تحقیقات منابع آب ایران
Iran-Water Resources Research

سال ششم، شماره ۳، پائیز ۱۳۸۹
Volume 6, No. 3, Fall 2010 (IR-WRR)
1-12

# ناحیه بندی حوضه‌های آبریز خراسان با استفاده از تحلیل خوشه‌ای هیبرید

بیژن قهرمان [۱*]، حمیرا شامکوبیان [۲]
و کامران داوری [۳]

## چکیده

به علت کمبود آمار و اطلاعات همیشه امکان استفاده از تحلیل فراوانی مکانی جهت تخمین چندک‌های سیلاب وجود ندارد. از آن‌جاکه استفاده از یک روش واحد برای ناحیه‌ای کردن معمولاً نتایج قابل قبولی را به دست نمی‌دهد، لذا معمولاً چندین روش منطقه‌ای به‌طور توأم مورد استفاده قرار می‌گیرد. در این مطالعه سه الگوریتم خوشه‌ای هیبرید که هر یک به طور جداگانه فرایند خوشه‌ای کردن را برای تعیین نواحی مشابه به کار می‌برند، مورد بررسی قرار گرفت. از الگوریتم‌های خوشه‌ای سلسله مراتبی متراکمی استفاده شد. الگوریتم‌های خوشه‌ای مورد استفاده شامل پیوند تکی، پیوند کامل و Ward، و الگوریتم خوشه‌ای تفکیکی شامل الگوریتم K-means است. تأثیر تحلیل خوشه‌ای هیبرید در ناحیه‌ای کردن با استفاده از آمار و اطلاعات ۶۸ حوضه آبریز استان‌های خراسان مورد بررسی قرار گرفت. همچنین چهار شاخص آزمون خوشه‌ای شامل ضریب کوفنتیک، عرض سیلهوت متوسط، شاخص Dunn و Davies-Bouldin جهت تعیین تعداد بهینه خوشه‌ها مورد استفاده واقع گردید. تحلیل خوشه‌ای هیبرید در حداقل‌سازی تلاش لازم جهت نیل به نواحی همگن مفید و مؤثر بود. نهایتاً هیبرید الگوریتم Ward و K-means برای استفاده در ناحیه‌ای کردن پیشنهاد گردید. چهار ناحیه همگن تشخیص داده شد.

***کلمات کلیدی:*** ناحیه بندی، تحلیل فراوانی سیلاب، گشتاورهای خطی، خوشه بندی.

تاریخ دریافت مقاله: ۳۰ مهر ۱۳۸۷
تاریخ پذیرش مقاله: ۲۸ مهر ۱۳۸۹

۱- استاد دانشکده کشاورزی، دانشگاه فردوسی مشهد، مشهد، ایران
۲- کارشناسی ارشد آبیاری، آب منطقه‌ای خراسان رضوی، مشهد، ایران
۳- دانشیار آبیاری،دانشگاه فردوسی مشهد، مشهد، ایران
*- نویسنده مسئول

# Regionalization of Khorasan Watersheds by Hybrid-Cluster Analysis

*B. Ghahraman [1*], H. Shamkooian [2] and K. Davary [3]*

## Abstract

Because of the scarcity of flood data, it is not always possible to use at-site frequency analysis for flood quantiles estimations. No single procedure has a global acceptance in regionalization. In this paper, three hybrid-clustering algorithms are investigated. Each of these algorithms use the partitional clustering procedure to identify groups of similar catchments by refining the clusters derived from agglomerative hierarchical clustering algorithms. Their effectiveness in regionalization are then compared. The hierarchical clustering algorithms used are single linkage, complete linkage, and Ward's algorithms. The partitional clustering algorithm used is the K-means algorithm. The effectiveness of the hybrid-cluster analysis in regionalization is investigated using data from 68 watersheds in former Khorasan Province, IRAN (now separated as three Provinces). Further, four cluster validity indices, namely cophenetic correlation coefficient, average silhouette width, Dunn's index, and Davies–Bouldin index are tested to determine their effectiveness in identifying optimal partition provided by the clustering algorithms. The hybrid-cluster analysis is found to be useful in minimizing the effort needed to identify homogeneous regions. The hybrid of Ward's and K-means algorithms is recommended for use. Four homogeneous zones were detected.

***Keywords:*** Regionalization, Flood frequency analysis, L-moments, Cluster analysis.

1 – Professor, College of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran, Email: bijangh@um.ac.ir
2- Senior Irrigation Specialist, Khorasan Razavi Water Authority, Mashhad, Iran
3- Associate Professor, College of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran
*- Corresponding Author

## 1-Introduction

Due to the scarcity of flood data, it is not always possible to use at-site frequency analysis for flood quantiles estimations. To contend with this problem, hydrologists use data from the nearby watersheds with similar flood producing mechanism. Such group of watersheds with similar flood responses constitute a homogeneous region. The procedure of identifying homogeneous regions is traditionally referred to as regionalization. Regional flood-frequency analysis (RFFA) is widely used for estimation of flood quantiles in the design and operation of water resources systems, land use planning and management, and flood insurance assessment and protection of populated areas. The RFFA is recommended for estimation of flood quantiles at sites with record length less than 5T, where T is the return period of interest (Reed et al., 1999).

Recently, increasing awareness on the use of hydro-climatic data has prompted several agencies to work on databanks related to the hydrological processes. Besides, in regionalization studies, there is a need to develop potential approaches that are useful to identify and interpret patterns inherent in hydrologic data. For this task, clustering algorithms is a promising approach which are found effective in recognizing the distribution of patterns in both large and small data sets. Introductory material on cluster analysis is found in Kaufman and Rousseeuw (1990), Everitt (1993), Gordon (1999) and others.

There is still no generally agreed upon catchment classification system. Such a classification framework should provide a mapping of landscape form and hydro-climatic conditions on the catchment's function (including partition, storage, and release of water), while explicitly accounting for uncertainty and for variability at multiple temporal and spatial scales. This framework should provide an organizing principle, create a common language, guide modeling and measurement efforts, and provide constraints on predictions in ungauged basins, as well as on estimates of environmental change impacts. Wagner et al. (2007) (i) reviewed existing approaches to define hydrologic similarity and catchment classification; (ii) discussed outstanding components or characteristics that should be included in a classification scheme; and (iii) provided a basic framework for catchment classification as a starting point for further analysis. Possible merits to describe form, hydro-climate, and function were suggested and discussed. Wagner et al. (2007) raised open questions such as: How can we best represent characteristics of form and hydro-climatic conditions? How does this representation change with spatial and temporal scale? What functions (partition, storage, and release) are relevant at what spatial and temporal scale? At what scale do internal structure and heterogeneity become important and need to be considered.

As discussed above, there are numerous methods available in literature for delineation of the homogeneous regions. Shu and Burn (2004) developed a couple of methods using fuzzy expert system (FES) to derive an objective similarity measure between catchments. In their research, the performance of the FES was improved by tuning of the membership function of the fuzzy sets using a genetic algorithm. Abida and Ellouze (2006) adopted a method on the shape of the empirical Cumulative Distribution Function (CDF) and similarities of physiographic and climatic characteristics for hydrologic delineation of homogeneous regions in Tunisia. To facilitate estimation of streamflow characteristics at an un-gauged site, hydrologists often define a region of influence (ROI) containing gauged sites hydrologically similar to the estimation site (Burn, 1990). This region can be defined either in geographic space or in the space of the variables that are used to predict streamflow (predictor variables). These approaches are complementary, and a combination of the two may be superior to either. So, Eng et al. (2007) proposed a hybrid region-of-influence (HRoI) regression method that combines the two approaches. To identify homogeneous region(s), Atiem and Harmancioglu (2006) redefined the region by dividing it into sub-regions. This was done into two steps: (a) removing some sites from the region and using a completely different assignment of sites to region(s) by re-adding site(s) to the identified homogeneous sub-region(s); and (b) adding a cluster feature to the regional sites.

The above methods are, however, either too subjective (the last one) or too complicated. The classic approach to the problem is routinely done by cluster analysis.

A cluster consists of one or more feature vectors. A feature vector (also referred to as 'data vector' or 'object') comprises of several attributes or variables. In hydrology, the attributes that have been used for RFFA include: (i) physiographic catchment characteristics such as drainage area, average basin slope, main stream slope, stream length, stream density; storage index, soil type index such as infiltration potential, runoff coefficient or effective mean soil moisture deficit, fraction of the basin covered by lakes, reservoirs or swamps; (ii) geographical location attributes such as latitude, longitude and altitude of catchment centroid; (iii) a measure of basin response time such as basin lag or time-to-peak; (iv) meteorological factors such as storm direction, mean annual rainfall, precipitation intensities; and (v) at-site flood statistics. Selection of a

suitable set for RFFA, is however an important key, since it may drastically change the homogeneous zones. Although a subjective approach is still common in hydrology (e.g. Rao and Srinivas, 2006), more structured approaches are present (e.g. Dinpashoh et al., 2004 by using factor analysis; and Lin et al., 2005 by using principal component analysis).

Clustering is a process by which a set of feature vectors is divided into clusters or groups such that the feature vectors within a cluster are as similar as possible and the feature vectors of different clusters are as dissimilar as possible. Clustering algorithms can be broadly classified into two categories: Hierarchical clustering and Partitional clustering.

Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones (agglomerative), or by splitting larger clusters to smaller ones (divisive). Burn et al. (1997) used the agglomerative hierarchical clustering algorithm for the regionalization of watersheds in Canada.

Partitional clustering procedures attempt to recover the natural grouping present in the data through a single partition. Examples of this class of algorithms include K-means (MacQueen, 1967), K-median, K-modes and K-medoids algorithms (KMA). In hydrology, K-means algorithm and its variations for RFFA have been used by Wiltshire (1986), Burn (1989), Bhaskar and O'Connor (1989), and Burn and Goel (2000).

While the hierarchical clustering procedures are not influenced by initialization and local minima, the partitional clustering procedures are influenced by initial guesses (number of clusters, cluster centers, etc.). The partitional clustering procedures are dynamic in the sense that feature vectors can move from one cluster to another to minimize the objective function. In contrast, the feature vectors committed to a cluster in the early stages cannot move to another in hierarchical clustering procedures.

## 2. Objectives

The basic objective of research presented in this paper is to investigate the potential of hybrid clustering methods in regionalization of watersheds for flood-frequency analysis.

## 3. Data and Method

Three hybrid-cluster algorithms, which are a blend of agglomerative hierarchical and partitional clustering procedures, are presented. The hierarchical clustering algorithms considered for hybridization are single linkage, complete linkage, and Ward's algorithms. The partitional clustering algorithm used is the K-means algorithm. The performance of the hybrid-cluster algorithms is evaluated using annual maximum flow data from watersheds in Khorasan Provinces. Further, four cluster validity indices, namely cophenetic correlation coefficient, average silhouette width, Dunn's index, and Davies–Bouldin index are tested to determine the effectiveness of these indices in identifying optimal partition provided by the clustering algorithms.

The hybrid-clustering algorithm for regionalization of watersheds uses Kmeans algorithm (a partitional clustering algorithm) to identify groups of homogeneous catchments by refining the clusters derived from agglomerative hierarchical clustering algorithm.

### 3.1. Hybrid algorithm

Let $\mathbf{X} = \{\mathbf{x}_i / i = 1, \cdots, N\}$ denote a set of $N$ feature vectors in m-dimensional attribute space (i.e. $\mathbf{x}_i = [x_{i1}, \cdots, x_{im}] \in R^m$), each of which characterizes one of the $N$ sites. Further, let $\mathbf{y}_i$ denote the $i$th rescaled feature vector in the m-dimensional attribute space

$\mathbf{y}_i = [y_{i1}, \cdots, y_{im}] \in \Re^m$ obtained by rescaling $\mathbf{x}_i$ using Eq. (1)

$$y_{ij} = \frac{w_j}{\sigma_j}\left[f\left(x_{ij}\right)\right] \qquad j = 1, \cdots, m \tag{1}$$

In Eq. (1), $f(\cdot)$ represents the transformation function; $x_{ij}$ denotes the value of attribute $j$ in the m-dimensional feature vector $\mathbf{x}_i$; $y_{ij}$ denotes the rescaled value of $x_{ij}$; $w_{ij}$ is the weight assigned to attribute $j$; $\sigma_j$ is the standard deviation of attribute $j$.

The K clusters formed in the step '*N–K*' of an agglomerative hierarchical clustering algorithm are used to initialize the K-means algorithm (Hartigan and Wong, 1979). The K-means algorithm (KMA) is an iterative procedure in which the feature vectors move from one cluster to another to minimize the value of objective function, F, defined in Eq. (2)

$$F = \sum_{k=1}^{K}\sum_{j=1}^{m}\sum_{i=1}^{N_k} d^2\left(y_{ij}^k - y_{\bullet j}^k\right) \tag{2}$$

In Eq. (2), $K$ denotes the number of clusters, $N_k$ represents the number of feature vectors in cluster $k$;

$y_{ij}^{k}$ denotes the rescaled value of attribute *j* in the feature vector *i* assigned to cluster *k*; $y_{\bullet j}^{k}$ is the mean value of feature *j* for cluster *k* (Eq. (3))

$$y_{\bullet j}^{k} = N_{k}^{-1} \sum_{i=1}^{N_k} y_{ij}^{k} \quad (3)$$

By minimizing *F* in Eq. (2), the distance of each feature vector from the center of the cluster to which it belongs is minimized. $d(\cdot)$ denotes an appropriate distance measure such as Euclidean which is suitable for clusters with spherical shape.

The optimal value attained by the objective function, *F*, depends on cluster centers used to initialize the KMA. There is no single procedure of initializing the cluster centers to yield a global minimum value for the objective function, F. Several methods of initialization are in use. Wiltshire (1986) randomly partitioned data to initiate the clustering algorithm. Bhaskar and O'Connor (1989) considered initial cluster centers as feature vectors that are separated by at least a specified minimum distance. Burn (1989) suggested choosing K of the N feature vectors as the starting centroids to ensure that each cluster has at least one member (Burn, 1989). In the present study, results from the hierarchical clustering algorithms namely, single linkage, complete linkage and Ward's algorithm are used to provide initial cluster centers for the KMA.

Every feature vector is assigned to a cluster center that is nearest to it among the K clusters. After assigning the feature vectors to the K-cluster centers, the center of each of the K-clusters is updated and the value of the objective function, F, is computed. This completes an iteration of K-means algorithm. The procedure of assigning feature vectors to nearest cluster centers and updating the cluster centers is repeated in each of the subsequent iterations. The algorithm is stopped at a point when change in the value of objective function between two successive iterations becomes sufficiently small.

### 3.2. Single linkage and complete linkage algorithms

The algorithms begin with *N* singleton clusters each comprising a rescaled feature vector. Among the *N* singleton clusters, two closest clusters $\mathbf{y}_i$ and $\mathbf{y}_j$ are identified and merged to form a new cluster $[\mathbf{y}_i, \mathbf{y}_j]$. In the single linkage algorithm the distance between the new cluster $[\mathbf{y}_i, \mathbf{y}_j]$ and any other singleton cluster $\mathbf{y}_k$ is the smaller of the distances between $\mathbf{y}_i$ and $\mathbf{y}_k$ or $\mathbf{y}_j$ and $\mathbf{y}_k$. On the other hand, in complete linkage algorithm the distance between the new cluster $[\mathbf{y}_i, \mathbf{y}_j]$ and any other singleton cluster $\mathbf{y}_k$ is the greater of the distances between $\mathbf{y}_i$ and $\mathbf{y}_k$ or $\mathbf{y}_j$ and $\mathbf{y}_k$.

At each step, the two closest clusters are identified and merged. As a consequence, the number of available clusters decreases by one with each additional step. The algorithms are terminated at the step when the number of clusters is equal to the specified value *K*.

### 3.3. Ward's algorithm

The objective function of Ward's algorithm, W (Ward, 1963) minimizes the sum of squares of deviations of the feature vectors from the centroid of their respective clusters (Eq. (4))

$$W = K \sum_{k=1}^{k} \sum_{j=1}^{m} \sum_{i=1}^{N_k} \left( y_{ij}^{k} - y_{\bullet j}^{k} \right)^2 \quad (4)$$

The Ward's algorithm starts with singleton clusters. At this point the cluster centers are the same as feature vectors. Therefore, the value of objective function is zero. At each step in the analysis, union of every possible pair of clusters is considered and the two clusters whose fusion give the smallest increase in W are merged.

Ward's algorithm is good at recovering cluster structure and it tends to form spherical clusters of equal size. This characteristic of the Ward's algorithm makes it useful in identification of homogeneous regions for regionalization (Hosking and Wallis, 1997).

### 3.4. Cluster validity

Cluster validity indices have been extensively used to determine optimal number of clusters (*K*) in a data set (Everitt, 1993; Halkidi et al., 2001). In this study, four cluster validity indices, namely cophenetic correlation coefficient (Sokal and Rohlf, 1962), average silhouesft6ttte width (Rousseeuw, 1987), Dunn's index (Dunn, 1973), and Davies–Bouldin index (Davies and Bouldin, 79) are tested to determine their effectiveness in identifying optimal partition provided by the clustering algorithms for RFFA.

The cophenetic correlation coefficient, abbreviated as CPCC by Farris (1969), is a validity measure for hierarchical clustering algorithms. A hierarchical clustering process can be represented as a nested sequence or tree, called dendrogram, which shows how the clusters that are formed at the various steps of the process are related. The CPCC is used to measure how

well the hierarchical structure from the dendrogram represents in two dimensions the multidimensional relationships within input data. The CPCC is defined as the correlation between the M=N(N-1)/2, original pairwise dissimilarities (proximity) between the feature vectors, and their cophenetic dissimilarities from the dendrogram. The cophenetic dissimilarity, $c_{ij}$, between two feature vectors *i* and *j* is the intercluster distance at which the two feature vectors are first merged in the same cluster.

$$\mathrm{CPCC} = \frac{(1/\mathrm{M})\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}c_{ij} - \mu_p\mu_c}{\sqrt{\left[\frac{1}{\mathrm{M}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 - \mu_p^2\right]\left[\frac{1}{\mathrm{M}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} c_{ij}^2 - \mu_c^2\right]}} \quad (5)$$

In Eq. (5), $\mu_p$ and $\mu_c$ are the means of elements in proximity and cophenetic matrices respectively, whereas $d_{ij}$ and $c_{ij}$ are respectively the (*i*,*j*)th elements of proximity and cophenetic matrices. The concordance between the input data and the dendrogram is close if value of the index is close to 1.0. A high value for CPCC is regarded as a measure of successful classification. A value of 0.8 or higher indicates that the dendrogram does not greatly distort the original structure in the input data (Romesburg, 1984).

The silhouette width (Rousseeuw, 1987) for a feature vector is a measure of how similar that feature vector is to feature vectors in its own cluster compared to feature vectors in other clusters. The silhouette width *s(i)* for the *i*th feature vector in a cluster *k* is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \quad (6)$$

In Eq. (6), $a(i)$ is the average distance of the *i*th feature vector to all other feature vectors in the cluster *k*; $b(i)$ is the minimum average distance of the *i*th feature vector to all the feature vectors in another cluster *j* ($j = 1, \cdots, K \quad j \neq k$). From this formula it follows that $-1 \le s(i) \le 1$.

If *s(i)* is close to 1, we may infer that the *i*th feature vector has been assigned to an appropriate cluster. On the other hand, when *s(i)* is close to -1, we may conclude that the *i*th feature vector has been misclassified. When *s(i)* is approximately zero, it indicates that the *i*th feature vector lies equally far away from the two clusters. For the given *K* clusters, the overall average silhouette width is the average of the silhouette widths for all the feature vectors in the dataset. The partition with the maximum overall average silhouette width is taken as the optimal partition.

Dunn's index (Dunn, 1973) and Davies– Bouldin index (Davies and Bouldin, 1979) are widely recognized for their ability to identify sets of clusters that are compact and well separated. The Davies– Bouldin index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation.

Suppose that the given set of N-feature vectors (in m-dimensional space) has been partitioned into K clusters $\{C_1, C_2, \cdots, C_k\}$ such that cluster $C_k$ has $N_k$ feature vectors and each feature vector is in exactly one cluster. The scatter within the *k*th cluster, $S_{k,q}$, is computed using Eq. (7) and the Minkowski distance of order *t* between the centroids that characterize clusters $C_j$ and $C_k$ is defined by Eq. (8).

$$S_{k,q} = \left(\frac{1}{N_k}\sum_{\mathbf{y}_i \in C_k} \|\mathbf{y}_i - \mathbf{z}_k\|_2^q\right)^{1/q} \quad (7)$$

$$d_{jk,t} = \|\mathbf{z}_j - \mathbf{z}_k\|_t \quad (8)$$

where $\mathbf{z}_k$ represents the center of cluster *k* and $S_{k,q}$ is the *q*th root of the *q*th moment of the points in cluster *k* with respect to their means. In this work the first moment (i.e. *q*=1) and Minkowski distance of order 2 (i.e. *t*=2) which are commonly adopted by practitioners (Rao and Srinivas, 2006), have been used. The Davies–Bouldin index is computed using Eq. (9). A small value for DB indicates good partition, which corresponds to compact clusters with their centers far apart.

$$DB = \frac{1}{K}\sum_{k=1}^{K}\max_{j,j\neq k}\left\{\frac{S_{k,q} + S_{j,q}}{d_{jk,t}}\right\} \quad (9)$$

Dunn's index is computed using Eq. (10)

$$\mathrm{D} = \min_{1\le i\le K}\left\{\min_{1\le j\le K, j\neq i}\left\{\frac{\delta(\mathrm{C_i}, \mathrm{C_j})}{\max_{1\le k\le K}\Delta(\mathrm{C_k})}\right\}\right\} \quad (10)$$

where $\delta(C_i, C_j)$ denotes the distance between clusters $C_i$ and $C_j$ (intercluster distance) computed by Eq. (11); $\Delta(C_k)$ represents the intracluster distance of cluster $C_k$ defined by Eq. (12). The value of *K* for which *D* is maximized is taken as the optimal number of clusters.

$$\delta(C_i, C_j) = \max_{\mathbf{y}_i \in C_i, \mathbf{y}_j \in C_j} \{ d(\mathbf{y}_i, \mathbf{y}_j) \} \quad (11)$$

$$\Delta(C_k) = \max_{y_i, y_j \in C_k} \{ d(\mathbf{y}_i, \mathbf{y}_j) \} \quad (12)$$

where $d(\mathbf{y}_i, \mathbf{y}_j)$ is the distance between rescaled feature vectors $\mathbf{y}_i$ and $\mathbf{y}_j$.

### 3.5. Regional homogeneity test

Following Hosking and Wallis approach (1993, 1997), the heterogeneity of the set of plausible regions obtained from the cluster analysis is assessed in this study using the homogeneity tests. They use the advantages offered by sampling properties of L-moment ratios. Three heterogeneity measures $H_1$, $H_2$, and $H_3$ suggested by Hosking and Wallis (1993) are used in the present study;

(i) A heterogeneity measure (HM) based on L-coefficient of variation (L-CV):

$$H_1 = \frac{(V - \mu_v)}{\sigma_v} \quad (13)$$

(ii) A measure based on L-CV and L-skewness:

$$H_2 = \frac{(V_2 - \mu_{v2})}{\sigma_{v2}} \quad (14)$$

(iii) A measure based on L-skewness and L-kurtosis:

$$H_3 = \frac{(V_3 - \mu_{v3})}{\sigma_{v3}} \quad (15)$$

In Eq. (13), *V* denotes the weighted standard deviation of the at-site sample L-CVs; In Eq. (14), $V_2$ represents the weighted average distance from the site to the group weighted mean in the two dimensional space of L-CV and L-skewness; In Eq. (15), $V_3$ refers to the weighted average distance from the site to the group weighted mean in the two dimensional space of L-skewness and L-kurtosis; $\mu_v$, $\mu_{v2}$ and $\mu_{v3}$ denote the means of the $N_{sim}$ values of *V*, $V_2$ and $V_3$ respectively; $\sigma_v$, $\sigma_{v2}$ and $\sigma_{v3}$ represent the standard deviations of the $N_{sim}$ values of *V*, $V_2$ and $V_3$ respectively. In the present study, the value of $N_{sim}$ is chosen as 500.

A straightforward computation of *V*, $V_2$, and $V_3$ can be found in Hosking and Wallis (1997; pp 63-67). A region is regarded as 'acceptably homogeneous' if the heterogeneity measure $HM < 1$, 'possibly homogeneous' if $1 < HM < 2$, and 'definitely heterogeneous' if $HM > 2$. Further details on the homogeneity test are found in Hosking and Wallis (1997).

### 3.6. Discordancy measure

The regions obtained from the cluster analysis are adjusted following the suggestions of Hosking and Wallis (1997) to improve their homogeneity. Adjustment of a region involves exclusion of those sites from the region, which are grossly discordant with respect to other sites in it. In this study the discordancy measure proposed by Hosking and Wallis (1993) is used to identify those sites in a cluster that are grossly discordant. For details, consult with Hosking and Wallis (1997; pp 45-49).

Hosking and Wallis (1993; 1997) provide critical values for the discordancy measure to declare a site unusual. In many instances the catchment discordancy may arise out of sampling variability. Therefore, the data at all the sites with the largest discordancy values should be closely scrutinized, regardless of the magnitude of the discordancy values, before deciding whether the sites are discordant.

### 3.7. Data used in the study

The area of three Khorasan provinces is about 313000 $km^2$ (see Fig.1). These provinces have an arid and semi-arid climate. The highest elevation in this area is the Binalood Peak, with a height of about 3300 masl. The lowest point in the area is the downstream of Sarakhs plain with an altitude of about 250 masl. A climatic diversity can be observed in the area due to deserts and high mountains. This results in a highly temporal and spatial rainfall distribution. In general rainfall decreases from north-west to south and south-east. Sixty eight watersheds in Khorasan Provinces were considered in this study.

The data was collected from 3 Khorasan Water Authorities. Attributes such as height, latitude, and longitude of the stations, area, form factor, slope, and length of the main river of each catchment is used. The full data is reported elsewhere (Shamkooian et al., 2009). The location attributes, latitude and longitude, are included in the feature vector to identify regions that are geographically contiguous. Each of the nine attributes were standardized by Eq. (1). Equal weight was assigned to all attributes, implying equal importance to all features.

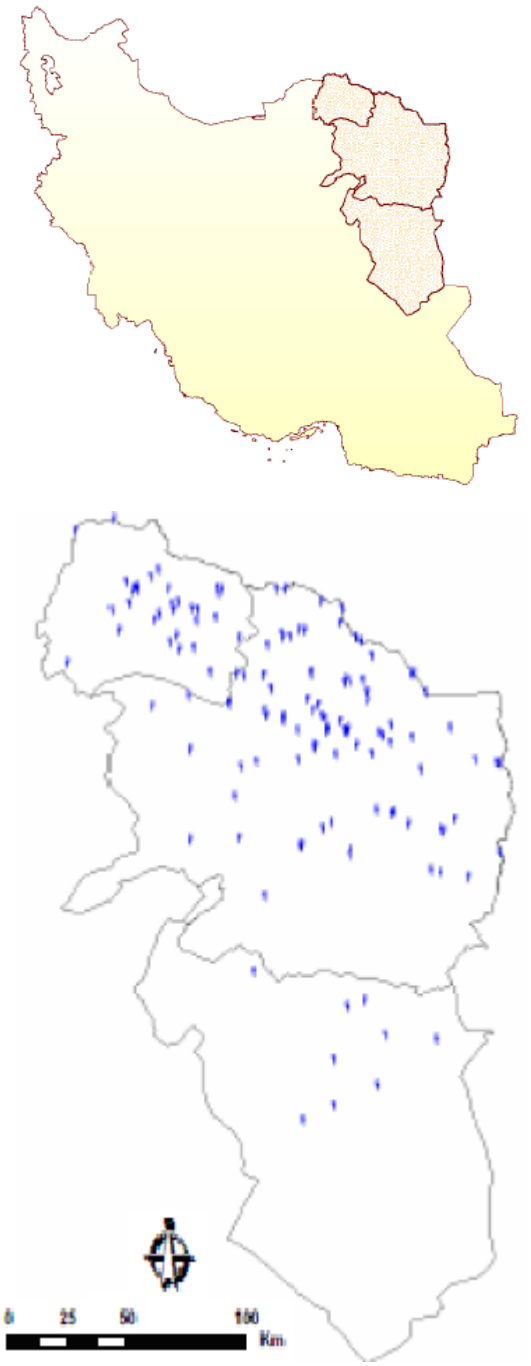


**Figure 1- Location of Khorasan Provinces (North, Razavi, and South) in Iran (top), and location of hydrometric stations in Khorasan, Iran (bottom).**

## 4. Results and discussion

### 4.1. Cluster analysis

The K clusters obtained from the agglomerative hierarchical clustering constituent of the hybrid model after 68-K merges are used to initiate the K-means algorithm. The value of the objective function (Eq. 2), in general, decreases with increase in the number of clusters. It has maximum value when all the feature vectors are lumped into a single cluster and has a minimum value of zero when K equals the number of feature vectors considered for the cluster analysis.

Variation in the optimal value of objective function for K ranging from 1 to 10 is presented in Table 1 for each of the three hybrid-cluster models and their respective hierarchical clustering constituents, namely single linkage, complete linkage and Ward's hierarchical clustering algorithms.

Of the three hierarchical clustering constituents of the hybrid model, Ward's algorithm gave the minimum value for the objective function (Table 1). Sum of F for all *K*s from 1 to 10 corresponding to SL, CL, and W are 4021, 3173, and 3070, respectively. As expected, the performance of each of the three hybrid-clustering algorithms in minimizing the objective function is better than that of the hierarchical clustering algorithm used to initialize them. In particular, the blend of Ward's algorithm and KMA gave the minimum value of objective function for values of *K* in the range of 3 to 10, 1 and 2 as exceptions. In essence, the overall performance of hybrid models in minimizing the objective function is better than that of the hierarchical and the *K*-means clustering model considered separately.

In a hybrid-clustering algorithm, the hierarchical clustering model is expected to provide more meaningful initial values to the KMA, so that the KMA provides better and meaningful output. However, one cannot guarantee a better output from KMA by hybrid-clustering. This point is evident from the results presented in Table 1 in selecting *K* equal to 5 and 6, for which the KMA initialized with the random K feature vectors in the data set, provided the smallest value for the objective function, yet the differences are 0.3 and 1.2% for *K*=5 and 6, respectively. In other words, one can always consider hybrid-clustering as a "*potential*" option to initialize *K*-means algorithm; however it is not always the best choice.

Before hybridization, the clusters obtained from single-linkage, complete-linkage, and Ward's clustering algorithms were examined. The clusters obtained from single linkage algorithm consisted of one large cluster and several small clusters, indicating that the algorithm is not suitable for regionalization. But the clusters obtained from Ward's algorithm are well separated. The clusters obtained from the hybrid of Ward's algorithm and KMA are found to be very similar to those resulting from Ward's algorithm (the difference is around 4%). In contrast, the clusters resulting from

**Table 1- Minimization of objective function (dimensionless) using hierarchical, K-means, and hybrid-clutering models.[a]**

| K | Hierarchical | | | K-means (KM) | Hybrid-clustering | | |
|---|---|---|---|---|---|---|---|
| | SL | CL | W | | SL+KM | CL+KM | W+KM |
| 1 | 603.00 | 603.00 | 603.00 | 603.00 | 603.00 | 603.00 | 603.00 |
| 2 | 482.19 | 449.25 | 449.25 | 449.25 | 449.25 | 449.25 | 449.25 |
| 3 | 462.62 | 386.03 | 375.62 | 360.44 | 374.02 | 365.72 | 360.44 |
| 4 | 430.44 | 318.46 | 310.11 | 298.05 | 296.93 | 297.28 | 297.28 |
| 5 | 404.68 | 293.41 | 278.10 | 268.92 | 273.42 | 269.43 | 269.78 |
| 6 | 354.05 | 276.60 | 254.20 | 241.63 | 251.01 | 248.67 | 244.58 |
| 7 | 347.81 | 253.83 | 231.43 | 223.74 | 230.17 | 225.81 | 221.21 |
| 8 | 332.80 | 221.69 | 209.31 | 199.67 | 205.72 | 207.73 | 199.29 |
| 9 | 317.75 | 190.44 | 188.23 | 186.73 | 185.50 | 187.94 | 176.69 |
| 10 | 285.91 | 180.01 | 170.63 | 164.04 | 178.36 | 165.30 | 162.54 |

[a] *K* represents number of clusters: SL denotes single linkage; CL refers to complete linkage; W denotes Ward's

single linkage algorithm are considerably modified by KMA (the difference is around 31%).

## 4.2. Cluster validity

The cluster validity indices, namely cophenetic correlation coefficient (CPCC), average silhouette width, Dunn's index (D) and Davies–Bouldin index (DB) are computed for the clusters obtained from the clustering methods using Eqs. (5)–(12) to determine optimal number of clusters in the dataset.

The CPCC is found to be considerably high for clusters obtained from single linkage algorithm. It is higher than in clusters obtained from complete linkage and Ward's algorithms (Table 2).

**Table 2- Cluster validity using Cophenetic correlation Coefficient (CPCC)**

| Algorithm | CPCC |
|---|---|
| Single linkage | 0.803 |
| Complete linkage | 0.655 |
| Ward's | 0.674 |

Following the definition of CPCC, an arguement may come up that the multi-dimensional relationship within the input data is represented better in the dendogram provided by single linkage algorithm than in the dendograms provided by complete linkage and Ward's algorithms. This is in contradiction to our earlier findings. Moreover, for the dataset considered herein, single linkage algorithm provides several singleton clusters and one very large cluster. This defeats the purpose of regionalization because such regions are highly heterogeneous.

The average silhouette width (ASW), which has a feasible range from -1 to +1, varied generally within a narrow range of 0.118 to 0.762 for the data over the variety of clustering options considered. The positive signs for all ASWs indicate that all feature vectors are assigned to appropriate clusters. The ASW is reasonably high for single linkage clustering with *K* in the range 2 to 6, for complete linkage clustering with *K* in the range 9 to 10, and for Ward's clustering with *K* in the range 7 to 8 (Table 3). However, these cases provide one large heterogeneous region (cluster) and very small regions, which due to heterogeneity are not suitable for regionalization. In general, the ASW of hybrid-clusters is marginally higher than that of the hierarchical clusters used to initialize the *K*means algorithm (Table 3), suggesting improvement in performance due to hybridization.

Among the hybrid- clustering models the ASW is found to be optimal, for the clusters obtained from the hybrid of single linkage and *K*-means algorithms with *K* equal to 8 (Table 3). But as stated above, adopting single linkage is not a good option. The ASW of the hybrid of Ward's and K-means with K equal to 4, is found to be maximum.

Fig. (2) supports that as K increases, the performances of both indicators D and DB increases. This is more pronounced up to K=6. As K is greater than 6, these indicators oscillate. So in the range of $K \leq 6$ it seems that L=4 is better for the hybrid of W+KM; D is maximum and DB is minimum. The Dunn's index and the Davies–Bouldin index also indicate the hybrid of Ward's and *K*-means algorithms to be the best (Fig. 2).

The heterogeneity measures of Hosking and Wallis (1993) described in Eqs. (13) to (15), were used to test if the clusters resulting from the Ward's and K-means hybrid-clustering for *K*=4 are statistically homogeneous. The results are presented in Table 4. All measures are negative which may be an indication that regions are cross correlated (see page 71 of Hosking

and Wallis -1997- for clarification). A number of statistical homogeneity tests are proposed in the scientific literature. It is documented that the inter-site correlation of floods is normally not negligible. The documents however does not specifically address the impact of cross-correlation on such statistical tests. Casterllarin et al. (2008) analyzed the effectiveness of a well-known homogeneity test for an inter-site cross-correlation through a series of Monte Carlo experiments. They proposed the following empirical corrector of the test based upon $H_1$ for providing an approximate indication in the presence of cross-correlation:

$$H_{1,adj} = H_1 + C \times \bar{\rho}^2 (R - 1) \quad (16)$$

**Table 3- Cluster validity using Silhouette width (dimensionless )—a comparison between hierarchical, K-means and hybrid-clustering models (K represents number of clusters; SL denotes single linkage; CL refers to complete linkage; W denotes Ward's)**

| K | Hierarchical | | | K-means (KM) | Hybrid-clustering | | |
|---|---|---|---|---|---|---|---|
| | SL | CL | W | | SL+KM | CL+KM | W+KM |
| 2 | 0.762 | 0.684 | 0.321 | 0.321 | 0.332 | 0.332 | 0.322 |
| 3 | 0.425 | 0.245 | 0.274 | 0.322 | 0.352 | 0.311 | 0.322 |
| 4 | 0.414 | 0.309 | 0.339 | 0.383 | 0.383 | 0.378 | 0.379 |
| 5 | 0.384 | 0.314 | 0.313 | 0.397 | 0.355 | 0.343 | 0.323 |
| 6 | 0.346 | 0.304 | 0.324 | 0.342 | 0.362 | 0.330 | 0.351 |
| 7 | 0.303 | 0.314 | 0.344 | 0.322 | 0.378 | 0.342 | 0.372 |
| 8 | 0.148 | 0.276 | 0.312 | 0.348 | 0.395 | 0.326 | 0.341 |
| 9 | 0.118 | 0.329 | 0.326 | 0.365 | 0.317 | 0.339 | 0.358 |
| 10 | 0.178 | 0.322 | 0.289 | 0.374 | 0.329 | 0.377 | 0.364 |

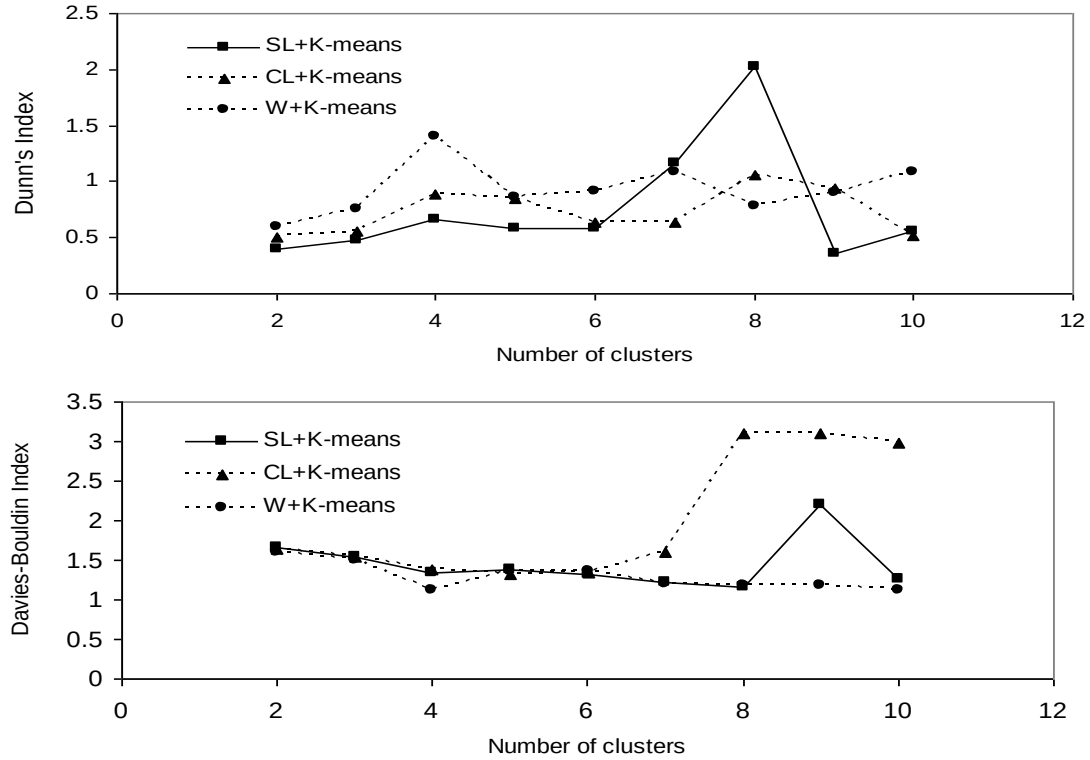


**Figure 2- Identification of optimal partition provided by the hybrid-clustering algorithms using Dunn's index and Davies–Bouldin index for the data set. The partition with the maximum value for Dunn's index and the minimum value for Davies–Boulding index is taken as the optimal partition.**

**Table 4- Heterogeneity measures for the Ward's and *K*-means hybrid-clustering for *K*=4**

| RN | NS | RS | Heterogeniety measure | | |
|---|---|---|---|---|---|
| | | | $H_1$ | $H_2$ | $H_3$ |
| 1 | 14 | 293 | -0.96 (-0.78)* | -0.61 | -0.34 |
| 2 | 5 | 142 | -0.99 (-0.96) | -0.15 | -0.23 |
| 3 | 19 | 381 | -0.89 (-0.61) | -0.17 | -0.26 |
| 4 | 30 | 725 | -0.98 (-0.80) | -0.19 | -0.08 |

RN=region number, NS=number of stations, RS=region size in station number (i.e. total record length).
* Numbers in parentheses are corresponding H values for uncorrelated regions (see Eq. 16).

where $H_{1,adj}$ is the adjusted value of the heterogeneity measure, $H_1$ is the value resulting from the homogeneity test, C is the empirical coefficient of the corrector that is assumed to be constant, $\bar{\rho}^2$ is the average squared correlation of concurrent flows, and R is the number of sequences in the region. In the present research, $\bar{\rho}^2$ is computed for the 4 regions as 0.11, 0.05, 0.14, and 0.08. Assuming that C=0.122 (Casterllarin et al., 2008), the values of $H_{1,adj}$ are reported in Table 4. By this correction, the results presented in Table 4 indicate an acceptable homogeneity in all regions.

In this study the discordancy measure proposed by Hosking and Wallis (1993) is used to identify those sites in a cluster that are grossly discordant (see Hosking and Wallis, 1997; pp 45-49). The values of the discordancy measure were in the range of 0.03-4.3. Compared to critical values (Hosking and Wallis, 1997), one station in region 3 and three stations in region 4 were found to be discordant (more clarification can be found in Shamkooian et al., 2009). These were ignored for RFFA studies. In the first and second region no sites were found to be discordant.

The location of cordant stations and their regions are presented in Fig. 3. This Figure supports that it is not possible to group the stations in closed boundaries to form separate clusters. The stations of one cluster are distributed in different locations. Such differentiation is not unusual in hydrology, however. Rao and Srinivas (2006) reported such distribution for FFA in Indiana, USA. A similar result is reported by Lin et al. (2005) for clustering autographic rain gages to study rainfall hyetographs in central Taiwan.

## 5. Summary, conclusions, and recommendations

Three hybrid-clustering algorithms, which are a blend of agglomerative hierarchical and partitional clustering procedures are investigated and used for regionalization of watersheds for flood-frequency analysis. The hierarchical clustering algorithms considered for hybridization are single linkage, complete linkage, and Ward's algorithms. The partitional clustering algorithm used is the K-means algorithm (KMA).

Of the three hybrid models presented, the combination of Ward's and *K*-means algorithms consistently provided good initial estimates of groups of watersheds.

In a hybrid-clustering algorithm, we expect the hierarchical clustering model to provide more meaningful initial values to the KMA, so that the KMA provides better and meaningful output. However, one cannot guarantee a better output from KMA by hybrid-clustering. In other words, one can always consider hybrid-clustering as a potential option to initialize K-means algorithm, however it is not always the best choice.

The effectiveness of four cluster validity indices, namely cophenetic correlation coefficient (CPCC), average silhouette width (ASW), Dunn's index, and Davies–Bouldin index are tested in identifying optimal partition provided by the clustering algorithms. The CPCC is found to be inefficient, whereas the ASW performed reasonably well. The Dunn's index and Davies–Bouldin index are found to be effective in identifying optimal partition, and is found to contain near-homogeneous clusters.

By these cluster validity indices, four clusters resulting from the Ward's and *K*-means hybrid -clustering were considered as optimal partition. The heterogeneity measures of Hosking and Wallis (1993) were used to test if these clusters are statistically homogeneous. The results indicated that all the regions are acceptably homogeneous. Finally these clusters can be used in regional flood frequency analysis. Using the discordancy test, one station in region 3 and three stations in region 4 were found to be discordant. Iin the first and second regions no sites were found to be discordant. By ignoring these stations we can use the regions for RFFA studies.

The presence of significant non-stationarity in a hydrologic time series as well as the climate change, cannot be ignored when estimating design values for future time horizons. There are so many researches published in literature dealing with such subject in

hydrology (i.e. Burn and Elnur, 2002; Cunderlik and Burn, 2003; Abdul Aziz and Burn, 2006; Sharif and Burn, 2006). However, none of them have dealt with the regionalization issue and there is still a gap which needs more consideration.

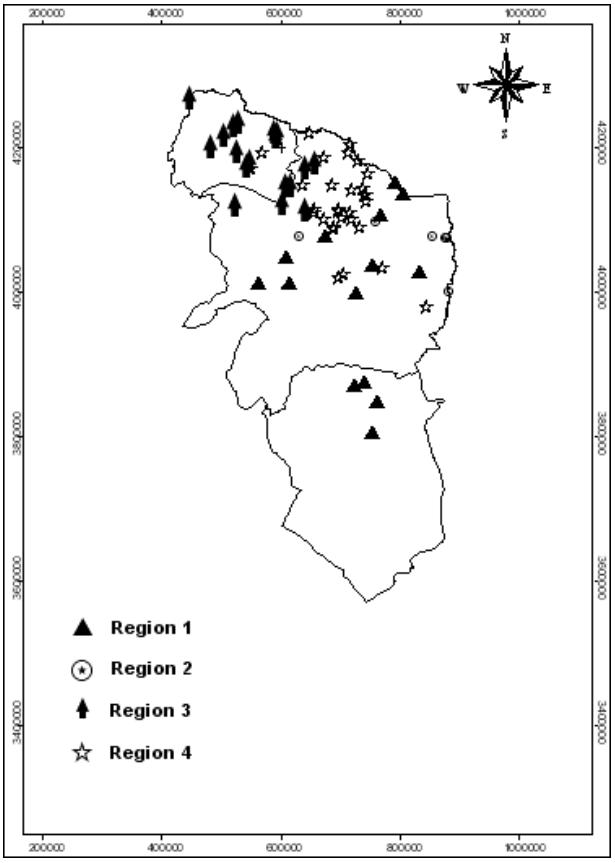


**Figure 3- Location of the regions defined by using the hybrid-cluster analysis. Three distinct parts are due to North Khorasan (top), Razavi Khorasan (middle), and South Khorasan (bottom), respectively.**

## List of abbreviations:

| | |
|---|---|
| ASW | average Silhoutte width |
| CL | complete linkage |
| CPCC | cophenetic correlation coefficient |
| D | Dunn's index |
| DB | Davies-Bouldin index |
| HM | heterogeneity measure |
| KMA | K-mean algorithm |
| RFFA | Regional flood frequency analysis |
| SL | single linkage |
| W | Ward |

## 6. References


Abdul Aziz, O.I., and Burn, D.H. (2006). "Trends and variability in the hydrological regime of the Mackenzie River Basain", *Journal of Hydrology*, 319, pp. 282-294.

Abida, H., and Ellouze, M. (2006). "Hydrological delineation of homogeneous regions in Tunisia", *Water Resources Management*, 20, pp. 961-977.

Atiem, I., and Harmancloglu, N.B. (2006). "Assessment of regional floods using L-moments approach: the case of the River Nile", *Water Resources Management*, 20, pp. 723-747.

Bhaskar, N.R., and O'Connor, C.A. (1989). "Comparison of method of residuals and cluster analysis for flood regionalization", *Journal of Water Resources Planning and Management*, 115 (6), pp. 793–808.

Burn, D.H. (1989). "Cluster analysis as applied to regional flood frequency", *Journal of Water Resources Planning and Management*, 115 (5), pp. 567–582.

Burn, D.H. 1990. "Evaluation of regional flood frequency analysis with a region of influence approach", *Warer Resources Research*, 26(10), pp. 2257-2265.

Burn, D.H., and Elnur, A.H. (2002). "Detection of hydrologic trends and variability", *Journal of Hydrology*, 255, pp. 107-122.

Burn, D.H., and Goel, N.K. (2000). "The formation of groups for regional flood frequency analysis", *Hydrological Sciences Journal*, 45 (1), pp. 97–112.

Burn, D.H., Zinji, Z. and Kowalchuk, M. (1997). "Regionalization of catchments for regional flood frequency analysis", *Journal of Hydrologic Engineering, ASCE*, 2(2), pp. 76-82.

Casterllarin, A., Burn, D.H., and Brath, A. (2008). "Homogeniety testing: how homogeneous do heterogeneous cross-correlated regions seem?", *Journal of Hydrology*, 360, pp. 67-76.

Cunderlik, J.M., and Burn, D.H. (2003). "Non-stationary pooled flood frequency analysis", *Journal of Hydrology*, 276, pp. 210-223.

Davies, D.L., and Bouldin, D.W. (1979). "A cluster separation measure". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1, pp. 224–227

Dinpashoh, Y., Fakheri-Fard, A., Moghaddam, M., Jahanbakhsh, S., and Mirnia, M. (2004). "Selection of variables for the purpose of regionalization of Iran's precipitation climate using multivariate methods", *Journal of Hydrology*, 297, pp. 109-123.

Dunn, J.C. (1973). "A fuzzy relative of the ISODATA process and its use in detecting compact well-

separated clusters". *Journal of Cybernetics*, 3, pp. 32–57.

Eng, K., Milly, P.C.D., and Tasker, G.D. (2007). "Flood regionalization: a hybrid geographic and predictor-variable region-of-influence regression method", *Journal of Hydrologic Engineering, ASCE*, 12(6), pp. 585-591.

Everitt, B. (1993). *Cluster Analysis*, Third ed. Halsted Press, New york, 280p.

Farris, J.S. (1969). "On the cophenetic correlation coefficient". *Systematic Zoology*, 18, pp. 279–285.

Gordon, A.D. (1999). *Classification*, Second ed. Chapman & Hall/CRC, London, 320p.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). "On clustering validation techniques". *Journal of Intelligent Information systems*, 17 (2/3), pp. 107–145.

Hartigan, J.A., and Wong, M.A. (1979). "Algorithm AS 136: a K-means clustering algorithm". *Applied Statistics*, 28, pp. 100–108.

Hosking, J.R.M., and Wallis, J.R. (1993). "Some statistics useful in regional frequency analysis". *Water Resources Research*, 29 (2), pp. 271–281 (Correction:*Water Resources Research* 31(1), pp. 251, 1995).

Hosking, J.R.M., and Wallis, J.R. (1997). *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, New York, USA., 224p.

Kaufman, L., and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York., 308p.

Lin, G.-F., Chen, L.-H., and Kao, S.-C. (2005). "Development of regional design hyetographs", *Hydrological Processes*, 19, pp. 937-946.

MacQueen, J. (1967). "Some methods for classification and analysis of multivariate observations". In: Le Cam, L.M., Neyman, J. (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. University of California Press, Berkeley, CA, pp. 281–297.

Rao, A.R., and Srinivas, V.V. (2006). "Regionalization of watersheds by hybrid-cluster analysis", *Journal of Hydrology*, 318, pp. 37-56.

Reed, D.W., Jakob, D., and Robson, A.J. (1999). *Selecting a pooling group*. In: Robson, A.J., Reed, D.W. (Eds.), *Statistical procedures for Flood Frequency Estimation*, Flood Estimation Handbook, vol. 3. Institute of Hydrology, Wallingford, UK (chapter 6).

Romesburg, H.C. (1984). *Cluster Analysis for Researchers*. Lifetime Learning Publications, Belmont, CA., 205p.

Rousseeuw, P.J. (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". *Journal of Computational and Applied Mathematics*, 20, pp. 53–65.

Shamkooian, H., Ghahraman, B., Davary, K., and Sarmad, M. (2009). "Flood frequency analysis using linear moments and flood index method in Khoranan provinces", *Journal of Water and Soil*, 23(1), pp. 31-43 (in Persian).

Sharif, M., and Burn, D.H. (2006). "Simulating climate change scenarios using an improved K-nearest neighbor model", *Journal of Hydrology*, 325, pp. 179-196.

Shu, C., and Burn, D.H. (2004). "Homogenous pooling delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement", *Journal of Hydrology*, pp. 291, 132.-149.

Sokal, R.R., and Rohlf, F.J. (1962). "The comparison of dendrograms by objective methods". *Taxon*, 11, pp. 33–40.

Wagner, T., Sivapalan, M., Troch, P., and Woods, R.. (2007). "Catchment classification and hydrologic similarity", *Geography Compass*, 1(4), pp. 901-931, doi: 10.1111/j.1749-8198.2007.00039.x.

Ward, Jr., J.H. (1963). "Hierarchical grouping to optimize an objective function". *Journal of American Statistical Association*, 58, pp. 236-244.

Wilshire, S.E. (1986). "Regional flood frequency analysis. II. Multivariate classification of drainage basins in Britain", *Hydrological Sciences Journal*, 31(3), pp. 335-346.