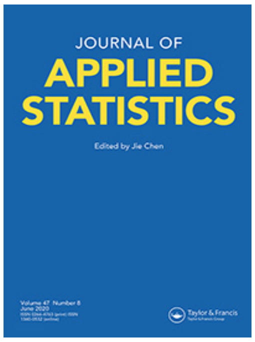
A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh

Article *in* Journal of Applied Statistics · October 2019

DOI: 10.1080/02664763.2019.1675606

Samsul Alam
North Carolina State University
6 PUBLICATIONS **62** CITATIONS

Sangita Paul
University of Dhaka
6 PUBLICATIONS **20** CITATIONS



Taylor & Francis
Taylor & Francis Group

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh

Mohammad Samsul Alam & Sangita Paul

To cite this article: Mohammad Samsul Alam & Sangita Paul (2020) A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh, Journal of Applied Statistics, 47:8, 1460-1481, DOI: 10.1080/02664763.2019.1675606

To link to this article: https://doi.org/10.1080/02664763.2019.1675606

Published online: 22 Oct 2019.

JOURNAL OF APPLIED STATISTICS
2020, VOL. 47, NO. 8, 1460–1481
https://doi.org/10.1080/02664763.2019.1675606

Taylor & Francis
Taylor & Francis Group

APPLICATION NOTE

A comparative analysis of clustering algorithms to identify the homogeneous rainfall gauge stations of Bangladesh

Mohammad Samsul Alam and Sangita Paul

Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka, Bangladesh

ABSTRACT

Dealing with individual rainfall station is time consuming as well as prone to more variation. It seems reasonable and advantageous to deal with a group of homogeneous stations rather than an individual station. Such groups can be identified using clustering algorithms, techniques used in the multivariate data analysis. Particularly, in this study, covering both hard and soft clustering approaches, three clustering algorithms namely Agglomerative hierarchical, *K*-means clustering and Fuzzy *C*-means methods are chosen due to their popularity. These algorithms are applied over precipitation data recorded by the Bangladesh Meteorology Department, and a comparison among the algorithms is made. Annual and seasonal precipitations from 1977 to 2012 recorded in 30 stations are used in this study. Optimal numbers of clusters in the four precipitation series are determined using the Gap statistic for *K*-means clustering and using the extended Gap statistic for Fuzzy *C*-means clustering, and are found as 3, 1, 3 and 2 for annual, pre-monsoon, monsoon and post-monsoon, respectively. This study investigates the clustering methods in terms of the similarity, members and homogeneity, among the clusters formed. The clusters are also characterized to see how they are distributed. Moreover, in terms of cluster homogeneity, Fuzzy *C*-means algorithm outperforms the other clustering methods.

ARTICLE HISTORY
Received 31 May 2018
Accepted 29 September 2019

KEYWORDS
Gap statistic; hierarchical clustering; *K*-means; Fuzzy *C*-means; L-moments; rainfall

1. Introduction

Bangladesh is a south-asian country where a tropical monsoon type climate prevails almost every year. Its climate is composed by a hot and rainy summer, and a dry winter. Among these, the rainy season contributes in a significant manner for the country's economy which is mainly based on the agricultural production. Moreover, in Bangladesh, rainfall is considered as the primary source of water supply for agricultural production as it releases a tremendous amount of latent heat which might have an enormous impact on the agricultural production. Therefore, for a sustainable agriculture development, an effective

CONTACT Mohammad Samsul Alam ✉ msalam@isrt.ac.bd Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

management of precipitation is a prerequisite for a developing country like Bangladesh. Such an effective management seems impossible for a process like rainfall which can vary at every point of a spatial domain. However, this apparent impossibility can be dealt by forming homogeneous zones of rainfall, as a result, such zoning is always in need of a country's governmental and non-governmental organizations.

Identification of homogeneous regions in terms of rainfall reception is important from different policy making perspectives. For example, although, different regions are subject to receive a different amount of precipitations, a classification into homogeneous regions is expected to reduce the difficulties involved in handling the precipitation gauge stations separately. Further, taking the reliable decision, at immediate situations like flood preparedness, determination of crop cycle, etc., will be feasible only when dealing with a smaller number of homogeneous groups. In addition, defining sub-regions of a spatial domain through precipitation is of keen interest for water resources management and land use planning [20,22]. Meddi *et al.* [15] stated that demarcation into sub-regions by precipitation is essential to understand the weather pattern and climate of a region. Moreover, regional flood frequency requires hydrological regionalization in terms of precipitation [10]. Therefore, undoubtedly, groups composed by the homogeneous rainfall gauge stations are always in need for spatial zoning as well as administrative policy making. For Bangladesh, it is expected that such homogeneous grouping will ease the execution of different initiatives to achieve various goals that are listed in the country's sustainable development goals (SDGs).

Clustering, a multivariate technique, allows researchers to group a set of units into distinguishable groups known as clusters. This technique is popularly known as unsupervised learning since it does not require previous knowledge regarding the groups to be formed. This multivariate technique is getting considerable popularity in identifying homogeneous groups of precipitation gauge stations [16,21]. Goyal and Gupta [5] stated that rainfall gauge stations grouped by the cluster analysis are similar in behavior. In addition, L-moments methods can be utilized along with the cluster analysis for assuring the homogeneity [7]. Malekinezhad and Zare-Garizi [14] stated that this technique is applied in all stages of regional analysis including the identification of homogeneous regions.

This study is designed to identify homogeneous regions in Bangladesh, particularly in terms of precipitation, using several algorithms covering both hard and soft clustering approaches. For such regionalization, from several alternatives, this study is planned to exploit agglomerative and *K*-means algorithms from hard clustering techniques, and Fuzzy *C*-means (FCM) algorithm from soft clustering techniques considering their popularity. Soft clustering method (Fuzzy *C*-means) is considered along with the hard clustering techniques because it allows the cases under study to belong all the clusters with a degree of membership whereas hard clustering techniques allow the cases to belong only one cluster. Moreover, Dikbas *et al.* [3] found that the regions identified using FCM are sufficiently homogeneous. Most importantly, it achieved much interest as a clustering algorithms over conventional methods in recent years due to its accurate identification of homogeneous stations [12]. Applying the considered clustering algorithms, this study will determine homogeneous regions based on the amount of annual precipitation as well as it's three components pre-monsoon, monsoon and post-monsoon.

2. Data

This study uses monthly rainfall data recorded in 35 stations operated by Bangladesh Meteorological Department (BMD). Among the stations, Tangail, Sayedpur, Chuadanga, Mongla and Ambagan are excluded from the analysis. This is because Dikbas *et al.* [3] stated that the stations to be used in the identification of homogeneous regions should have data for at least 30 years. Figure 1 shows the names and locations of the precipitation stations considered in this study.

The monthly rainfall observed by BMD from different stations are utilized in this study to get annual, pre-monsoon, monsoon and post-monsoon rainfall. For each of the stations, annual precipitation is calculated by summing the monthly rainfall over the months January to December. In a similar fashion, for each station, the pre-monsoon, monsoon and post-monsoon precipitations are computed by summing the monthly rainfall over the months February to May, June to September and October to January, respectively. Data, for different precipitation series, were constructed over the time span 1977–2012 which yields 36 observations for each of the stations. Therefore, in every situation, a clustering algorithm is applied over a data of dimension 30 × 36. That indicates, for each of the 30 stations finally retained, a time series data of 36 years is used in the analysis.

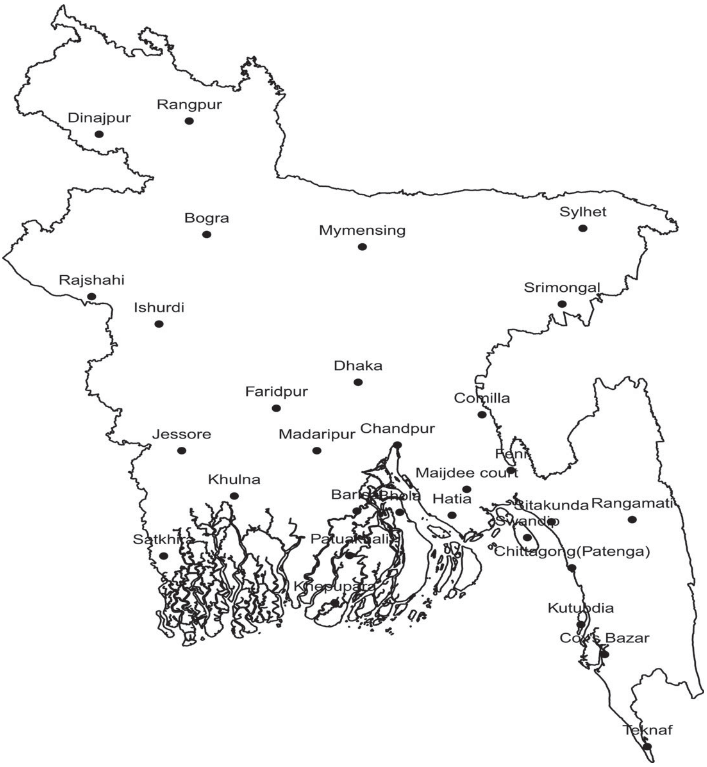


Figure 1. Precipitation gauge stations operated by BMD.

3. Statistical methods

In application of the clustering algorithms, the number of clusters to be form is one of the crucial issues and has to be determined at the time of analysis. This study uses Gap statistic [23] to determine the optimal numbers of clusters to be form in case of annual, pre-monsoon, monsoon, post-monsoon precipitations. Then, the considered clustering algorithms are used to identify the members (or rainfall gauge stations) of the clusters. Finally, homogeneity of the clusters is checked using L-moments ratio [8,9]. This section is devoted in the brief description of the aforementioned statistical methods chronologically.

3.1. Clustering algorithms

Cluster analysis or clustering is the process of grouping a set of objects in such a way that the objects in the same group are more similar to each other than to those in other groups. The groups which contain similar objects are called clusters. It is an exploratory data mining process, and a common technique for statistical data analysis, used in many fields. The clustering algorithms used in this study are agglomerative, K-means and Fuzzy C-means. Among these three algorithms, the agglomerative and K-means are well known frequently used in the identification of homogeneous region based on the rainfall measurement [16]. On the other hand, the Fuzzy C-means algorithm is relatively new and growing technique in this field [3].

However, the agglomerative algorithm starts with an individual as cluster. Then, individuals are clustered based on a similarity measure usually the Euclidean distance [11] and ends with a single cluster. Similarly, K-means method depends on a similarity measure, and Euclidean distance is used most widely. However, for a given number of clusters, say $K = k$, with n feature vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, each of dimension p, it starts by assigning the feature vectors in k clusters randomly. Then, for each cluster, the initial centroid of the clusters are calculated by taking means of the features. Next, an unit is compared with each of the centroids through the similarity measure. Finally, the unit is assigned with the cluster for which much similarity is observed, and process is repeated until no unit changes their current cluster. A comprehensive discussion on these algorithms is available in Johnson and Wicharn [11]. On the other hand, Fuzzy C-means algorithm uses distance or similarity measure also but depends on an idea known as the degree of membership.

3.1.1. Fuzzy C-means algorithm

Fuzzy clustering method was proposed by Dunn [4] based on the fuzzy logic. Later the clustering method was further developed and extended by Bezdek [1]. The mostly used fuzzy clustering method is the Fuzzy C-means method (FCM) which was proposed by Bezdek [1]. The FCM method minimizes the objective function,

$$J_m(U, V : X) = \sum_{r=1}^{c} \sum_{i=1}^{n} (u_{ri})^m d_{ri}^2(x_i, v_r), \tag{1}$$

where c is the number of clusters to be formed, x_i is a p-dimensional feature vector of individual i which forms the ith column of $p \times n$ data matrix X, u_{ri} is the value of $c \times n$

membership matrix U at cell (r, i), v_r is a p-dimensional vector defined as

$$v_r = \frac{\sum_{i=1}^{n} u_{ri}^m x_i}{\sum_{i=1}^{n} u_{ri}^m},$$

that constitutes rth column of the $p \times c$ matrix V of cluster centroids, $m \in [1, \infty]$ is the fuzzy-ness weight that controls the membership shared within the fuzzy clusters and d_{ri}^2 is the Euclidean distance between ith feature vector x_i and rth cluster mean v_r.

FCM method depends on fuzzy-ness parameter m, and works well when $m \in [1.5, 2.5]$ [18]. In the literature, there is practice, for example, Goyal and Gupta [5], of using $m = 2$. Therefore, this study uses $m = 2$ for clustering the rainfall gauge stations. Besides this, minimization of the objective function in Equation (1) is done under the constraints,

$$\sum_{r=1}^{c} u_{ri} = 1 \quad \forall\, i \in \{1, \ldots, n\}, \tag{2}$$

$$0 < \sum_{i=1}^{n} u_{ri} < n \quad \forall\, r \in \{1, \ldots, c\}. \tag{3}$$

Moreover, particularly, for minimization of (1), u_{ri} and v_r are computed in this study by the way stated in Rao and Srinivas [19] at each of the iterations.

3.2. Gap statistic

Suppose the data $\{x_{ij}\}$, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, p$, are obtained by measuring p features from n independent observations. Also, suppose that the observations belong to different clusters. But, the number of clusters in which these data belong to is unknown. For an optimal choice of this number, the Gap statistic splits the data $\{x_{ij}\}$ into c (while applying K-means algorithm $c = k$) clusters say $I_1, I_2, \ldots, I_c$ where I_r is the set of indices of observations in cluster r. For each of these clusters, let

$$D_r = \sum_{i,i' \in I_r} d_{ii'} \tag{4}$$

be the sum of pairwise distances for all points in cluster r where $d_{ii'} = \sum_j (x_{ij} - x_{i'j})^2$ is the squared Euclidean distance between the observations i and i'. Based on these total distances D_r, the pooled within-cluster sum of squares around the cluster means defined as,

$$W_c = \sum_{r=1}^{c} \frac{1}{2n_r} D_r, \tag{5}$$

where n_r is the number of observations in cluster r. The quantity W_c is then compared with its expectation under an appropriate null reference distribution of the data.

The optimal number of clusters is then determined by the value of c for which the quantity

$$\text{Gap}_n(c) = E_n^* \left\{\log(W_c)\right\} - \log(W_c), \tag{6}$$

where E_n^* indicates expectation under a sample of size n from the reference distribution, is maximum after taking the sampling distribution into account. This task is done by plotting the value of $\mathrm{Gap}_n(c)$ in Equation (6) against different values of c. However, implementation of the Gap statistic is carried out through Monte Carlo approach in which $E_n^*\{\log(W_c)\}$ is estimated from B copies of $\log(W_c^*)$. Each of these $\log(W_c^*)$ is calculated by taking a Monte Carlo sample of size n from the chosen reference distribution. Two approaches have been suggested by Tibshirani *et al.* [23] for generating the reference distribution. The first is to generate each of the features uniformly over its respective range whereas the second is to generate the reference features by utilizing the principle component of the data. Later approach is carried out through a uniform distribution which is over a box aligned with the principle component of the data. Specifically, the second approach performs a singular value decomposition of the data to obtain as transformed data. This transformation is made by $X' = XV$ where X is the data matrix and V is such that $X = UDV^T$. Then, a new matrix Z' is formed by generating features from uniform distributions where ranges are obtained from the columns of X'. Finally, to get the reference data Z, a back transformation though $Z = Z'V^T$ is performed. Tibshirani *et al.* [23] suggested that the standard deviation $\mathrm{sd}(c)$ in B replicates of $\log(W_c^*)$ can be utilized in decision making in terms of

$$s_c = \sqrt{(1 + 1/B)}\mathrm{sd}(c).$$

The optimal number of clusters, $\hat{c}$, is then determined by the smallest c such that $\mathrm{Gap}_n(c) \geq \mathrm{Gap}_n(c+1) - s_{c+1}$.

An extension of this Gap statistic for using in the case of Fuzzy clustering has been proposed by Yue *et al.* [24]. This extension considers the softness of Fuzzy clustering incorporating the membership degrees while computing the within-cluster sum of squares defined in Equation (5). For a given number of clusters (say c), this extension defines the following two quantities using the elements u_{ik} of the Fuzzy membership matrix U

$$u_{\cdot i} = \sum_{r=1}^{c} u_{ri} \quad \text{and} \quad u_{r\cdot} = \sum_{i=1}^{n} u_{ri}, \tag{7}$$

u_{ri} is the value at cell (r, i) of the matrix U defined in Equation 1. As stated earlier in Equation (2) the quantities $u_{\cdot i}$ are all equal to 1 in this study. The later quantity, similar to the hard clustering, gives the number of objects in the cluster r, which is written as $u_{r\cdot} = n_r$. More on the values of the quantities in Equation (7) has been given in Yue *et al.* [24]. In this extension, the quantities in Equations (4) and (5) are calculated, respectively, as

$$D_r = \sum_{i,i'=1}^{n} \min\left\{u_{ri}^m, u_{ri'}^m\right\} d_{ii'}^2; \quad r = 1, 2, \ldots, c, \text{ and } W_c = \sum_{r=1}^{c} \frac{1}{n_r} D_r; \quad r = 1, 2, \ldots, c.$$

3.3. L-moments based homogeneity test

L-moments are latest advances in mathematical statistics based on probability weighted moments [6] that ease the estimation of frequency analysis [17]. However, in regional frequency analysis using the L-moments method, homogeneity is tested using heterogeneity measure namely H statistic [2]. The same idea is used in this study to check the homogeneity of clusters identified by the different clustering algorithms. Hosking and Wallis [8] described three versions H_1, H_2 and H_3 of H statistic whose are based on the three

L-moments ratios L-coefficient of variation (L_{CV}), L-coefficient of skewness (L_{CS}) and L-coefficient of kurtosis (L_{CK}) respectively. Among these, the H_1 is based on L_{CV} and chosen in this study since it indicates heterogeneity or potential heterogeneity more often than H_2 and H_3 [13]. Moreover, four parameters Kappa distribution is fitted in this study to the regional datasets following Hosking and Wallis [8,9] to evaluate the homogeneity.

The computation of H_1 statistic requires the sample variance of L_{CV} which is obtained using

$$V_1 = \frac{\sum_i^p n_i \left(L_{CV}^i - \bar{L}_{CV}\right)^2}{\sum_{i=1}^p n_i}, \tag{8}$$

where n_i is the number of data points in the feature vector (rainfall gauge station) i among the available p vectors, L_{CV}^i is the L_{CV} computed for feature vector i, and $\bar{L}_{CV}$ is the average of L_{CV}^i taken over all p feature vectors. The detail of computing the L_{CV}^i in Equation (8) is available in Hosking and Wallis [8,9]. Then, H_1 is computed as,

$$H_1 = \frac{V_1 - \mu_{V_1}}{\sigma_{v_1}}, \tag{9}$$

where μ_{V_1} and σ_{v_1} are average and standard deviation calculated from the simulated data using four parameters Kappa distribution which is considered the expected under the assumption of homogeneity.

The hypothesis of homogeneity is evaluated based on the value of H_1 computed through the Equation (9). Hosking and Wallis [8] summarized that the sites (rainfall gauge stations in the same cluster) are acceptable homogeneous if $H_1 < 1$, possibly homogeneous if $1 \leq H_1 < 2$ or definitely heterogeneous if $H_1 \geq 2$.

3.4. Similarity of clustering methods

Suppose that n units are classified by two clustering methods A and B into r and k clusters, respectively. The resulting clustering can be summarized in a $r \times k$ contingency table with $n_{i\cdot} = \sum_{j=1}^k n_{ij}$, $n_{\cdot j} = \sum_{i=1}^r n_{ij}$ and $n = \sum_{i=1}^r \sum_{j=1}^k n_{ij}$ as that in Table 1. Now, by defining four quantities, a, b, c and d, respectively, as

a = number of pairs of objects in the same cluster by both A and B,

b = number of pairs of objects in the same cluster by A but not by B,

c = number of pairs of objects in the same cluster by B but not by A,

d = number of pairs of objects in not in the same cluster by both A and B,

the rand index (RI) for assessing the similarity between the two clustering methods is defined as,

$$RI = \frac{a + d}{a + b + c + d}. \tag{10}$$

Table 1. Contingency table summarizing units classified into different clusters by two clustering methods.

	B				
A	1	2	...	k	Total
1	n_{11}	n_{12}	...	n_{1k}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2k}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	...	n_{rk}	$n_{r\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot k}$	n

To compute the RI in Equation (10), one needs to compute the values of a, b, c and d that can be obtained as,

$$a = \sum_{i=1}^{r}\sum_{j=1}^{k}\binom{n_{ij}}{2}, b = \sum_{i=1}^{r}\binom{n_{i\cdot}}{2} - a, c = \sum_{i=1}^{k}\binom{n_{\cdot j}}{2} - a, \text{ and } d = \binom{n}{2} - a - b - c.$$

Mutual information (MI), based on the classification shown in the Table 1, between the clustering methods A and B is calculated as,

$$MI = H(A) + H(B) - H(A,B) = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}}{n}\log\frac{n_{ij}n}{n_{i\cdot}n_{\cdot j}}, \tag{11}$$

where $H(A)$ is the entropy of method A, $H(B)$ is the entropy of method B and $H(A,B)$ is joint entropy of methods A and B. These quantities are computed by,

$$H(A) = -\sum_{i=1}^{r}\frac{n_{i\cdot}}{n}\log\frac{n_{i\cdot}}{n},\ H(B) = -\sum_{j=1}^{k}\frac{n_{\cdot j}}{n}\log\frac{n_{\cdot j}}{n}, \text{ and}$$

$$H(A,B) = -\sum_{i=1}^{r}\sum_{j=1}^{k}\frac{n_{ij}}{n}\log\frac{n_{ij}}{n}.$$

4. Results and discussion

This study analyzes 30 rainfall gauge stations of BMD. The measurements of rainfall over the time span 1977–2012 are used for every stations. As a result, 36 measurements from the 36 years are analyzed for each of the annual, pre-monsoon, monsoon and post-monsoon rainfall. The data used in this study are presented in Figure 2.

Cluster analysis, particularly, the K-means and Fuzzy C-means, has the number of clusters to be formed as a parameter. This requirement, in case of different precipitations (annual, pre-monsoon, monsoon and post-monsoon), is accomplished by determining the optimal numbers by the Gap statistic and extended Gap statistic, respectively, for K-means and Fuzzy C-means algorithms. Then, cluster analysis is performed applying agglomerative hierarchical, K-means and Fuzzy C-means over the four precipitation series. Then a

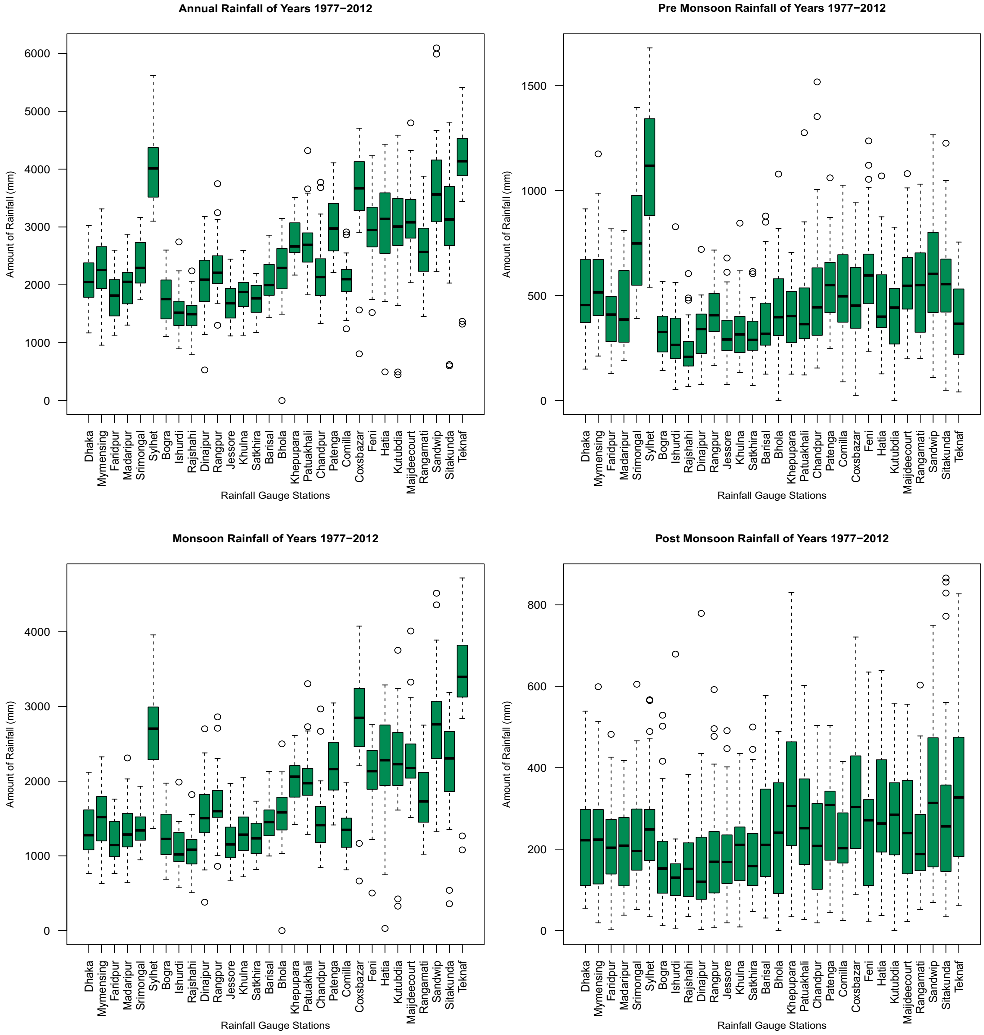


Figure 2. Distribution of annual, pre-monsoon, monsoon and post-monsoon rainfall over the years 1977–2012 recorded in different rainfall gauge stations.

homogeneity test of the clusters identified by the clustering methods is to performed using H_1 statistic based on the L-moments ratio. In addition, the clustering methods are to be assessed for the similarity in the cluster formation using the rand index (RI) and mutual information (MI). Finally, the clusters are characterized schematically for the purpose of understanding the rainfall mechanism.

The statistical software `R` is utilized to carryout the whole analysis of this study. However, the application of extended gap statistic using Fuzzy *C*-means algorithm is done by writing own codes. On the other cases, built-in functions and few libraries are used. Particularly, `R` libraries `fcluster`, `aricode` and `homtest` are used for implementing Fuzzy *C*-means clustering, rand index and mutual information, and homogeneity test, respectively.

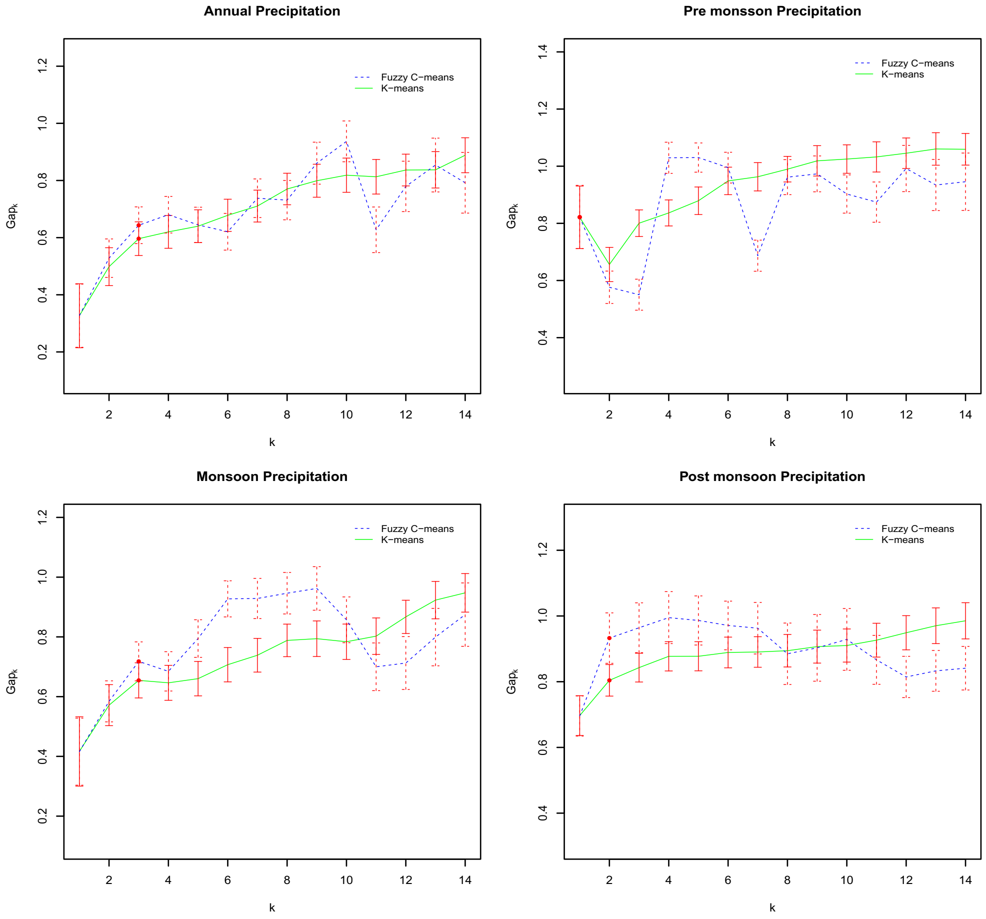


Figure 3. The Gap statistics and associated confidence intervals in case of Annual, Pre-monsoon, Monsoon and Post-monsoon precipitations.

4.1. Optimal numbers of clusters

To determine the optimal numbers of clusters using Gap statistic described in Section 3.2, the minimum and maximum numbers are set 1 and 14, respectively, in this study. Results obtained through the Gap statistic in case of the four precipitation series are summarized schematically in Figure 3 where vertical lines representing the corresponding confidence intervals. Moreover, the associated quantitative results are presented in Table 2.

The Gap statistic is calculated using both *K*-means and Fuzzy *C*-means methods where the second approach (see Section 3.2) of generating the reference distribution is used. Since hierarchical clustering, particularly, agglomerative clustering technique starts each station as an individual cluster and finishes with a single cluster, hence this algorithm does not have number of clusters as its parameter. As a result, the Gap statistic is not applied using the agglomerative clustering method. However, for the purpose of comparison, clustering output is summarized by keeping similar numbers of clusters as that found by the Gap statistic using the *K*-means clustering.

Table 2. The Gap statistics with their associated standard errors in case of Annual, Pre-monsoon, Monsoon and Post-monsoon precipitations.

	Number of clusters	Precipitation series							
		Annual		Pre-monsoon		Monsoon		Post-monsoon	
	k	$GAP_n(k)$	s_k	$GAP_n(k)$	s_k	$GAP_n(k)$	s_k	$GAP_n(k)$	s_k
K-Means	1	0.327	0.112	0.821	0.109	0.417	0.116	0.697	0.061
	2	0.498	0.066	0.656	0.060	0.571	0.069	0.804	0.048
	3	0.596	0.059	0.801	0.047	0.654	0.059	0.843	0.044
	4	0.620	0.057	0.836	0.046	0.646	0.059	0.877	0.044
	5	0.640	0.057	0.879	0.048	0.660	0.058	0.877	0.044
	6	0.678	0.056	0.948	0.048	0.707	0.057	0.889	0.047
	7	0.710	0.056	0.963	0.049	0.739	0.056	0.890	0.046
	8	0.770	0.055	0.990	0.045	0.788	0.054	0.894	0.049
	9	0.799	0.058	1.019	0.053	0.794	0.060	0.907	0.050
	10	0.819	0.060	1.025	0.050	0.784	0.059	0.910	0.050
	11	0.813	0.060	1.033	0.053	0.803	0.061	0.926	0.051
	12	0.837	0.056	1.045	0.053	0.867	0.056	0.949	0.052
	13	0.837	0.064	1.060	0.057	0.923	0.063	0.970	0.054
	14	0.888	0.061	1.059	0.055	0.947	0.065	0.985	0.055
Fuzzy *C*-Means	1	0.327	0.111	0.822	0.111	0.416	0.112	0.696	0.061
	2	0.528	0.067	0.576	0.057	0.584	0.069	0.933	0.077
	3	0.643	0.064	0.550	0.054	0.718	0.065	0.964	0.076
	4	0.680	0.064	1.029	0.054	0.685	0.066	0.995	0.079
	5	0.645	0.062	1.030	0.051	0.794	0.063	0.986	0.075
	6	0.621	0.064	0.995	0.054	0.927	0.060	0.971	0.074
	7	0.738	0.068	0.687	0.054	0.929	0.067	0.963	0.078
	8	0.731	0.069	0.962	0.061	0.946	0.069	0.885	0.093
	9	0.861	0.073	0.973	0.063	0.962	0.073	0.903	0.102
	10	0.937	0.071	0.903	0.067	0.857	0.077	0.929	0.094
	11	0.627	0.080	0.874	0.070	0.700	0.080	0.867	0.074
	12	0.779	0.088	0.992	0.081	0.712	0.088	0.814	0.063
	13	0.854	0.094	0.934	0.090	0.799	0.096	0.833	0.062
	14	0.792	0.106	0.946	0.100	0.875	0.106	0.841	0.066

In case of annual precipitation, for both K-means and Fuzzy C-means methods, from top left panel of Figure 3, it is observed that $\text{Gap}_n(k)$ defined in Equation (6) increases notably at $k = 2, 3, 8$ and 14 for K-means whereas at $k = 2, 3, 7, 9, 10, 12$ and 13 for Fuzzy C-means. Following the criterion (see Section 3.2) suggested by Tibshirani *et al.* [23] this study chooses $k = 3$ for both K-means and Fuzzy C-means clustering algorithms. Similarly, investigating the upper right panel of Figure 3 and the values for Pre-monsoon precipitation in Table 2, it is observed that, following criterion proposed by Tibshirani *et al.* [23], 1 cluster is sufficient for the 30 stations considered in this study. This observation is true for both the K-means and Fuzzy C-means. Therefore, in case of pre-monsoon rainfall, no clustering will be made using the K-means and Fuzzy C-means algorithm as all the stations belong to the same cluster. However, the hierarchical clustering will be applied to see how these stations form the cluster. The same criterion selects $k = 3$ and $k = 2$, in case of both (K-means and Fuzzy C-means) the clustering algorithms, for monsoon and post-monsoon rainfall, respectively.

4.2. Clustering of precipitation series

The precipitation gauge stations of Bangladesh are than classified into 3, 1, 3 and 2 clusters, respectively, for annual, pre-monsoon, monsoon and Post-monsoon precipitation using

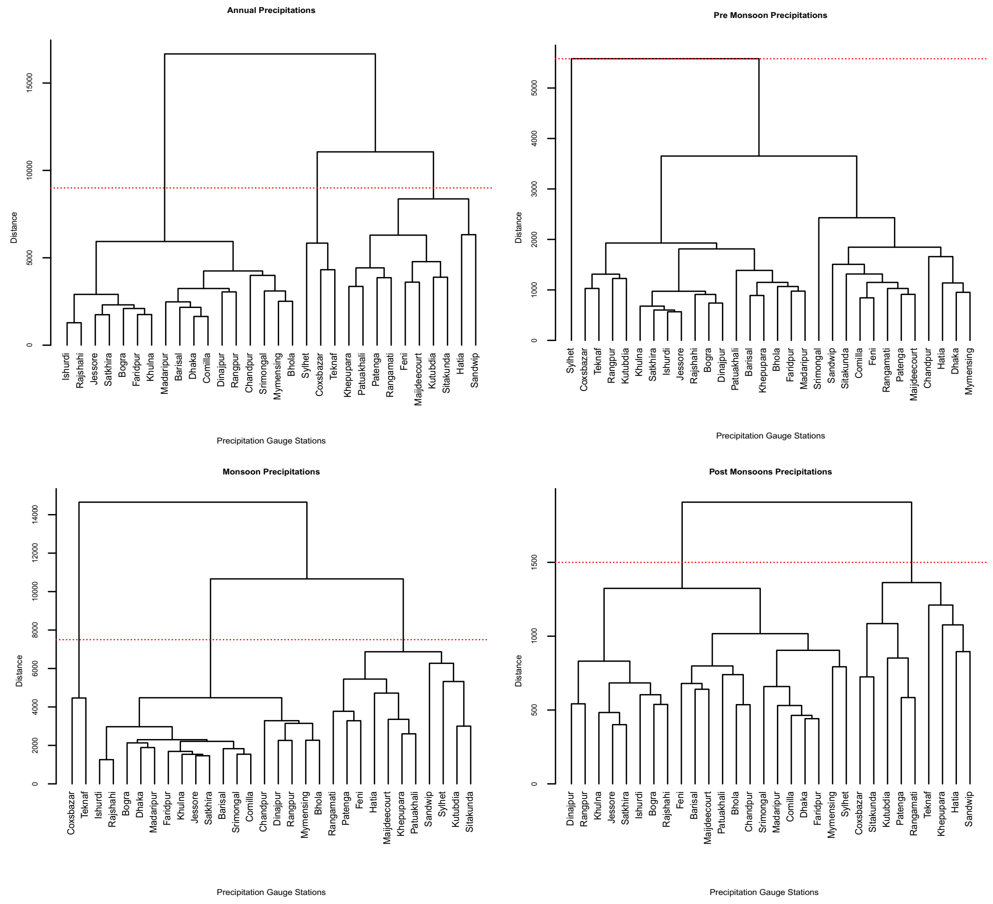


Figure 4. Dendogram using the agglomerative hierarchical clustering for Annual, Pre-monsoon, Monsoon and Post-monsoon precipitations.

the three clustering algorithms. In doing so, this study first applies hierarchical clustering, in particular, the agglomerative clustering to construct clusters of precipitation gauge stations.

4.2.1. Agglomerative clustering

Clusters formed in case of the four precipitation series annual, pre-monsoon, monsoon and post-monsoon by the agglomerative hierarchical clustering with complete linkage are shown in Figure 4 through dendrograms. A horizontal line is drawn over every dendrogram to identify clusters. To do so, numbers of clusters for different precipitation series, respectively, are kept as same as that found for K-means algorithm by the Gap statistic. Moreover, the spatial distribution of the clusters found is shown in Figure 5 with different symbols for different clusters. Stations that belong to the first cluster (from left to right in the top left map of Figure 5) of annual rainfall are distributed almost all over the country except the south eastern part. Precipitation gauge stations that are located in the south eastern part of Bangladesh including the Sylhet, a station located in the north eastern part, form the two other clusters in case of annual rainfall. More specifically, this study observes that,

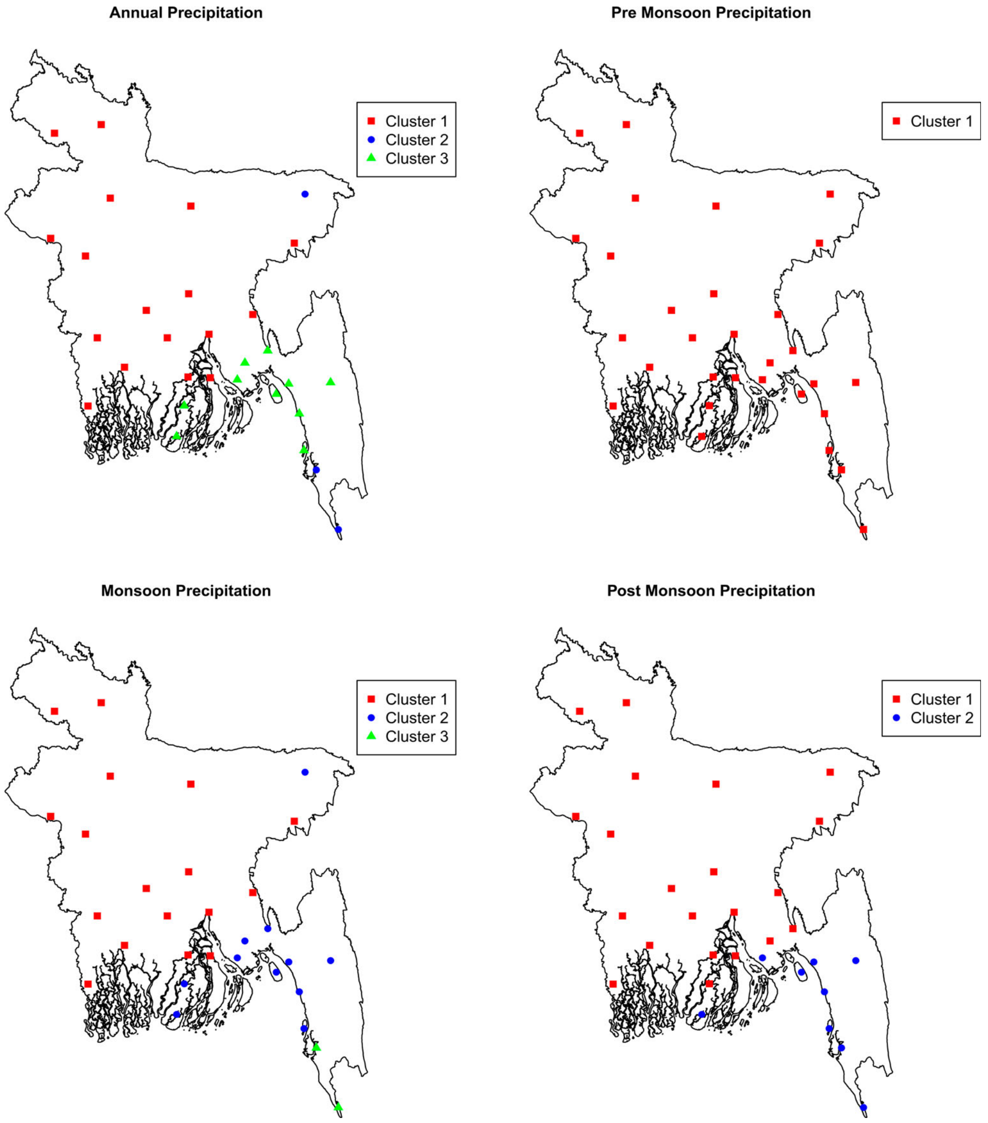


Figure 5. Spatial distribution of the clusters formed by agglomerative hierarchical clustering method in case of Annual, Pre-monsoon, Monsoon and Post-monsoon precipitations.

in case of the annual rainfall, the Sylhet station is similar to the stations Cox's Bazar and Teknaf from the south eastern part of the country, and these three stations form a cluster. The right top map shows the spatial distribution of the stations for pre-monsoon rainfall. All the stations are presented by the same color and symbol as they belong to a single cluster. Notably, this study finds, from the maps in the left panel of Figure 5, spatial distributions of clusters for annual and monsoon rainfall are almost identical. The only difference is observed in case of the Sylhet station which changes its belonging in monsoon rainfall than it had in case of annual rainfall. Henceforth, it can easily be stated that monsoon rainfall is highly influential in the cluster formation of the annual rainfall. For post-monsoon

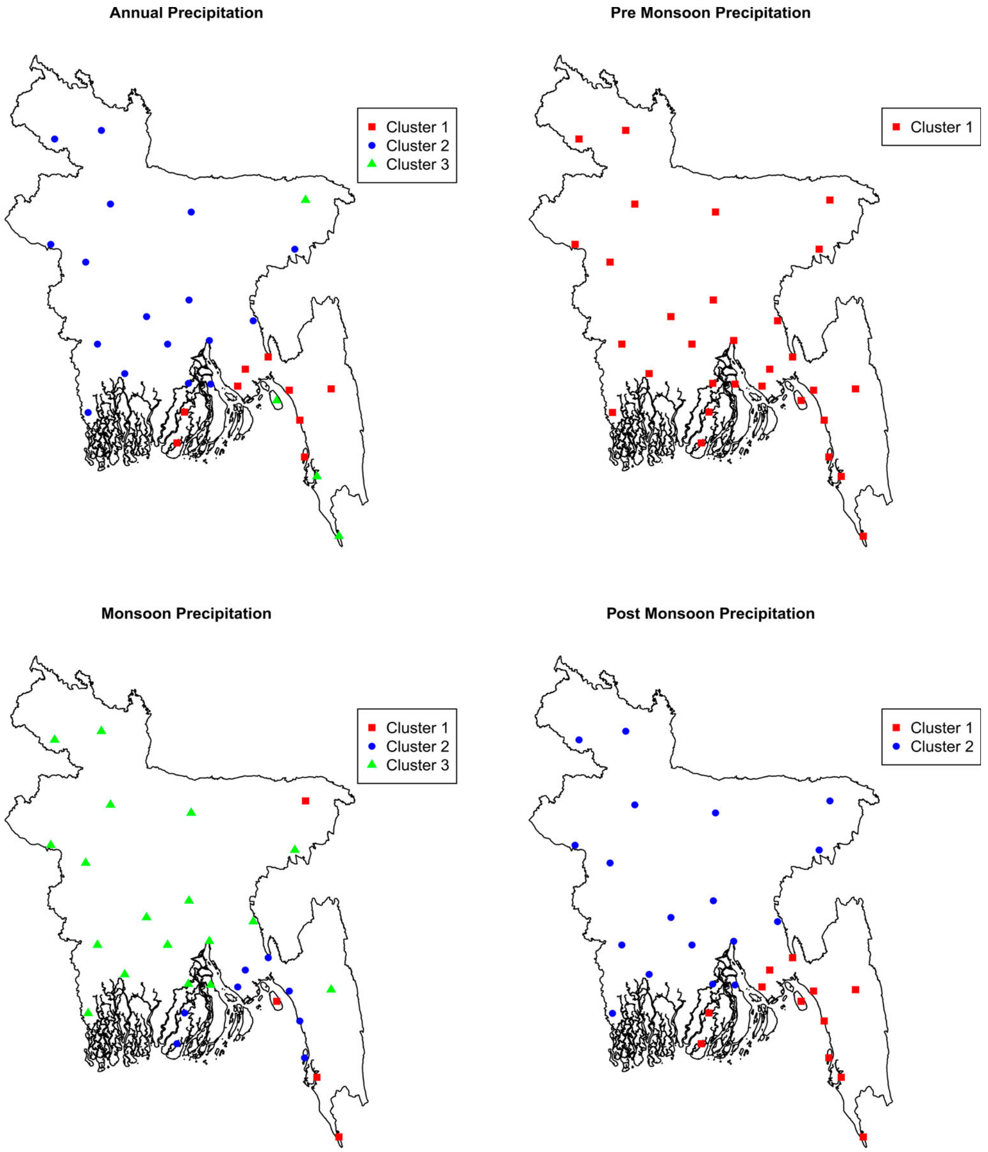


Figure 6. Spatial distribution of the clusters formed by *K*-means clustering method in case of Annual, Pre-monsoon, Monsoon and Post-monsoon precipitations.

rainfall, this study observes that stations from the south eastern part of Bangladesh along with the Khepupara, a station from the south western part, form a cluster. Rest of the stations belong to the other cluster whose spatial distribution encompasses almost all over Bangladesh.

4.2.2. K-means clustering

The *K*-means algorithm has been applied over the precipitation series to identify which stations belong to which clusters. To do so, the number of clusters, in case of different precipitation series, are formed according to the optimal numbers found through the Gap statistic which is presented in Section 4.1. For a given number of clusters, K-means algorithm requires the selection of initial centroids for the clusters. This study selects initial centroids randomly. However, to ensure the stability of desired clusters, the random selection of the initial centroids is repeated for 100 times. Among these, the initial centroid for which the total within cluster sum of squares is minimum has been used to get the final results.

Spatial distributions of the clusters formed for different precipitation series are depicted in Figure 6. Specifically, graph for annual rainfall reflects that overall Bangladesh can be separated, approximately, into 3 distinguish spatial zone from west to east by the 3 clusters formed. However, rainfall gauge station in Sylhet belongs to the cluster formed by the stations in Chittagong (Patenga), Cox's Bazar and Teknaf. These three stations are located on the bank of the Bay of Bengal. Except these three, 7 stations from the south eastern part and 2 stations from south western part form another cluster. The other cluster, in case of the annual rainfall, has the remaining 17 stations as its members. It is note worthy to mention that, for monsoon rainfall, K-means algorithm identified 3 clusters almost identical to that obtained in case of annual rainfall. Here, there are also 18 stations in cluster 3 which captures most of the spatial domain of Bangladesh. Particularly, this cluster is distributed over the middle and western part of Bangladesh. The other two clusters have been found almost similar two those observed in case of annual rainfall. The only difference is observed due to the Rangamati station. This station behaves like the stations located in the plain land of the western part. Likewise monsoon rainfall, for post-monsoon rainfall, there is a large cluster of 18 rainfall stations which are distributed over north eastern to southern part. On the other hand, all the stations in south east part including two from south western part form the other cluster.

4.2.3. *Fuzzy C-means clustering*

Although, Fuzzy clustering allows a rainfall station to belong all the clusters under consideration with different membership degree, this study assigns the stations in cluster for which the membership degree is greater. This particular isolation, in this study, is made to compare Fuzzy clustering with the other hard clustering techniques considered. However, the membership degrees for different clusters of the rainfall stations are shown in Table A1.

Likewise the hierarchical and K-means clustering algorithms, spatial distributions of the clusters formed by the Fuzzy C-means clustering are summarized in the map of Bangladesh, and respected maps of the four precipitation series are presented all together in Figure 7. The map for annual precipitation depicts that cluster 1 and 3 have spatial zoning. On the other hand, cluster 2 is formed by the stations those are distributed in different areas of the country. The cluster 3 consists of stations in the south western part except Rangamati, and the Sylhet station. Unlike agglomerative and K-means methods, here, it is observed that the cluster 2 includes stations from different (see the blue dots) parts of Bangladesh. The right top map which drawn for pre-monsoon rainfall presents all the stations with the same symbol as they all belong to a single cluster. The bottom left map representing monsoon rainfall shows that the cluster 1 has 15 rainfall stations as its members. This cluster includes all but Rangpur, Sylhet, Patuakhali, Khepupara and stations in

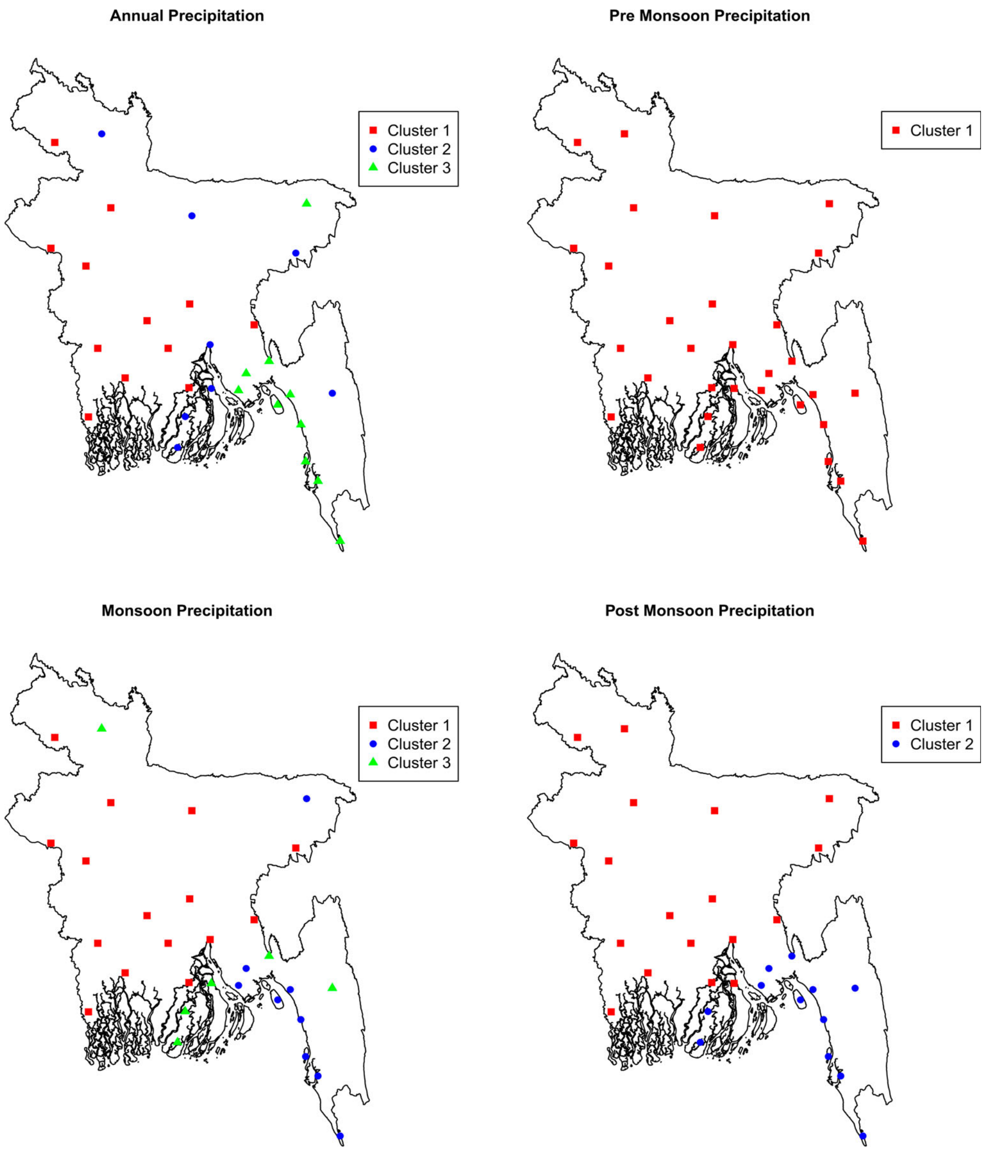


Figure 7. Spatial distribution of the clusters formed by Fuzzy *C*-means clustering method in case of Annual, Pre-monsoon, Monsoon and Post-monsoon precipitations.

the south eastern part of Bangladesh. The other two clusters comprise 6 and 9 stations, respectively. Between these two, Sylhet station belongs to the cluster formed by the stations in Chittagong divisions except Feni and Rangamati. These two rainfall gauge stations belong to the cluster formed by the Rangpur and stations in the Barisal division. Finally, for post-monsoon rainfall, Fuzzy clustering forms two non-overlapping spatial regions based on the clusters formed. According to this clustering algorithm, stations in the south east part of Bangladesh belong to one cluster, and the rest of the stations belong to the other cluster.

Table 3. Homogeneity of the clusters due to Fuzzy *C*-means, Hierarchical and *K*-means clustering for Annual, Pre-monsoon, Monsoon and Post-monsoon based on the values of H_1.

		Clustering method					
		Hierarchical		*K*-means		Fuzzy *C*-means	
Precipitations	Cluster	Stations	H_1	Stations	H_1	Stations	H_1
Annual	1	17	0.52^a	9	2.17^c	12	0.45^a
	2	3	-0.10^a	17	0.52^a	8	1.87^b
	3	10	2.05^c	4	0.16^a	10	1.42^b
Pre-monsoon	1	30	1.14^b	30	1.14^b	30	1.14^b
Monsoon	1	17	0.56^a	4	-0.53^a	15	1.30^b
	2	11	1.66^b	8	1.82^b	9	0.91^a
	3	2	-0.28^a	18	0.51^a	6	0.44^a
Post-monsoon	1	21	-0.05^a	12	0.04^a	18	0.45^a
	2	9	0.74^a	18	0.45^a	12	0.04^a

[a] Acceptably homogeneous
[b] Possibly homogeneous
[c] Definitely heterogeneous

4.3. Homogeneity of the clusters

Clusters obtained, for the four precipitation series, by different clustering techniques are then tested for within-cluster homogeneity. This has been done using one of the test statistics due to Hosking and Wallis [8] which is based on the L-moments and known as H_1 statistic. The values of H_1 observed for various cases are shown in Table 3, and homogeneity is assessed through the criterion mentioned in Section 3.3.

Results obtained for annual rainfall show that agglomerative clustering forms 2 acceptably homogeneous clusters and 1 definitely heterogeneous cluster. Similar findings are observed for the clusters obtained through *K*-means clustering. However, none of the clusters, in the case of Fuzzy *C*-means clustering, are definitely heterogeneous. Besides this, it is observed that 2 clusters are possibly and 1 cluster is acceptably homogeneous. Since, for pre-monsoon rainfall, one clustered is formed, results of homogeneity test are identical for all the algorithms. Moreover, that cluster is found to be possibly homogeneous. Most uniform results have observed in the case of monsoon rainfall from the three methods. Though, clusters are different in terms of the members they have, 2 clusters, for all the methods, are acceptably and 1 cluster is possibly homogeneous. Likewise, this study finds that the three algorithms identified, separately, 2 possibly homogeneous clusters in case of post-monsoon rainfall. Additional to this, it is observed that the *K*-means and FCMs formed clusters of same sizes.

4.4. Similarity of the clustering methods

How similarly the units are classified by the three clustering methods is assessed by two common indices, namely, rand index (RI) and mutual information (MI). The results found are summarized in Table 4. Clustering methods are compared for annual, monsoon and post-monsoon rainfall. The comparison is not made for the pre monsoon rainfall, because this study observed that all the rainfall gauge stations belong to the same cluster.

Clustering of the annual rainfall that has been obtained by the hierarchical clustering method (agglomerative) is similar to that obtained by the *K*-means clustering method. On

Table 4. Rand Index (RI) and Mutual Information (MI) for the clustering obtained by the hierarchical clustering (HC), *K*-means (KM) and fuzzy *C*-means (FCM) algorithms.

	HC-KM		HC-FCM		KM-FCM	
	RI	MI	RI	MI	RI	MI
Annual	0.97	0.84	0.73	0.54	0.73	0.55
Monsoon	0.89	0.65	0.82	0.58	0.80	0.58
Post-monsoon	0.81	0.39	0.69	0.32	0.78	0.44

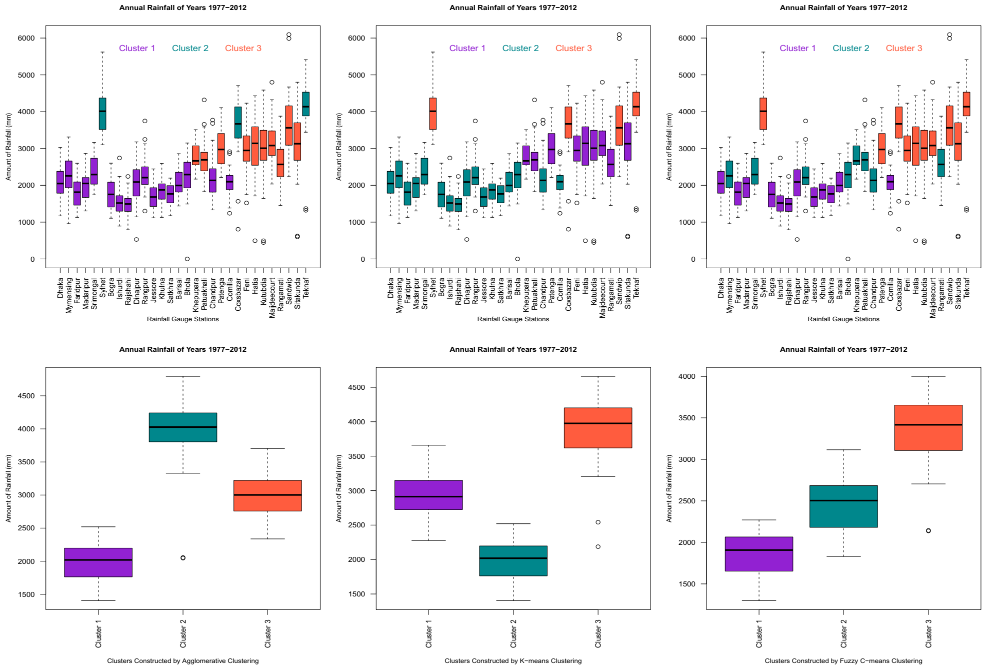


Figure 8. Clusters obtained for annual rainfall by the agglomerative, *K*-means and fuzzy *C*-means algorithms.

the other hand, the clustering of annual rainfall that has been observed by the FCM is not too similar to the clustering found by the two other algorithms. In addition, values of the MI for annual precipitation, when compared between the hierarchical clustering and *K*-means with fuzzy *C*-means, indicate minimal mutual information between HC-FCM and KM-FCM. Besides these, it has been observed that the three clustering methods are more similar in the case of monsoon rainfall whereas they are less similar in case of post-monsoon rainfall.

4.5. Characterization of clusters

The clusters obtained in case of different precipitation series by the three algorithms are explored by constructing boxplots for individual stations and cluster centroids.

In Figures 8–10, boxplots are presenting how the individual stations and clusters are standing apart from each other in terms of different precipitation series. For a given

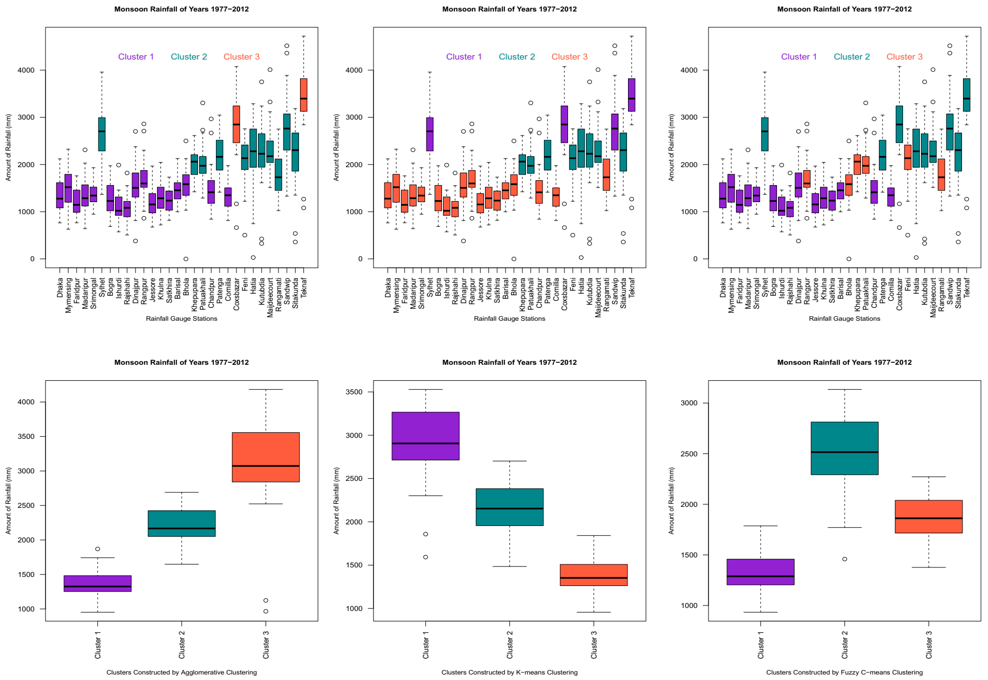


Figure 9. Clusters obtained for monsoon rainfall by the agglomerative, *K*-means and fuzzy *C*-means algorithms.

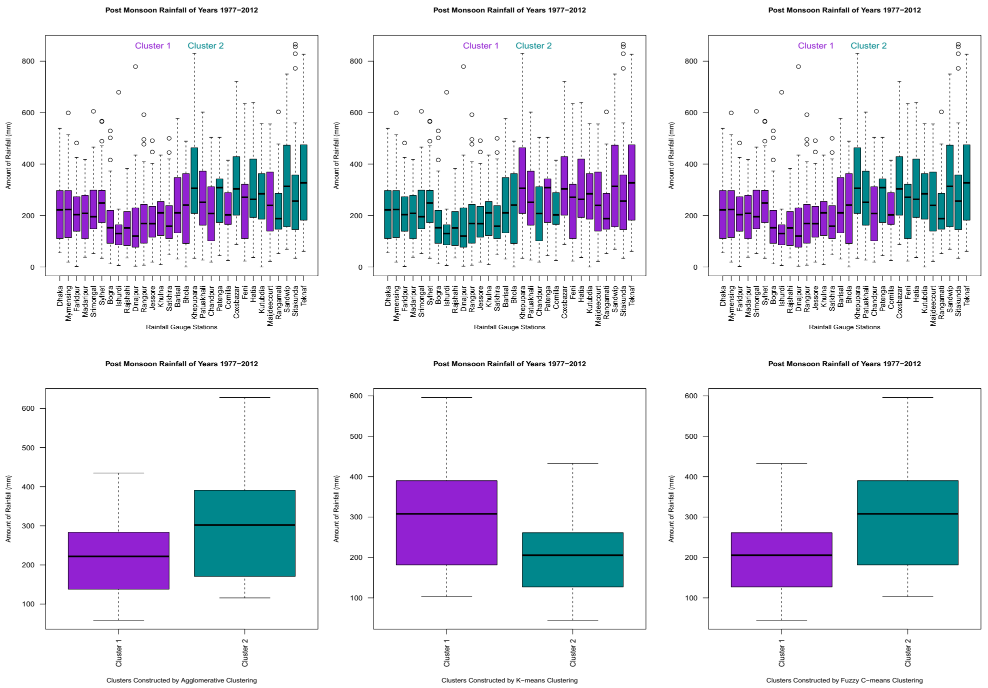


Figure 10. Clusters obtained for post-monsoon rainfall by the agglomerative, *K*-means and fuzzy *C*-means algorithms.

precipitation series, scrutinizing the boxplots for various clusters, it can be stated that the rainfall distributions of various clusters are different in terms of their location and scale. However, for post-monsoon rainfall, in case of clusters found by the K-means algorithm, imbrication has been observed in the distributions of rainfall. This may create ambiguity while making policies for the post-monsoon season.

5. Conclusions

This study creates clusters, in case of different precipitation series, by 30 stations located over the spatial domain of Bangladesh. This is done aiming to ease the duties of government and non-government organizations while making policies where rainfall plays a vital role. To do that, three familiar clustering algorithms namely agglomerative, K-means and Fuzzy C-means are exploited. This study not only identifies the clusters but also assesses their within-cluster homogeneity. Specifically, rainfall data are analyzed considering aggregated and segregated time spans of a year. Moreover, the spatial distribution, in different situations, is presented in this study using 30 rainfall gauge stations of the BMD. Finally, the clusters obtained in different scenarios by the three methods are explored schematically.

This study finds that the 30 stations that are considered belong to 3, 1, 3 and 2 clusters while dealing with annual, pre-monsoon, monsoon and post-monsoon rainfall, respectively. This situation remains unchanged when the gap statistic is applied using the K-means clustering and the extended gap statistic is applied using the Fuzzy C-means clustering. The identified clusters, in different situations, are presented for their spatial distribution which will, indeed, help policy makers to perceive the mechanism of the rainfall in Bangladesh. Particularly, while making policies for agriculture, flood management, reduction of bank erosion, etc., the findings reported in this study will be substantially worthwhile. Moreover, the characterization of rainfall distribution in different clusters will provide a clear and complete picture of the rainfall in Bangladesh. Indeed, stakeholders of precipitation will be benefited through the insight depicted in this study.

Though the three algorithms create an equal number of clusters in the case of different precipitation series, clusters are not identical in terms of cluster size and within-cluster homogeneity. In addition, none of the clusters obtained from Fuzzy C-means clustering is heterogeneous. In contrast, within-cluster heterogeneity is found in one cluster in case of agglomerative and K-means algorithms. Moreover, an ambiguity may arise if the two clusters for post-monsoon rainfall are formed using the K-means algorithm. Therefore, it seems reasonable to say that the Fuzzy C-means algorithm can be preferred for identifying the homogeneous clusters in different precipitation series of Bangladesh over available alternatives.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[2] Y.D. Chen, G. Huang, Q. Shao, and C.-y. Xu, *Regional analysis of low flow using l-moments for Dongjiang Basin, South China*, Hydrolog. Sci. J. 51 (2006), pp. 1051–1064.

[3] F. Dikbas, M. Firat, A.C. Koc, and M. Gungor, *Classification of precipitation series using fuzzy cluster method*, Int. J. Climatol. 32 (2012), pp. 1596–1603.

[4] J.C. Dunn, *A fuzzy relative of the isodata process and its use in detecting compact well seperated clusters*, J. Cyber. 3 (1974), pp. 32–57.

[5] M.K. Goyal and V. Gupta, *Identification of homogeneous rainfall regimes in Northeast Region of India using fuzzy cluster analysis*, Water Resour. Manag. 28 (2014), pp. 4491–4511.

[6] J.A. Greenwood, J.M. Landwehr, N.C. Matalas, and J.R. Wallis, *Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form*, Water. Resour. Res. 15 (1979), pp. 1049–1054.

[7] N.B. Guttman, *The use of L-moments in the determination of regional precipitation climates*, J. Clim. 6 (1993), pp. 2309–2325.

[8] J.R.M. Hosking and J.R. Wallis, *Some statistics useful in regional frequency analysis*, J. Clim. 29 (1993), pp. 271–281.

[9] J.R.M. Hosking and J.R. Wallis, *Regional Frequency Analysis: An Approach Based on L-moments*, Cambridge University Press, Cambridge, 1997.

[10] F. Johnson and J. Green, *A comprehensive continent-wide regionalisation investigation for daily design rainfall*, J. Hydrol. Reg. Stud. 16 (2018), pp. 67–79.

[11] R.A. Johnson and D.W. Wicharn, *Applied Multivariate Statistical Analysis*, 6th ed., Pearson Prentice Hall, Upper Saddle River, NJ, 2007.

[12] A. Kulkarni and R.H. Kripalani, *Rainfall patterns over India: Classification with fuzzy c-means method*, Theor. Appl. Climatol. 59 (1998), pp. 137–146.

[13] J. Kyselážş, J. Picek, and R. Huth, *Formation of homogeneous regions for regional frequency analysis of extreme precipitation events in the Czech Republic*, Studia Geophysica et Geodaetica 51 (2007), pp. 327–344.

[14] H. Malekinezhad and A. Zare-Garizi, *Regional frequency analysis of daily rainfall extremes using L-moments approach*, Atmósfera 27 (2014), pp. 411–427.

[15] M. Meddi, H. Meddi, S. Toumi, and M. Mehaiguen, *Regionalization of rainfall in North-western Algeria*, Geographia Technica 17 (2013).

[16] R. Modarres, *Regional precipitation climates of Iran*, J. Hydrol. (NZ) 45 (2006), pp. 15.

[17] L.V. Noto and G. La Loggia, *Use of L-moments approach for regional flood frequency analysis in Sicily, Italy*, Water Res. Manag. 23 (2009), pp. 2207–2229.

[18] N.R. Pal and J.C. Bezdek, *On cluster validity for the fuzzy c-means model*, IEEE. Trans. Fuzzy. Syst. 3 (1995), pp. 370–379.

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

[20] T. Raziei, I. Bordi, and L. Pereira, *A precipitation-based regionalization for Western Iran and regional drought variability*, Hydrol. Earth Syst. Sci. 12 (2008), pp. 1309–1321.

[21] A. Sarhadi and M. Heydarizadeh, *Regional frequency analysis and spatial pattern characterization of dry spells in Iran*, Int. J. Climatol. 34 (2014), pp. 835–848.

[22] A.S. Shirin and R. Thomas, *Regionalization of rainfall in Kerala state*, Proc. Technol. 24 (2016), pp. 15–22.

[23] R. Tibshirani, G. Walther, and T. Hastie, *Estimating the number of clusters in a data set via the gap statistic*, J. R. Stat. Soc. 63 (2000), pp. 411–423.

[24] S. Yue, P. Wang, J. Wang, and T. Huang, *Extension of the gap statistics index to fuzzy clustering*, Soft. Comput. 17 (2013), pp. 1833–1846.

Appendix

Table A1. The membership degrees (in percentage) of the stations by FCM for clusters considered in case of Annual, Monsoon and Post-monsoon precipitations.

	Precipitation series							
	Annual			Monsoon			Post-monsoon	
Station	1	2	3	1	2	3	1	2
Dhaka	48	39	13	70	7	24	69	31
Mymensing	37	45	18	49	11	40	62	38
Faridpur	59	30	11	79	5	16	77	23
Madaripur	53	35	12	68	7	25	74	26
Srimongal	33	47	20	69	6	24	66	34
Sylhet	21	27	52	12	63	25	51	49
Bogra	58	30	12	70	7	23	85	15
Ishurdi	54	31	14	73	7	20	82	18
Rajshahi	54	31	15	74	7	19	84	16
Dinajpur	47	38	15	44	12	43	77	23
Rangpur	36	44	19	31	14	55	77	23
Jessore	58	30	12	78	5	17	84	16
Khulna	60	29	11	76	5	18	81	19
Satkhira	58	30	12	76	6	18	84	16
Barisal	50	38	12	64	7	29	51	49
Bhola	38	44	19	43	13	44	54	46
Khepupara	24	39	38	15	31	54	26	74
Patuakhali	27	41	32	18	28	54	37	63
Chandpur	39	43	19	47	12	41	64	36
Patenga	21	32	47	14	45	41	28	72
Comilla	49	39	12	73	6	21	75	25
Coxsbazar	17	23	60	9	72	19	18	82
Feni	21	32	47	15	41	44	27	73
Hatia	24	33	43	18	44	38	31	69
Kutubdia	22	31	46	13	56	31	27	73
Maijdeecourt	18	29	53	12	53	35	33	67
Rangamati	29	43	28	26	18	56	32	68
Sandwip	20	27	52	13	62	26	20	80
Sitakunda	20	28	52	13	55	32	24	76
Teknaf	20	26	53	15	60	25	24	76