visual interpretation of group differences (Fig. 5).

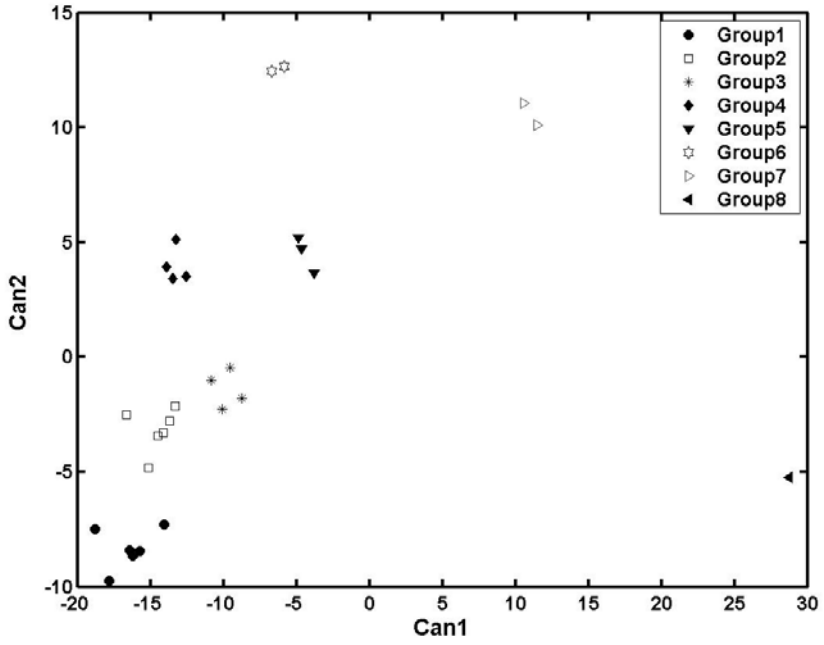

Fig 5. Illustration of the spatial separation of the canonical scores

The rainfall climates of Iran

The spatial distribution of the 8 rainfall groups over the entire country is illustrated in Fig. 6. These groups can show the different rainfall climates of Iran.

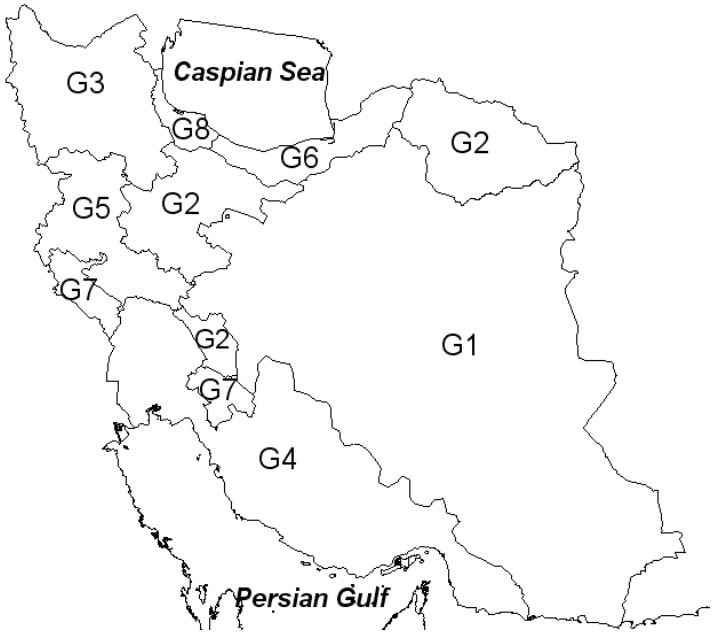

Fig. 6. Spatial distribution of rainfall groups over Iran

The monthly rainfall pattern of these 8 groups can show the difference between groups. Fig. 7 shows the mean monthly rainfall of one station in each group. For the first group which is located in the central or margin of arid and semi arid region of Iran, the main proportion of rainfall occurs in winter and spring while for the second group, the proportion of fall rainfall is higher and somehow equal to spring rainfall. In the third group, located in the north western of Iran, the main proportion of rainfall occurs in spring with the equal rainfall in fall and winter. Winter rainfall is the main proportion of rainfall in the fourth group including southwestern region of Iran. The margin of Persian Gulf (Bandarabbas) is somehow different with this part as it receives some summer rainfall (Masoodian, 1998). The fifth group which is located in the western hill slopes of Zagros Mountains receives equal winter and spring rainfall. The seventh group is similar to second and fifth group but the amount of rainfall is relatively more. This difference in the amount of rainfall, which can be the effect of higher elevation or snow precipitation, may cause Ilam and Yasuj stations to be grouped in another cluster. Comparing the sixth and eighth clusters including three stations in the margin of Caspian Sea, introduce different climate conditions in the west and east parts of the margin of Caspian Sea. The climate of western part is more humid than eastern semi arid region while the main difference of these clusters to the other clusters is summer rainfall which is not observed in any other part of Iran. From the above classification of monthly rainfall, it can be concluded that there are 3 main intera-annual variation of rainfall with different main rainfall season namely winter, winter-spring and summer rainfall.

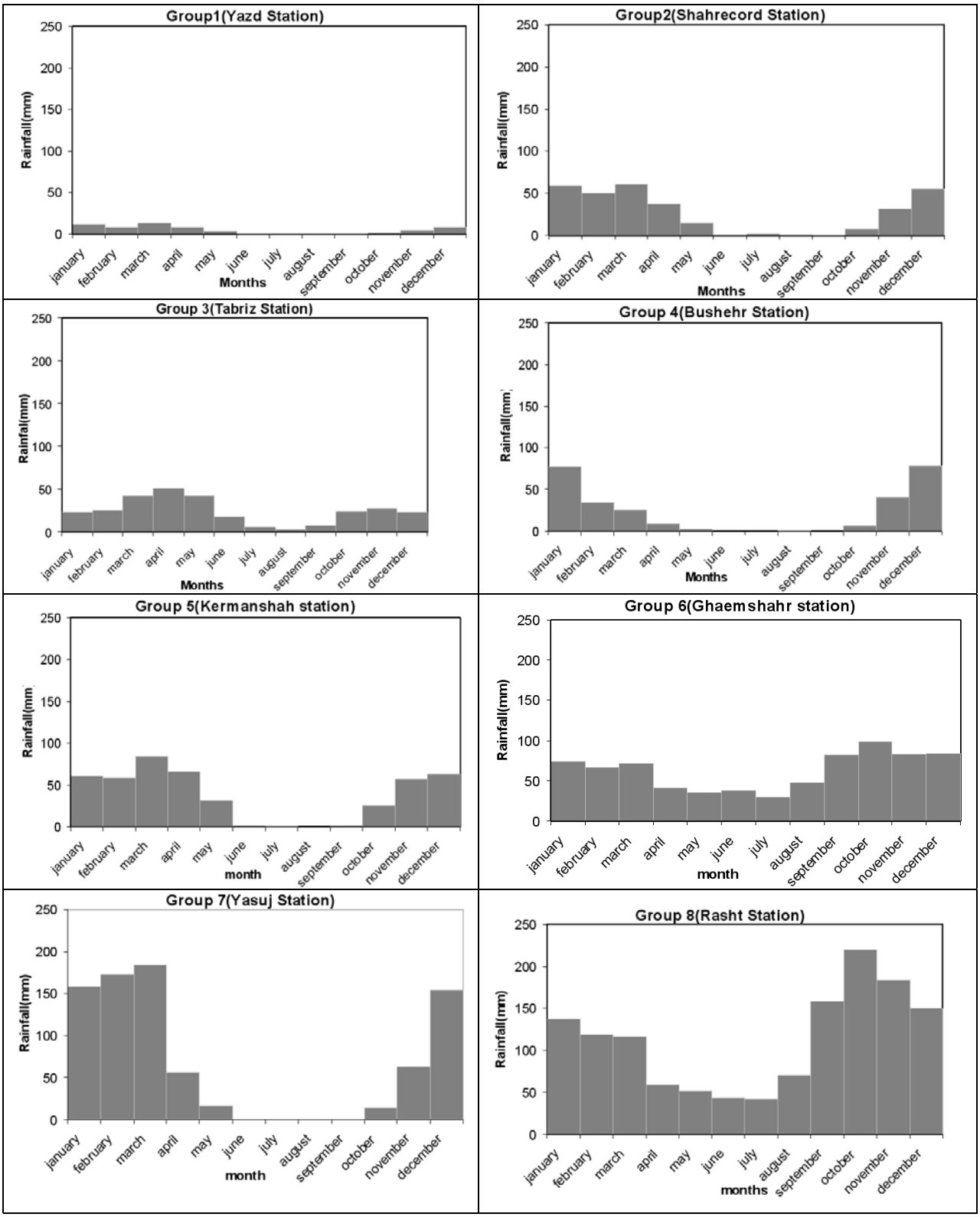

Fig. 7. Monthly rainfall distribution for one station in each rainfall group

Conclusions

This study aimed to classify annual rainfall over Iran into spatial groups. It was found that a hierarchical cluster analysis could classify this spatial pattern. The comparison of derived clusters and geographical conditions are very well matched with each other. Different clusters of central arid zone of Iran, the cold and rainy region of northwestern Iran, the impact of Zagros Mountains on rainfall of western Iran and the margins of Caspian Sea and Persian Gulf could show that the spatial pattern of Iran is influenced by sea neighborhood, elevation and latitude (Masoodian, 1998). Furthermore, it was also indicated that the results of Ward method is better than average method because the clusters derived from Ward method match very well with the result of the factors influencing rainfall in Iran. In addition, summer rainfall could only happen in the margin of Caspian Sea. The final conclusion of this study is that there are 3 main seasonal rainfall patterns over Iran, based on the total area of each regime, namely winter, winter-spring and summer regimes.

Acknowledgement

The authors wish to sincerely thank Dr Honarbakhsh for helpful comments on the very early version of this paper.

References

Acerman, M. C. 1985. Predicting the mean annual flood from basin characteristics in Scotland. Hydrological Sciences Journal. 30: 37-49.

Acerman, M. C. and Sinclair, C. D. 1986. Classification of drainage basins according to their physical characteristics: an application for flood frequency analysis in Scotland. Journal of Hydrology. 84: 365-380.

Gottschalk, L. 1985. Hydrological regionalization of Sweden. Hydrological Sciences Journal. 30: 65-83.

Jackson, I. J., Weinand, H., 195.Classification of tropical rainfall stations: A comparison of clustering techniques. Int. J. Climatol., 15, 985-994.

Jobson, J. D. 1992. Applied Multivariate Data Analysis, Vol. II: Catagorical and Multivariate Methods. Springer-Verlag, 731 pp.

Kaufman, L., and Rousseuw P. J. 1990. Finding groups in data: An introduction to cluster analysis. Wiley, New York, 344 pp.

Masoodian, S. A. 1998. An analysis of Tempo-Spatial variation or precipitation in Iran. Ph. D thesis in climatology, University of Isfahan, Iran.

Milligan, G. W., 1980. An examination of the effects of six types of errors perturbation on Fifteen clustering algorithms. Physcometrika, 45, 325-342.

Moron, V., 1997. Trend, decadal and interannual variability in annual rainfall of subequatorial and tropical North Africa (1990-1994). Int. J. Climatol. 17: 785-805.

Mosley, M. P. 1981. Delimitation of New Zealand hydrology regions. Journal of Hydrology. 49: 173-192.

Ramos, M. C., 2001. Divisive and hierarchical clustering techniques to analyze variability of rainfall distribution patterns in a Mediterranean region. J. hydrol. 57: 123-138.

Romero, R., G. Summer, C. Ramis and Genoves, A. 1999. A classification of the atmospheric circulation patterns producing significant daily rainfall in the Spanish Mediterranean area. Int. J. Climatol. 19, 765-785.

Salinger, M. J., Mullan, A. B., 1999. New Zealand: Temperature and precipitation variations and their links with atmospheric circulation 1930-1994. Int. J. Climatol. 18: 1049-1071.

SAS/STAT, User's Guide, Version 8, 1999. SAS Institute Inc., Cary, NC, USA.

Singh, C. V. 1999. Principal components of monsson rainfall in normal, flood and drought years over India. Int. J. Climatol. 19, 639-952.

Stahl, K. and Demuth, S. 1999. Methods for regional classification of streamflow drought series: Cluster analysis. Technical report to the ARIDE project, No. 1.

Tasker, G. d. 1982. Comparing methods of hydrologic regionalization. Water Resources Bulletin. 18: 965-970.