Assessment of some combinations of hard and fuzzy clustering techniques for regionalisation of catchments in Sefidroud basin
Ali Ahani and S. Saeid Mousavi Nadoushani

ABSTRACT
Cluster analysis methods are a type of well-known technique for regionalisation of catchments to perform regional flood frequency analysis. In this study, a fuzzy extension of hybrid clustering algorithms is evaluated. Self-organizing feature maps and four hierarchical clustering algorithms were used to provide the initial cluster centres for fuzzy c-means (FCM) algorithm. The hybrid approach was used for regionalisation of catchments in Sefidroud basin based on feature vectors including five catchment attributes: longitude and latitude, drainage area, runoff coefficient and mean annual precipitation. The results showed that according to the values of both the objective function and the cluster validity indices, the performances of FCM algorithm often was improved by using the proposed hybrid approach. Also, it was evident from the results that in the case of minimizing the objective function, the combination of Ward’s algorithm and FCM provided best results, but according to the cluster validity indices, other hybrid algorithms such as combinations of single linkage or complete linkage and FCM algorithm presented the most desirable results. In addition, according to the results, there are two well-defined homogeneous regions in Sefidroud basin identified by all the examined hybrid algorithms.

Key words | cluster analysis, flood frequency analysis, fuzzy c-means algorithm, L-moments, regionalisation

INTRODUCTION
In hydrology, estimation of the average return period of an event with specified magnitude is called frequency analysis (Hosking & Wallis 1997). Flood frequency analysis (FFA) methods are used to relate the magnitude of floods to their return periods (Rao & Srinivas 2008). In at-site FFA, site-specific hydrologic data is used to estimate flood quantiles at the target site. However, in some cases, the at-site FFA may not be useful to provide reliable flood estimates because of paucity of gauging stations and flood data in target sites. The regional flood frequency analysis (RFFA) methods can solve this problem. In these methods, a number of sites with similar flood generating mechanisms forms a group named homogeneous region. After that, a good-fitting frequency distribution is chosen as regional frequency distribution and then, the site-specific frequency distributions and quantile estimates can be provided for the interested sites.

Over the past five decades, several researchers studied the RFFA. One of the earliest and most widely used RFFA methods is the index-flood procedure of Dalrymple (1960). A suitable frequency distribution is fitted to the flood data gathered from all sites of a homogeneous region. This distribution is named regional frequency distribution. Based on a number of researches about ‘regional PWM (probability
weighted moments) algorithm, Hosking & Wallis (1997) concluded that a suitable regional analysis will yield useful and relatively accurate quantile estimates for many realistic applications. Hosking & Wallis (1997) also proposed an RFFA approach based on L-moments (Hosking 1990). The index-flood method and its L-moment based extension are based on an assumption of scale invariance. This assumption means that the frequency distributions of all sites in a homogeneous region are identical and only a site-specific factor named the index flood differs from site to site. However, it is not always a valid assumption and, therefore, Basu & Srinivas (2013) proposed a mathematical approach in L-moment framework to address this issue. Based on their results, the proposed approach shows a better performance than methods based on index-flood approach.

The first main step in RFFA is regionalisation. Different attributes of sites can be used for regionalisation. For a site with n selected attributes, an n-dimensional feature vector is defined to represent the site. Hydrologists proposed several methods for regionalisation, each method has its advantages and limitations. One of the most widely used regionalisation methods is cluster analysis (e.g. Mosley 1981; Tasker 1982; Wiltshire 1986; Bhaskar & O’Connor 1989; Burn 1989; Nathan & McMahon 1990; Hosking & Wallis 1997; Hall & Minns 1999; Burn & Goel 2000; Hall et al. 2002; Jingyi & Hall 2004; Rao & Srinivas 2006a, 2006b; Sadri & Burn 2011; Basu & Srinivas 2015). Cluster analysis is a procedure for grouping the data points based on the similarities between them. Each data point is represented by a feature vector. The objective is to reach the maximum possible similarity between the feature vectors of a group or cluster and also to reach the maximum possible dissimilarity between the feature vectors of different clusters (Rao & Srinivas 2008).

Hard clustering and fuzzy clustering are two main categories of clustering algorithms. In order to form the hard clusters, each feature vector should belong to only one cluster, but in fuzzy clusters each feature vector can be assigned to all the clusters with degrees of membership in the interval [0,1]. Hard clustering algorithms can be categorized as hierarchical algorithms and partitional algorithms. Some of the well-known hierarchical algorithms include: average linkage (Sokal & Michener 1958; Lance & Williams 1967; Sneath & Sokal 1973), complete linkage (Johnson 1967; Lance & Williams 1967; McQuitty 1967; Sneath & Sokal 1973), single linkage (Florek et al. 1951a, 1951b; Sneath 1957, 1966; Jardine et al. 1967; Johnson 1967; Lance & Williams 1967; Gower & Ross 1969; Sneath & Sokal 1973), and Ward’s algorithm (Ward 1963). Also K-means (KM), K-medoids and K-modes are some of the most widely used partitional algorithms. Fuzzy clustering algorithms are founded on fuzzy sets (Zadeh 1965) and fuzzy c-means (FCM) algorithm is the most well-known in regionalisation studies (e.g. Hall & Minns 1999; Rao & Srinivas 2006b; Srinivas et al. 2008; Sadri & Burn 2011; Asong et al. 2015; Basu & Srinivas 2015).

In RFFA studies, several researchers used hierarchical algorithms (e.g. Mosley 1981; Tasker 1982; Nathan & McMahon 1990; Burn et al. 1997) and KM algorithm (e.g. Wiltshire 1986; Bhaskar & O’Connor 1989; Burn 1989; Burn & Goel 2000) for regionalisation of catchments in various study areas. Rao & Srinivas (2006a) proposed a hybrid clustering approach for regionalisation of catchments. They introduced three hybrid algorithms to utilize the advantages of both hierarchical and partitional algorithms and also to tackle their limitations. Three hierarchical algorithms including single linkage, complete linkage and Ward’s algorithm were used to provide the initial cluster centres for KM algorithms. They applied these algorithms to the watersheds of Indiana in the USA and concluded that the combination of Ward’s algorithm and KM algorithm provides the best results according to some cluster validity indices and heterogeneity measures proposed by Hosking & Wallis (1997).

Hall & Minns (1999) assessed the efficiency of an FCM algorithm for regionalisation of 101 gauging stations in two areas in the UK. Also, Rao & Srinivas (2006b) applied an FCM algorithm to regionalisation of catchments of Indiana state in the USA. They used fuzzy cluster validity measures to find the optimum number of regions and suggested that measures related to both degrees of membership and data structure are more efficient than measures only related to degrees of membership. Sadri & Burn (2011) used an FCM algorithm to regional drought frequency analysis in three Canadian provinces. They also used multivariate L-moments (L-comoments) and a bivariate copula method for regional frequency analysis of droughts. Asong et al. (2013) proposed a new approach for identifying homogeneous regions for regionalisation of precipitation characteristics for the Canadian Prairie Provinces. They utilized a combination of
principle component analysis, canonical correlation analysis and FCM clustering to form homogeneous precipitation regions. Basu & Srinivas (2015) proposed an approach to solve the problem of loss of information in defuzzification of the fuzzy clusters to form the final regions in RFA. Also, they stated that the threshold strategies may yield quantile under-estimation.

Furthermore, about 30 years ago, Kohonen (1982) introduced the self-organizing feature map (SOFM) or self-organizing map. SOFM is a type of artificial neural network (ANN) that can be used as a clustering technique and therefore some hydrologists used it for regionalisation in RFFA studies (e.g. Hall & Minns 1999; Hall et al. 2002; Jingyi & Hall 2004). Besides RFFA, SOFM has found several applications in improving performances of river flow forecasting models (Tiwari et al. 2013), rainfall–runoff models (Nourani & Parhizkar 2015) and groundwater quality models (Nourani et al. 2015). Also, SOFM has been used in some studies on water distribution systems (e.g. Chang et al. 2011; Mounce et al. 2016).

In the area of RFFA, Hall & Minns (1999) used SOFM and FCM algorithms for regionalisation of catchments in the southwest of England and Wales. The results indicated the feasibility of employing these methodologies on a country-wide basis. Hall et al. (2002) studied the application of data mining techniques for the regionalisation of hydrological variables. They used one dimensional (1-D) SOFM for regionalisation of three data sets from three study areas. Also, Jingyi & Hall (2004) utilized this type of ANN for regionalisation of 86 gauging stations in two provinces in China. Srinivas et al. (2008) presented a two-level SOFM-based clustering approach by combining SOFM and FCM. The results showed that this combination performs better in estimating flood quantiles at ungauged sites. In addition, Di Prinzio et al. (2011), Razavi & Coulibaly (2013), Toth (2015) and Farsadnia et al. (2014) used combinations of SOFM and various statistical methods and different data-driven techniques to RFFA.

In this study, our objective is to evaluate the potential of a hybrid approach to improve the performance of FCM algorithm for regionalisation of catchments. The used approach may reduce the variation in the clustering results because of the random initialization of FCM algorithm. In this approach, four hierarchical clustering algorithms (average linkage, complete linkage, single linkage and Ward’s algorithm) and SOFM are used to provide the initial cluster centres for an FCM algorithm. Effects of proposed hybrid algorithms on the performance of FCM are examined by assessing the changes in the values of the objective function, fuzzy cluster validity indices and cluster sizes. In addition, the capabilities of these hybrid algorithms to form the homogeneous regions are evaluated based on the heterogeneity measures.

METHODS

Study area and data sets

The study area is Sefidrood basin located in the north of Iran (Figure 1). The total basin area is about 63,945 km² and includes portions of ten provinces: Alborz, Ardebil, Azarbaijan-e-ghari (west Azarbaijan), Azarbaijan-e-sharghi (east Azarbaijan), Ghazvin, Guilan, Hamedan, Kordestan, Mazanderan and Zanjan. Shahroud and Ghezel-ozan rivers join together and constitute Sefidrood river that flows to the Caspian Sea. Thirty-nine gauging stations over the basin were selected for use in RFFA. The annual peak flows or flood data were obtained from the electronic file provided by Iran Water Resources Management Company (WRMC).

The dependence of flood related statistics includes mean value of annual peak flows (MAF), median value of annual peak flows (MEF), mean annual flood per unit area of drainage basin (MAF/A), median annual flood per unit area of drainage basin (MEF/A), mean annual flood divided by the mean annual precipitation (MAF/P), and median annual flood divided by the mean annual precipitation (MEF/P) on the different attributes by calculating the correlations. There were noticeable values of correlation between some of these statistics and three attributes including drainage area, mean annual precipitation and runoff coefficient. Finally, five attributes were selected to form the feature vectors and evaluate dissimilarity (similarity) between the sites for use in the regionalisation. These attributes include two geographic location attributes: longitude and latitude; and three other catchment characteristics including drainage area, mean annual precipitation and runoff coefficient. In addition, the correlation matrix was used to evaluate the independence between

Downloaded from https://iwaponline.com/jh/article-pdf/18/6/1033/390709/jh0181033.pdf by guest
three latter attributes. There was not any significant correlation between these three attributes. The statistical characteristics of the considered attributes in interested sites are presented in Table 1. The attributes used to form the feature vectors were obtained from the technical reports provided by Iran WRMC.

The logarithmic transformation was applied to the drainage area values and then the selected attributes were standardized because of their different magnitudes and dimensions. For the $j$th attribute in the $i$th initial feature vector $y_i$ in an $n$-dimensional attribute space, the $j$th element of the $i$th rescaled feature vector is defined as:

$$x_{ij} = \frac{y_{ij} - \bar{y}_j}{S_j} \quad \text{for } j = 1, \ldots, n$$  \hspace{1cm} (1)

where $x_{ij}$ represents the rescaled value of $y_{ij}$, and $\bar{y}_j$ and $S_j$ denote the mean value and the standard deviation of the $j$th attribute respectively.

### Hierarchical clustering

Hierarchical clustering algorithms are multi-stage or nested clustering procedures that can be categorized into two main groups of algorithms: agglomerative algorithms and divisive algorithms. In the regionalisation studies, the agglomerative
algorithms have been used more than the divisive algorithms (e.g. Mosley 1981; Tasker 1982; Acreman & Sinclair 1986; Nathan & McMahon 1990; Burn et al. 1997; Hosking & Wallis 1997; Rao & Srinivas 2006a). For an interested set of N data points, the agglomerative algorithms begin with N clusters and each of them includes one specified feature vector. In the first step, the two nearest or most similar clusters are identified and merged. Distance or similarity between any two sites is evaluated by using a distance or dissimilarity measure such as Euclidean distance. The agglomerative clustering continues finding and merging the two nearest clusters in each step until the interested number of clusters is obtained. In contrast, divisive algorithms begin with one cluster including all feature vectors. Then, in the first step the furthest feature vector from the cluster centroid is found and separated to form a distinct cluster. The process continues until the requested number of clusters is formed (Rao & Srinivas 2008).

In this study, four agglomerative clustering algorithms are used to initialize the clusters for FCM algorithm. These algorithms include: average linkage, complete linkage, single linkage and Ward’s algorithm. Due to the abundance of available information on the application of these algorithms, this information is not presented in this paper. For details, see Rao & Srinivas (2008).

SOFMs

A SOFM is a type of ANN with special capabilities in unsupervised learning and pattern recognition and therefore can be used as an efficient tool for data clustering. Two types of SOFM were proposed by Willshaw & von der Malsburg (1976) and Kohonen (1982). The latter model has found more applications in scientific researches. Also, several researchers used SOFM in regionalisation studies (Hall & Minns 1999; Hall et al. 2002; Jingyi & Hall 2004; Srinivas et al. 2008; Di Prinzio et al. 2011; Razavi & Coulibaly 2013; Toth 2013; Farsadnia et al. 2014).

The Kohonen network consists of two layers: an input layer including n-dimensional observations x and an output layer (also called Kohonen layer) consisting of c nodes for the c clusters, each of which is associated with an n-dimensional weight (Everitt et al. 2011). The number of dimensions of the output layer does not have a principal limitation, but the one and two dimensional structures are the most widely used structures in regionalisation studies. The structure of SOFM has been described in Haykin (2009) and Everitt et al. (2011) in more detail.

In this study, the R package named ‘kohonen’ developed by Wehrens & Buydens (2007) was used to perform clustering by SOFM.

Fuzzy clustering and FCM algorithm

Fuzzy clustering is a type of clustering method in which each feature vector or data point can be assigned to more than one cluster and even all clusters simultaneously with membership values in the interval [0,1].

The matrix $U = [\mu_{ik}]$ with c rows and N columns represents the fuzzy partitions, where $\mu_{ik} \in [0, 1]$ denotes the degree of membership of the $k$th feature vector, $x_k$, belongs to the $i$th cluster ($1 \leq i \leq c$). The matrix $U$ is called the fuzzy partition matrix. Conditions for a fuzzy partition matrix are given by:

$$\sum_{i=1}^{c} \mu_{ik} = 1 \quad \forall k = 1, \ldots, N$$  \hspace{1cm} (2)

$$0 < \sum_{k=1}^{N} \mu_{ik} < N$$  \hspace{1cm} (3)

The latter condition implies that there can be no empty cluster and there can be no class that contains all the data points (Ross 2009).

Two approaches can be used to defuzzify the fuzzy partition matrix: the maximum membership and the nearest-centre classifier (Ross 2009), but the results of Rao & Srinivas (2006b) show that fuzzy clustering can improve the homogeneity of the formed regions for RFFA.

A fuzzy cluster includes the sites whose memberships in the cluster is greater than a specified threshold value. As suggested by Rao & Srinivas (2008), the value of $1/c$ is believed to be a reasonable option for the threshold fuzzy membership. The total number of the feature vectors contained in all regions can increase by using fuzzy clusters and so, more flood data can be contained in the regions. Consequently, the reliable flood estimates can be provided with higher return periods.
FCM algorithm was proposed by Bezdek (1981) and because of its similarities to KM algorithm, it is sometimes called fuzzy KM. As with the KM algorithm, FCM needs the number of clusters to be specified as an input parameter to the algorithm, but the objective function of FCM contains an additional parameter named ‘fuzzifier’ (de Oliveira & Pedrycz 2007).

For a data matrix $X = [x_1, \ldots, x_N]$, $x_i \in \mathbb{R}^n$, FCM algorithm is based on minimizing a squared error objective function defined as (de Oliveira & Pedrycz 2007):

$$J(X; U, V) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m d^2(x_k, v_i)$$  \hspace{1cm} (4)

where $U = [\mu_{ik}]$ is a fuzzy partition matrix of $X$, $v_i$ is the $i$th cluster prototype, which has to be determined, and $d^2(x_k, v_i)$ is a squared inner-product distance norm between the $k$th feature vector and the $i$th prototype. In the classical FCM algorithm, the prototypes are centres, $V = [v_1, \ldots, v_c], v_i \in \mathbb{R}^n$; therefore, the distance can be formed as below:

$$d^2(x_k, v_i) = ||x_k - v_i||^2 = (x_k - v_i)^T A(x_k - v_i)$$  \hspace{1cm} (5)

where A is the distance measure (if there is no prior knowledge, $A = I$), and parameter $m$ is called the ‘fuzziness index’ or ‘fuzzifier’; it is used to control the fuzziness of each feature vector membership, $m \in [1, \infty)$. There is no theoretical basis for the optimal selection of $m$, and a value of $m = 2.0$ is often chosen (Bezdek 1981). Also, when $m \rightarrow 1$ FCM is transformed to classical KM.

The FCM algorithm is an iterative process to minimize Equation (4) while updating the cluster centres $v_i$ and the memberships $\mu_{ik}$ by de Oliveira & Pedrycz (2007):

$$v_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m x_k}{\sum_{k=1}^{N} (\mu_{ik})^m} \hspace{1cm} \forall i = 1, \ldots, c$$  \hspace{1cm} (6)

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left( \frac{d(x_k, v_j)}{d(x_k, v_i)} \right)^{2/(m-1)}}, \hspace{1cm} 1 \leq i \leq c, \hspace{0.2cm} 1 \leq k \leq N.$$  \hspace{1cm} (7)

when the difference between the updated and the previous $U$ reaches a predefined minimum threshold or the objective function is below a certain tolerance value, the algorithm can be terminated. Also, the maximum number of iterations can be considered as a termination criterion (de Oliveira & Pedrycz 2007).

Fuzzy cluster validity indices

Specifying the number of clusters ($c$) is an important problem in fuzzy clustering. Selections of a different number of initial clusters result in different clustering partitions. The problem for finding an optimal $c$ is usually called cluster validity (Wang & Zhang 2007).

Compactness and separation are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme. Compactness measures closeness of feature vectors within each cluster. On the other hand, separation indicates how distinct two clusters are. It computes the ‘distance’ between two different clusters (Wang & Zhang 2007).

Several validity indices have been proposed in the literature. These indices can be categorized into two main groups. The first group indices make use only of membership values and have the advantage of being easy to compute. On the other hand, the indices of the second group involve both partition matrix $U$ and the data set structure. Now, it is widely accepted that a better definition of a validity index always considers both partition matrix $U$ and the data set itself (Wang & Zhang 2007). This conclusion also is verified in RFFA by Rao & Srinivas (2006a). In this study partition coefficient (PC) and partition entropy (PE) from the first group and extended Xie and Beni’s index and Kwon index from the second group were used to validate clusters.

Bezdek (1974a, 1974b) presented a performance measure named PC. The index was defined as:

$$V_{PC} = \frac{1}{N} \sum_{i=1}^{c} \sum_{h=1}^{N} \mu_{ih}^2$$  \hspace{1cm} (8)

The index values range in $[1/c, 1]$.

The PE was proposed by Bezdek (1981) as below:

$$V_{PE} = -\frac{1}{N} \sum_{i=1}^{c} \sum_{h=1}^{N} \mu_{ih} \log \mu_{ih}$$  \hspace{1cm} (9)
where $a$ is the base of the logarithm. The index is computed for values of $c$ greater than 1 and its values range in $[0, \log_a c]$.

Pal & Bezdek (1995) modified a validity function proposed by Xie & Beni (1991) with $m = 2$ as follows:

$$V_{XB,m} = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^m ||x_k - v_i||^2}{\min_{i,j} ||v_i - v_j||^2}$$ (10)

They stated that a good partition produces a small value for the compactness, and that well-separated vs will produce a high value for the separation (Wang & Zhang 2007). Also, Kwon (1998) proposed an extended version of Xie and Beni’s index to eliminate its tendency to monotonically decrease when the number of clusters approaches the number of data points by adding a punishing function to the numerator of Xie and Beni’s original validity index. The Kwon index was defined as (Kwon 1998):

$$V_K = \frac{\sum_{i=1}^{c} \sum_{k=1}^{N} \mu_{ik}^2 ||x_k - v_i||^2 + \frac{1}{c} \sum_{i=1}^{c} ||v_i - v_i||^2}{\min_{i,j} ||v_i - v_j||^2}$$ (11)

The second term in the numerator in Equation (11) is an ad hoc punishing function, used to eliminate the decreasing tendency when the number of clusters $c$ becomes very large and close to the number of data points $N$ (Kwon 1998).

In general, optimal partition corresponds to a minimum value of $V_{PE}$, $V_{XB,m}$ and $V_K$, while a maximum value of $V_{PC}$ implies an optimal partition.

Computation of an FCM algorithm in this study was performed by using the R package ‘e1071’ developed by Meyer et al. (2014).

**Discordancy and heterogeneity measures**

Given a group of sites, the aim of screening the data is to identify those sites that are grossly discordant with the group as a whole (Hosking & Wallis 1997). Hosking & Wallis (1997) defined a discordancy measure, $D$, in terms of the L-moments of the site’s data. They suggested critical values of $D$ to identify discordant sites (for $N \geq 14$, $D_{critical} = 3$). If the value of $D$ for a specified site exceeds the critical value, it is considered as a discordant site.

The identification of homogeneous regions is one of the most important stages of RFFA. Heterogeneity (or homogeneity) measures are used to assess whether formed regions are heterogeneous or homogeneous. Hosking & Wallis (1997) introduced three heterogeneity measures denoted by $H_1$, $H_2$ and $H_3$, to estimate the degree of heterogeneity in a group of sites and to assess whether the sites might reasonably be treated as a homogeneous region. A region is regarded as ‘acceptably homogeneous’ if $H < 1$, ‘possibly heterogeneous’ if $1 \leq H < 2$, and ‘definitely heterogeneous’ if $H \geq 2$ (Hosking & Wallis 1997). However, in some studies, a region is considered as ‘possibly homogeneous’ if $1 \leq H < 2$ (e.g. Rao & Srinivas 2006a, 2006b; Wallis et al. 2007; Srinivas et al. 2008).

Because of the abundance of available information about L-moments theory and its application in RFFA, that information is not presented in this paper. More details can be found in Hosking & Wallis (1997).

The R package ‘lmomRFA’ written by Hosking (2014) was used to compute discordancy measure and heterogeneity measures for sites and formed regions.

**A hybrid approach for fuzzy clustering**

Let $X$ represent a set of $N$ standardized feature vectors in $n$-dimensional attribute space, each represents one of the $N$ sites. Feature vectors are grouped in $c$ (the interested number of clusters) clusters by using SOFM and four hierarchical clustering algorithms including average linkage, complete linkage, single linkage and Ward’s algorithm. The cluster centres provided by different hierarchical algorithms and SOFM are used to specify initial cluster centres for FCM algorithm, separately. Then, an iterative FCM algorithm with the initialized cluster centres is applied to $N$ rescaled feature vectors to form $c$ fuzzy clusters. The iterative algorithm is repeated until a termination criterion is met. This approach can be considered as a fuzzy extension of hybrid clustering algorithms presented by Rao & Srinivas (2006a).

For each specified number of clusters, $c$, values of the objective function and fuzzy cluster validity indices obtained for each proposed algorithm can be compared with the
average corresponding values related to the FCM initialized randomly in order to evaluate the performance of each hybrid algorithm. In fact, these average values are calculated for a reasonable number of performing FCM algorithms with randomly initialized cluster centres. Although specified values of $c$ and $m$ optimal clustering is determined by minimizing the objective function, the minimum value may depend on selecting initial cluster centres. There is no simple and certain way to determine most optimum initial cluster centres. Therefore, if an approach of selecting initial cluster centres can provide a final value of objective function lower than an average value of objective function provided by other options, then it can be used as a reasonable option to perform clustering.

Besides the items mentioned above, cluster or region size and heterogeneity measures can be used to compare performances of the different combinations of the clustering algorithms with each other. The number of sites belonging to each region is an important issue in RFFA. Jakob et al. (1999) suggested the $5T$ rule as a rule of thumb. The $5T$ rule means that the sum of the observed data of all stations (in station-years) in a region should be at least equal to $5T$, where $T$ is the return period of interest. In addition, approximate homogeneity is sufficient to ensure that regional frequency analysis is much more accurate than at-site analysis (Hosking & Wallis 1997). Also, larger homogeneous regions can provide flood estimates with higher return periods.

R (2014) programming language and its GUI were used to perform the required computations and plotting graphs.

RESULTS AND DISCUSSION

In the initial stage, the discordancy measure $D$ was applied to all the 39 selected sites. According to the values of the measure $D$, two sites were identified as discordant ($D > 3$) and data related to them were removed from data sets. Thus, the study was followed on the 37 remained sites. Further, according to the $5T$ rule (Jakob et al. 1999), regionalisation of the 37 remaining sites was examined for a range of cluster numbers, $c = 2, 3, \ldots, 7$.

Five clustering algorithms including average linkage, complete linkage, single linkage, Ward’s algorithm, and SOFM were applied to the 37 feature vectors to initialize cluster centres for FCM algorithm. In this paper, these combinations are denoted by ALFCM, CLFCM, SLFCM, WFCM and SOFMFCM, respectively.

The ability of the hybrid approach to minimize FCM objective function is a key point to evaluate its performance. The values of FCM objective function obtained by each hybrid algorithm can be used to compare its performance with other hybrid algorithms. To evaluate the performances of the hybrid algorithms in comparison with FCM initialized randomly, the values of the objective function were used. As mentioned before, for a number of clusters $c$, an average value of the FCM objective function was obtained in a number of implementations of FCM algorithm initialized randomly. Each implementation included a required number of iterations to form $c$ steady clusters. The computation procedure can be summarized as below:

- For $p$th implementation, the objective function value denoted by $J_p$ is calculated.
- For $p$th implementation, $\bar{J}_p$ represents the mean value of the objective function in $p$ implementation. $\bar{J}_p$ is calculated as below:

$$\bar{J}_p = \frac{\sum_{i=1}^{p} J_i}{p}$$

(12)

- In each step $p$, a relative difference value between $\bar{J}_p$ and $\bar{J}_{p-1}$ are calculated and represented by $RD_{\bar{J}_p}$. $RD_{\bar{J}_p}$ is defined as follows:

$$RD_{\bar{J}_p} = \left| \frac{\bar{J}_p - \bar{J}_{p-1}}{\bar{J}_{p-1}} \right|$$

(13)

- The process is continued until $RD_{\bar{J}_p} < 0.0001$ in six sequent implementation cycles and then the process is terminated and the final value of $p$ is denoted by $p_f$.

Figure 2 shows, for $c = 2, 3, \ldots, 7$, how $RD_{\bar{J}_p}$ varies with incremental changes in values of $p$. For $c = 2, 3, \ldots, 7$, values of $p_f$ were approximately equal to 12, 250, 810, 670, 530 and 610, respectively. In this paper, the average values of the objective function and the cluster...
validity indices related to $p_f$ implementations of the randomly initialized FCM are denoted by arFCM.

FCM objective function values were calculated for $c = 2, 3, \ldots, 7$ and the fuzzifier values $m \in [1.1, 2.5]$ with increments equal to 0.1. Values of $J$ of FCM algorithm for mentioned ranges are presented in Figure 3. Generally, the value of $J$ reduces when $c$ increases for a specified value of the fuzzifier $m$. It also decreases when the value of fuzzifier $m$ increases for a specified number of clusters.

According to Figure 3(a), for $c = 2$, the values of the objective function $J$ calculated for clusters formed by the five different hybrid algorithms are identical. Also, the mean values of $J$ related to arFCM in the range $m \in [1.1, 2.5]$ are equal to those of the hybrid algorithms. In fact, there are two well-separated clusters based on the defined feature vectors and therefore, all the examined approaches and algorithms result in the same clusters and the same values of $J$.

As seen in Figure 3(b), for $c = 3$, a noticeable increase in absolute values of slopes of curves related to SLFCM, CLFCM and ALFCM algorithms occurs in $m \in [2.2, 2.3]$. In addition, for $c = 4$, it is seen in Figure 3(c) that the value of $J$ computed for SOFMFCM algorithm decreases sharply in the range $m \in [1.5, 1.6]$. Also, in this figure, when the value of $m$ increases from 1.1 to 1.2, a fast change is observed in the slope of the graph related to SLFCM. Furthermore, a similar change is observed in Figure 3(d) in the case of SOFMFCM algorithm. These noticeable changes occur because of sudden significant changes in configuration of the clusters provided by the examined clustering algorithms. The effects of the changes in configuration of the clusters on the values of the cluster validity measures can be seen in Figure 4.

Four cluster validity indices including partition coefficient ($V_{PC}$), partition entropy ($V_{PE}$), extended Xie and Beni’s index ($V_{XB,m}$) and Kwon index ($V_k$) were used to validate the clusters formed by the hybrid algorithms and arFCM. The computed values of these indices are seen in Figure 4. It is evident from Figure 4 that $V_{PC}$ shows a monotonic decrease with an increase in the values of fuzzifier, and on the other hand $V_{PE}$ shows a monotonic increase with an increase in the values of $m$. In fact, both $V_{PC}$ and $V_{PE}$ always suggest $m = 1.1$ as the optimal value of fuzzifier. This is consistent with the results presented by Rao & Srinivas (2006b). Based on the results related to these two indices, the quality of resulting clusters always decreases with an increase in their fuzziness. In fact, these indices seem to be appropriate to look for well-separated clusters with crisp boundaries. Also, as reported by Rao & Srinivas (2006b), because of lack of a connection to data structure, appropriate clustering according to $V_{PC}$ and $V_{PE}$, may not result in homogeneous regions. Also, it is seen in Figure 4 that $V_{PC}$ and $V_{PE}$ may not clearly distinguish between performances of the different clustering algorithms. So, it seems to be more reasonable to pay more attention to the two other indices, $V_{XB,m}$ and $V_k$, in order to compare performances of the clustering algorithms.

For $c = 2$, values of the four cluster validity indices obtained by all the examined algorithms are nearly identical, as expected (Figure 4(a)). This result confirms the results related to the variation of the objective function for $c = 2$. In fact, all the used clustering algorithms form two identical clusters, because there are two potential well-defined clusters based on the feature vectors of the sites.

According to Figure 4, for $c = 3$, the values of $V_{XB,m}$ and $V_k$ for SLFCM, CLFCM and ALFCM algorithms in the range $m \in [1.1, 2.1]$ are more suitable than that of arFCM, while values of these indices for the other two algorithms are not desirable in the same range. However, in the range $m \in [2.3, 2.5]$, all the considered algorithms present approximately the same results. For $c = 4, 5, 6, 7$, based on the values of $V_{XB,m}$ and $V_k$, all the hybrid algorithms often present more appropriate performances than that of arFCM, even though there are a few exceptions.

In the case of $V_{XB,m}$ and $V_k$, it can be seen in Figure 4 that the values of each index for the different clustering algorithms become closer to each other when the values of the fuzzifier increases. On the other hand, the differences between the values of each index for the different clustering algorithms increases with an increase in the number of clusters.

Similar to the results related to the variation of the objective function, some sudden changes are observed in the absolute values of slopes of the graphs related to the cluster validity indices. Generally, these changes increase with an increase in the number of the clusters. Especially, for $c = 7$, a considerable number of shifts in the values of $V_{XB,m}$ and $V_k$ are seen. These changes occur because of
noticeable changes in the configuration of clusters formed by these algorithms in the mentioned ranges as seen in Figures 5–8.

As seen in Table 2, the mean values of $J$ and the cluster validity indices computed for all the hybrid algorithms for $V_{XB,m}$ over the range $m \in [1, 2.5]$ are more desirable than that of arFCM (lower values of $J$ and $V_{PE}$, $V_{XB,m}$ and $V_K$ measures, and higher values of $V_{PC}$). These results can show the efficiency of the hybrid approach to improve the performance of the FCM algorithm.

As evidenced by the results shown in Table 3, using each of the hybrid algorithms produced results better than those
obtained by arFCM, at least in 62% of applications. In the case of minimizing the objective function, the WFCM algorithm provided the best results.

To specify final regions and identify sites assigned to them, for a specified value of $c$ ($c = 2, 3, \ldots, 7$) a membership threshold $1/c$ suggested by Rao & Srinivas (2008) was considered. In fact, for each region or cluster, the sites belonging to it with a degree of membership greater than or equal to $1/c$ were specified as final members of the interested region. In the case of a site assigned to more than one region, its flood quantiles are computed as a weighted mean of its quantiles in different regions that it is assigned to them.

Figure 3 | Optimal value of $J$ of FCM obtained by the proposed hybrid clustering algorithms and arFCM for different values of fuzzifier ($m$) for (a) $c = 2$, (b) $c = 3$, (c) $c = 4$, (d) $c = 5$, (e) $c = 6$, and (f) $c = 7$. The vertical axis shows the values of objective function ($J$) for the hybrid algorithms, and the mean value of the objective function in $p$ implementations for arFCM ($J_{arFCM}$).
Values of cluster validity indices for the proposed hybrid clustering algorithms and arFCM for different values of fuzzifier $m \in [1, 2.5]$ for (a) $c = 2$, (b) $c = 3$, (c) $c = 4$, (d) $c = 5$, (e) $c = 6$, and (f) $c = 7$.

Figure 4 | Values of cluster validity indices for the proposed hybrid clustering algorithms and arFCM for different values of fuzzifier $m \in [1, 2.5]$ for (a) $c = 2$, (b) $c = 3$, (c) $c = 4$, (d) $c = 5$, (e) $c = 6$, and (f) $c = 7$. 
As seen in Figures 5–9 for \( c = 2 \), the total number of sites belonging to all the clusters using each of the hybrid algorithms is equal to 37. This implies that each site is assigned only to one region and no site is assigned to both regions. For \( c = 2 \), the value of the threshold is equal to 1/2. In addition, the condition \( \sum_{i=1}^{C} H_{ik} = 1 \) presented in Equation (2) must be satisfied. Therefore, it is very unlikely that a site belongs to the two regions simultaneously when \( c = 2 \). In other words, if we use the threshold value \( 1/c \) for \( c = 2 \), the final results of regionalisation using a \( c \)-means algorithm would be very close to the results of the classic KM algorithm.

Figure 5 | Number of sites belonging to the regions formed by SLFCM algorithm for different values of fuzzifier \( m \in [1, 2, 5] \) for \( c = 2, \ldots, 7 \).
For $c = 3$, SLFCM, CLFCM and ALFCM result in an identical regionalisation. On the other hand, the regionalisations provided by WFCM and SOFMFCM are the same, but different from that related to the three former algorithms. For $c = 4$ and $c = 6$, the performances of SLFCM, CLFCM, ALFCM and WFCM algorithms in forming the regions are the same. There is only one exception related to SLFCM in $m = 1.1$ for $c = 4$. It should be noted that the region or cluster numbers may vary from algorithm to algorithm. As seen in the figures, for $c = 4$, the configuration of regions formed by SOFMFCM is different from that of other algorithms. For $c = 5$ only regionalisations performed by CLFCM and WFCM are completely
the same over the range \( m \in [1, 1.5] \) and finally, for \( c = 7 \), there are fewer similarities between the regionalisations performed by the five hybrid algorithms.

Figure 5 shows some sudden changes in configuration of the regions formed by the SLFCM algorithm for \( c = 5 \) and \( c = 7 \). Also, it is evident from Figure 9 that configuration of regions formed by an SOFMFCM algorithm changes in some ranges considerably. These changes or shifts correspond to those shown in Figures 3 and 4 related to the values of the objective function and the cluster validity indices.

In general, as evidenced by Figures 5–9, for each value of \( c \), the total number of sites belonging to all clusters
increases with an increase in value of fuzzifier \((m)\), i.e. increase in fuzziness of clusters. This result confirms the results presented by Rao & Srinivas (2006b). In addition, based on the results shown in the figures, SOFMFCM shows a higher ability to increase the total number of sites belonging to all the regions, especially for \(c = 6\) and \(c = 7\).

Descriptive statistics of the sizes of the regions and the number of the singleton clusters obtained by each used algorithm are presented in Table 4. As seen in Table 4, SLFCM and ALFCM algorithms tend to form regions with noticeable differences in their sizes (i.e. a small number of large regions and several small regions). These findings are

Figure 8 | Number of sites belonging to the regions formed by WFCM algorithm for different values of fuzzifier \(m \in [1, 2.5]\) for \(c = 2, \ldots, 7\).
consistent with the results reported by Rao & Srinivas (2006a). On the other hand, CLFCM, WFCM and SOFMFCM algorithms often form regions with sizes relatively close to each other. Also, they do not tend to form regions including one site (singleton). In general, singleton regions are not appropriate to perform RFFA and therefore the clustering algorithms that tend to form singleton clusters are not reasonable choices for regionalisation.

In some cases, it can be seen that the results related to the objective function values and the cluster validity measures of two hybrid algorithms are different, but their final clusters (regions) are the same. This is related to the finalization process. Finalization of the clusters using a membership threshold decreases differences between clusters formed by different algorithms. To understand this concept, consider two feature vectors $j$ and $k$ that for $c = 4$, their memberships in four clusters are $\mu_j = [0.97, 0.01, 0.01, 0.01]$ and $\mu_k = [0.28, 0.24, 0.24, 0.24]$ for $i = 1,2,3,4$ respectively. After finalization of the clusters using a threshold 1/4, assignments of the feature vectors to both of them are the same ([1,0,0,0]). For this reason, some differences seen between values of $J$ or cluster validity indices computed for the different clustering algorithms are not recognizable in the final clusters formed by them.

Homogeneity of the regions formed by each hybrid algorithm were assessed by the heterogeneity measures $H_1, H_2$ and $H_3$ introduced by Hosking & Wallis (1997). According to the results, for $c = 2$, all the regions formed by all the hybrid algorithms over the range $m \in [1.1, 2.5]$ are homogeneous. As mentioned before, for $c = 2$, there are two well-defined regions that can be identified by all the hybrid algorithms and so, the heterogeneity measures are identical for the regions provided by the different hybrid algorithms. In general, $c = 2$ can be chosen as the optimum number of the regions in order to partition Sefidroud basin into homogeneous regions for RFFA.

For the regions formed by SLFCM and ALFCM, in many cases the values of the heterogeneity measures are relatively identical, because the regions formed by these two algorithms are similar to each other. In general, for $c = 2,3,4$ and $c = 6,7$ there is no noticeable difference between homogeneity of the regions provided by these two algorithms. Also, the regions formed by the WFCM algorithm are often very similar to the regions provided by the CLFCM algorithm. Therefore the computed heterogeneity measures are very close to each other for $c = 4$, $c = 5$ and $c = 6$. However, it should be noted that computations of $H$ measures are performed using simulation of a number of homogeneous regions with specified regional L-moment ratios, and therefore some limited variations may be seen in values of heterogeneity measures computed for a specified region in two different iterative computations.

Table 5 shows the ratio of the homogeneous regions formed by each hybrid algorithm according to each heterogeneity measure $H$ to all the regions provided by it. The first three columns of the table show the ratio of homogeneous regions ($H<1$), each based on one of the heterogeneity measures ($H_1, H_2, H_3$). The two last columns of the table are related to the assessment of homogeneity of the formed regions based on the values of the three heterogeneity measures ($H_1, H_2, H_3$), simultaneously. It is seen from the table that if the regions with $H<1$ are considered as ‘acceptably homogeneous’, then CLFCM, WFCM and SOFMFCM algorithms provide a maximum number of homogeneous regions respectively. However, if the criteria $H<2$ suggested by Rao & Srinivas (2006a,
2006b), Wallis et al. (2007) and Srinivas et al. (2008) is used to identify 'possibly homogeneous' regions, then the percentage of the homogeneous regions provided by each hybrid algorithm grows considerably, especially in the cases of WFCM and SOFMFCM, and all the formed regions can be considered as homogeneous.

In general, it seems to be appropriate to divide Sefidroud basin into two regions based on the feature vectors related to the sites. As already mentioned, there are two well-defined regions that can be identified by all the examined hybrid algorithms. The regions are shown in Figure 10. These two regions also satisfy the homogeneity conditions completely.
In addition, the lower number of regions usually results in forming larger regions so that it can help to provide flood quantiles related to the higher return periods. Although, the increase in the number of regions often causes the increase in the total number of sites belonging to all the regions, it usually results in smaller regions. Therefore, \( c = 2 \) is a reasonable choice for regionalisation in Sefidroud basin. For the current case study, there is no difference between performances of the examined hybrid algorithms for \( c = 2 \). However, for the other values of \( c \), WFCM, SOFMFCM and CLFCM algorithms generally show better performances in forming medium-size homogeneous regions.

As seen in Figure 10, the first and second formed regions include 26 and 11 sites respectively. Twenty-five sites of the first region are located in the west and middle area of Sefidroud basin on the Ghezel-ozan river subbasin. On the other hand, all the sites of the second region are located in the east area of Sefidroud basin, eight of them on the Shahroud river subbasin, and the three remaining sites around the main reach of Sefidroud river. Besides the geographical contiguity, a look at the feature vector related to the sites of the second region shows that these sites have smaller drainage areas and greater values of the mean annual precipitation than those of the sites of the first region. Also, they often include higher values of the runoff coefficient compared to the sites of the first region.

According to the results obtained in the current study, compared to the non-hybrid approach, the hybrid approach can often improve the performance of a FCM algorithm to form more desirable clusters based on the values of the objective function and the cluster validity indices. Furthermore, while in the non-hybrid approach the initial cluster centres for FCM algorithm must be defined, it is not required in the hybrid approach, because the initial cluster centres used by FCM algorithm are specific feature vectors provided by the hierarchical clustering algorithms or SOFM. In the non-hybrid approach, the initial cluster centres for FCM algorithm are often defined randomly and so, the final formed clusters may change in each implementation of clustering algorithm with a change in the initial cluster centres. This problem is also solved by using the hybrid approach. This way, while it is not required to specify initial cluster centres because of using the hierarchical clustering algorithms or SOFM, the dynamic structure of FCM algorithm can improve the quality of the final clustering in the hybrid approach. In fact, contrary to the hierarchical clustering algorithms, in FCM algorithm, the data points can move from one cluster to another to minimize the objective function.

### Table 4 | Main statistical characteristics of sizes of the regions formed by the hybrid algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Mean</th>
<th>Std. dev.</th>
<th>Max</th>
<th>Min</th>
<th>Singletons</th>
<th>Total number of sites belonging to all clusters in ( m = 2.5 ) for ( c = 7 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLFCM</td>
<td>9.595</td>
<td>5.978</td>
<td>26</td>
<td>1</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>CLFCM</td>
<td>9.578</td>
<td>5.757</td>
<td>26</td>
<td>3</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td>ALFCM</td>
<td>9.560</td>
<td>5.799</td>
<td>26</td>
<td>1</td>
<td>3</td>
<td>65</td>
</tr>
<tr>
<td>WFCM</td>
<td>9.567</td>
<td>5.044</td>
<td>26</td>
<td>5</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td>SOFMFCM</td>
<td>9.564</td>
<td>4.851</td>
<td>26</td>
<td>3</td>
<td>0</td>
<td>67</td>
</tr>
</tbody>
</table>

### Table 5 | Ratio of homogeneous regions formed by each hybrid algorithm to all the regions provided by it

<table>
<thead>
<tr>
<th>Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>SLFCM</td>
</tr>
<tr>
<td>CLFCM</td>
</tr>
<tr>
<td>ALFCM</td>
</tr>
<tr>
<td>WFCM</td>
</tr>
<tr>
<td>SOFMFCM</td>
</tr>
</tbody>
</table>

Downloaded from https://iwaponline.com/jh/article-pdf/18/6/1033/390709/jh0181033.pdf by guest
CONCLUSIONS

As mentioned before, using fuzzy clusters in regionalisation can increase the total number of feature vectors included in all regions. This causes a growth in data contained in each region. Consequently, reliable flood estimates can be provided with higher return periods. This may be an important motive to use FCM in RFFA studies. On the other hand, specifying initial cluster centres for an FCM algorithm affects the final resulting clusters. In this paper, a hybrid approach was proposed to improve performance of FCM algorithm for regionalisation of catchments. SOFM and four hierarchical algorithms including average linkage, complete linkage, single linkage and Ward’s algorithm were used to determine the initial cluster centres for FCM algorithm.

To evaluate the performances of the proposed hybrid algorithms, for each specified number of clusters, \( c \), values of the objective function of FCM, \( J \), and the four fuzzy cluster validity indices obtained for each proposed algorithm were compared with average corresponding values calculated for a reasonable number of performing FCM algorithm with randomly initialized cluster centres. Results showed that according to the values of both objective function and the cluster validity indices, the performances of the hybrid algorithms are often more appropriate than those of randomly initialized FCM. In the case of minimizing the objective function, the combination of Ward’s algorithm and FCM provided the best results, but according to cluster validity indices other hybrid algorithms such as SLFCM and CLFCM presented the most desirable results. Also, the results showed that the PC and the PE may be not efficient enough to determine optimal fuzzy clusters.

The comparison of sizes of the regions formed by the hybrid algorithms showed that SLFCM and ALFCM algorithms tend to form the regions with noticeable differences in their sizes (i.e. a small number of large regions and several small regions). On the other hand, CLFCM, WFCM and SOFMFCM algorithms often form regions with sizes relatively close to each other. Also, they do not tend to form regions including one site (singletons). According to these results, it seems to be better to use the algorithms of the latter group (CLFCM, WFCM and SOFMFCM) when our goal is to perform an RFFA in order to estimate flood.

Figure 10 | Final assignment of the sites to two optimal regions in Sefidroud basin.
quantiles for all sites included in an interested area. However, if flood estimation for a specified site is required, then it may be more reasonable to evaluate the performances of the different hybrid clustering algorithms and to choose the one that can provide reliable flood estimations with the largest return period for the interested site.

In this study, regionalisation is performed by using geographical location, physiological and meteorological attributes. The heterogeneity measures were computed based on flood data and so there may be no direct relation between the performance of clustering algorithm and homogeneity of regions. However, it was seen from the results that the homogeneity of the regions formed by WFCM, SOFMFCM and CLFCM algorithms were at a higher rate than those of SLFCM and ALFCM algorithms.

In the case of Sefidrood basin, as already mentioned, there are two well-defined regions identified by all the examined hybrid algorithms. Also, these two regions are completely homogeneous based on the values of the heterogeneity measures. Further, \( c = 2 \) is an appropriate choice for the number of regions because it results in forming larger regions in comparison with those provided by the greater values of \( c \). For the current case study, there is no difference between performances of the examined hybrid algorithms for \( c = 2 \). However, in general, for the other values of \( c \), WFCM, SOFMFCM and CLFCM algorithms shows better performances in forming medium-size homogeneous regions.

For future work, it may be very useful to evaluate the effects of feature selection and feature weighting methods on the performance of hybrid clustering algorithms, especially in RFFFA studies.

REFERENCES


