J. Hydrol. Hydromech., 65, 2017, 4, 359–365
DOI: 10.1515/johh-2017-0024

A novel fuzzy clustering approach to regionalise watersheds with an automatic determination of optimal number of clusters

Javier Senent-Aparicio*, Jesús Soto, Julio Pérez-Sánchez, Jorge Garrido

Civil Engineering Department, UCAM Universidad Católica San Antonio de Murcia (UCAM), Campus de los Jerónimos, nº 135, 30107 Murcia, Spain.
* Corresponding author. Tel.: +34 968 278 818. E-mail: jsenent@ucam.edu

Abstract: One of the most important problems faced in hydrology is the estimation of flood magnitudes and frequencies in ungauged basins. Hydrological regionalisation is used to transfer information from gauged watersheds to ungauged watersheds. However, to obtain reliable results, the watersheds involved must have a similar hydrological behaviour. In this study, two different clustering approaches are used and compared to identify the hydrologically homogeneous regions. Fuzzy C-Means algorithm (FCM), which is widely used for regionalisation studies, needs the calculation of cluster validity indices in order to determine the optimal number of clusters. Fuzzy Minimals algorithm (FM), which presents an advantage compared with others fuzzy clustering algorithms, does not need to know a priori the number of clusters, so cluster validity indices are not used. Regional homogeneity test based on L-moments approach is used to check homogeneity of regions identified by both cluster analysis approaches. The validation of the FM algorithm in deriving homogeneous regions for flood frequency analysis is illustrated through its application to data from the watersheds in Alto Genil (South Spain). According to the results, FM algorithm is recommended for identifying the hydrologically homogeneous regions for regional frequency analysis.

Keywords: Fuzzy Clustering; Regionalisation; Alto Genil; Hydrological homogeneity; Regional flood frequency analysis.

INTRODUCTION

Identification of hydrologically homogeneous watersheds is one of the aspects that must be taken into account when we make the estimation of hydrological variables that are involved in water resources planning and management. In general, the networks of gauging stations are scarce, so there is not enough information to accurately determine the amount of water resources that are available in a particular watershed or the design flow of a given flow control structure. These limitations of information are intended to be overcome through the determination of hydrologically homogeneous watersheds and regional flood frequency analysis (RFFA). In a hydrologically homogeneous region, each of the catchments have the same rescaled distribution function of the annual/seasonal maxima; thus, the samples from the individual catchments can be merged in a larger, regional sample, and analyse it in order to achieve more robust estimations of quantiles or return periods. The degree of regional homogeneity/heterogeneity is assessed by tests of regional homogeneity (Hosking and Wallis, 1997). The identification of hydrologically homogeneous regions is usually the most important and difficult step of the RFFA (Smithers and Schulze, 2001). The watersheds are often grouped by collecting the geographically close stations into the same group. However, it is not possible to say that the regions generated with this approach are hydrologically homogeneous.

A review of the approaches generally used for regionalisation of watersheds is found in Rao and Srinivas (2008). Those include the region of influence (ROI) approach and its extensions, the method of residuals, the canonical correlation analysis, the hierarchical approach and its extension to ROI framework and the cluster analysis. Since the end of twentieth century, cluster analysis and fuzzy clustering (FC) procedures, specifically, gained recognition as a major mechanism in regionalisation of watersheds for RFFA (see Table 1), as those procedures allow watersheds to partially resemble each other.

Table 1. Previous studies on regionalization of watersheds for RFFA.

Authors	Study area and data	Method of regionalization
Bargaoui et al. (1998)	Tunisia (39 stations)	Iphigenie and ISODATA fuzzy clustering methods
Hall and Minns (1999)	United Kingdom (101 stations)	Fuzzy c-means algorithm
Jingyi and Hall (2004)	China (86 stations)	Fuzzy c-means algorithm
Rao and Srinivas (2006)	Indiana, USA (245 stations)	Fuzzy c-means algorithm
Isik and Singh (2008)	Turkey (1410 stations)	Hierarchical clustering, k-means algorithm and flow duration curve method
Srinivas et al. (2008)	Indiana, USA (245 stations)	Self-organizing feature map
Gaál et al. (2009)	Slovakia (56 stations)	Cluster analysis
Raju and Nagesh Kumar (2011)	Rajasthan, India (25 stations)	K-means cluster analysis, fuzzy cluster analysis and Kohonen Neural Networks
Goyal and Gupta (2014)	Northeast India (68 stations)	Fuzzy c-means algorithm and K-Mean algorithm
Basu and Srinivas (2014)	Ohio, USA (305 stations)	Kernel based Fuzzy c-means
Basu and Srinivas (2015)	Mid-Atlantic, USA (114 stations)	Defuzzification approach and threshold strategy
Kumar et al. (2015)	Godavari, India (17 stations)	Artificial neural network and fuzzy inference system
Goyal and Sharma (2016)	Western India (81 stations)	Fuzzy c-means algorithm
Agarwal et al. (2016)	United States of America (530 stations)	Wavelet-based Multiscale Entropy

FC provides more information about the structure in data than hard clustering, so FC is a better choice for RFFA (Rao and Srinivas, 2006).

One of the steps in regionalisation by cluster analysis involves the selection of a clustering algorithm to partition feature vectors. Once clusters are formed, they are interpreted visually and by using cluster validity indices (e.g. partition coefficient; partition entropy; Xie-Beni; Fukuyama-Sugeno) to determine the optimum number of regions (see details below in the paper). According to Rao and Srinivas (2008) cluster validity indices must be used with extreme prudence due to the fact that they are developed and validated in applications other than regionalisation of watersheds. Hence, the identification of the optimal number of regions in regionalisation studies should be purposed in future research.

In this work, we have checked the use of FM for the identification of hydrologically homogeneous regions due to the fact that this algorithm presents an advantage compared with others fuzzy clustering algorithms: it does not need to know a priori the number of clusters, so cluster validity indices are not used. Therefore, the objective of this study was to apply a novel clustering algorithm with the ability to (1) produce a partition that represents a meaningful interpretation of structure in the data, obtained without applying cluster validity indices to the optimal number of clusters, and (2) test the regions for homogeneity by using statistical homogeneity tests. Results obtained with the application of this algorithm have been compared with those obtained with the application of the widely used Fuzzy C-Means algorithm.

METHODS

In the present study, practical applicability of two fuzzy clustering techniques, namely, Fuzzy C-Means (FCM) and Fuzzy Minimals (FM) is analysed for grouping 20 watersheds of Alto Genil watershed (South Spain). These techniques are explained briefly below.

Fuzzy C-Means algorithm

If we want to classify a sample, but we do not know the classes of available clusters and even the number of the available classes for the sample, we can use the unsupervised classification or clustering method to find out the classes using some measure of similarity. Similarity can be defined as proximity of the points in the textures space according to a distance function.

Usually partitioned clustering (cluster analysis), called 'hard clustering', assigns each datum to exactly one cluster; on the other hand, fuzzy cluster analysis allows gradual memberships. In this way, we can deal with data belonging to more than one cluster at the same time.

The origin of fuzzy clustering emerges with the work of Bellman et al. (1966) and Ruspini (1969), based on the ideas of Zadeh (1965). Dunn (1974) formalised the FCM algorithm, which was later generalised by Bezdek (1981). FCM algorithm calculates group membership probabilities or degrees, taking into account distance between objects and group prototypes. The aim of the FCM algorithm is to find an optimal fuzzy c-partition and corresponding prototypes, minimising the objective function:

$$J(U,V) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m \| x_k - v_i \|^2 \qquad (1)$$

where, $X = \{x_1, \ldots, x_n\}$ is a data set, each data point x_k is an input vector, $V = (v_1, v_2, \ldots, v_c)$ is a matrix of unknown cluster centers, U is a membership matrix, u_{ik} is the membership value of x_k in cluster $i(i = 1, \ldots, c)$, and the weighting exponent m in $[1, \infty]$ is a constant that influences the membership values and is generally called a 'fuzzifier'.

In each iteration, it is necessary to amend the cluster centroids using Eq. 2, and given the new centroids, it is also necessary to amend membership values using Eq. 3. The stop condition of the algorithm uses the error between the previous and current membership values.

$$v_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m} \qquad (2)$$

$$u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{\| x_k - v_i \|^2}{\| x_k - v_j \|^2}\right)^{\frac{2}{m-1}}\right]^{-1} \qquad (3)$$

Cluster validation indices

Cluster validation indices are used to determine the optimal number of clusters (c) in a data set. In this study, four cluster validation indices, namely Partition Coefficient (V_{PC}), Partition Entropy (V_{PE}), Fukuyama-Sugeno Index (V_{FS}) and Xie-Beni Index (V_{XB}) are computed for different values of c. The validity indices V_{PC} and V_{PE} were proposed by Bezdek (1974), and are defined as:

$$V_{PC}(U) = \frac{1}{n}\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}^2 \qquad (4)$$

$$V_{PE}(U) = \frac{1}{n}\left[\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\log_a(u_{ik})\right] \qquad (5)$$

The range of variation of V_{PC} and V_{PE} is $[1/c, 1]$ and $[0, \log_a c]$, respectively. An optimal partition corresponds to a maximum value of V_{PC}, which suggests minimum overlap between cluster elements. On the contrary, minimum value of V_{PE} suggests a good partition, which corresponds to a harder partition.

Fukuyama and Sugeno (1989) developed a new validity index specifically for the FCM method. In this measure, the minimum value of V_{FS} suggests optimal partitioning.

$$V_{FS}(U,C:X) = \sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^m \| v_i - x_k \|^2 - \sum_{k=1}^{n}\sum_{i=1}^{c}u_{ik}^m \| v_i - \bar{v} \|^2 \qquad (6)$$

Xie and Beni (1991) proposed a validity index that focussed on two properties: compactness and separation. In their equation for V_{XB} (Eq. (7)), the numerator reveals the compactness of the fuzzy partition, while the denominator reveals the strength of the separation between clusters. A minimum value of V_{XB} suggests optimal clustering.

$$V_{XB}\left(U,V:X\right)=\frac{\sum_{j=1}^{n}\sum_{i=1}^{c}u_{ij}^{2}\parallel x_j-v_i\parallel^2}{n\left[min_{i\neq k}\left(\parallel v_i-v_k\parallel^2\right)\right]} \tag{7}$$

Fuzzy Minimals algorithm

While most of the analytic fuzzy clustering methods used are derived from Bezdek's FCM algorithms, Bezdek showed that the solution obtained through the FCM algorithm may or may not provide the desired solution, suggesting that the method was not based on a completely reliable criteria. Flores-Sintas et al. (1998, 1999, 2000) analysed this possibility and found, based on local geometrical properties, the way to reformulate the FCM algorithm.

In Soto et al. (2008), the computation of such membership probabilities was improved by a new membership function which also reflects the relative position of an object with respect to each group. The objective function is defined as:

$$J_{FM}=\sum_{x\in X}\mu_{xv}d_{xv}^{2} \tag{8}$$

where d_{xv} is the Euclidean distance from the x to prototype v, and μ represents the membership functions obtained by:

$$\mu_{xv}=\frac{1}{1+r^2d_{xv}^2} \tag{9}$$

where r is the factor which normalises the Euclidean distance, making the sample mean density equal to 1 (Flores-Sintas et al., 1998). Finally, the objective function will be minimised in:

$$v=\frac{\sum_{x\in X}\left(\mu_{xv}\right)^2 x}{\sum_{x\in X}\left(\mu_{xv}\right)^2} \tag{10}$$

which will be the prototypes.

That algorithm is not derived from FCM due to the fact that it does not need to consider a concrete number of prototypes, although it has a similar way of working (Timón et al., 2016). This algorithm is called Fuzzy Minimals (FM) and is completely described in Soto et al. (2008).

Hosking and Wallis homogeneity test

The homogeneity of an identified region was determined using a heterogeneity measure (H) based on L-moments. L-moments are analogues to the traditional moments (mean, standard deviation, skewness, kurtosis), but they are computed on the basis of the linearly arranged data sample, and they show more favourable statistical features in comparison with the traditional moments (Hosking and Wallis, 1993). For heterogeneity testing, a four-parameter kappa distribution is fitted to the regional data set generated from a series of 500 simulations of region data by numerical simulation. The heterogeneity measure compares the dispersion between observed and simulated data. Hosking and Wallis (1997) suggest that a group of sites may be regarded as 'acceptably homogeneous' if $H<1$, 'possibly heterogeneous' if $1\leq H<2$, and 'definitely heterogeneous' if $H\geq 2$.

STUDY AREA AND DATA

Alto Genil Watershed

The Alto Genil watershed area, located in the South of Spain within the Guadalquivir River watershed, is characterised by being surrounded by mountain systems with steep slopes, due to the presence of the massif of Sierra Nevada, which provides a relative abundance of surface water. It is a typical Mediterranean landscape characterised by fragile natural ecosystems, insufficient rainfall for fast vegetation recovery and long-term human exploitation. Water resources circulate through its main course, the Genil river, along with its tributaries, Dilar, Monachil, Aguas Blancas, Darro, Cubillas and Velillos river, draining an area next to the 3000 km^2 (Fig. 1). Parts of these resources are regulated by Canales, Quéntar, Cubillas and Colomera reservoirs, where flow rates are continuously measured.

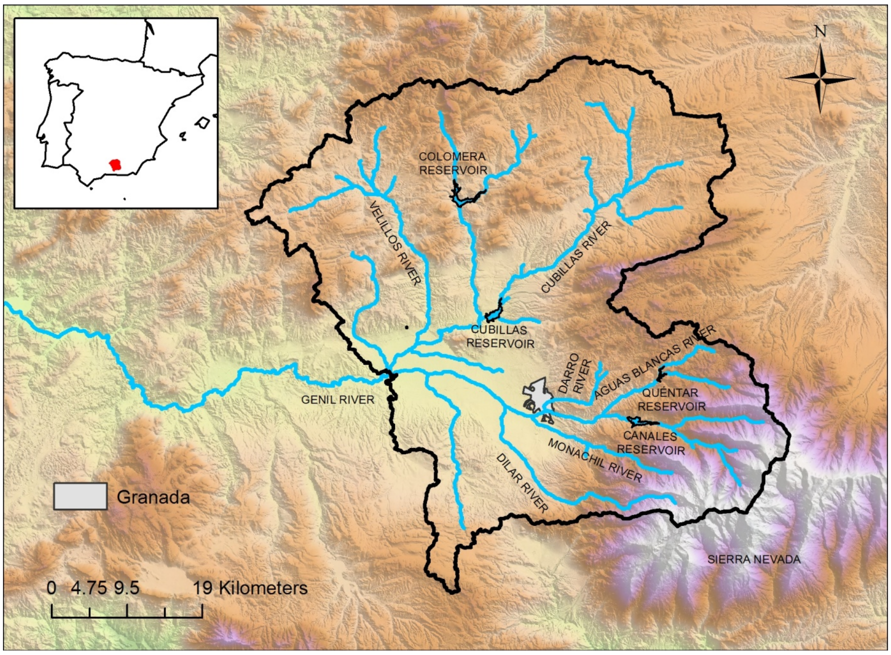

Fig. 1. Study area.

Most of these water resources are used for urban water supply in the city of Granada and surrounding areas (500,000 inhabitants), reserving the surplus, the wastewater and groundwater for irrigation. In this study, the Alto Genil watershed was discretized into 20 subcatchments.

Data used in the study

Many attributes have been used for RFFA. These attributes can be classified according to (i) geographical location attributes such as latitude, longitude and mean basin elevation; (ii) physiographic watershed attributes such as stream density, stream length, drainage area and average basin slope; (iii) meteorological factors such as mean annual rainfall and precipitation intensities; (iv) soil cover attributes such as runoff coefficient and infiltration potential; and (iv) at-site flood statistics. In order to choose what attributes could be used in fuzzy clustering analysis, we have taken into account that, according to Burn et al. (1997), the use of flood-related attributes in regionalisation of watersheds and the posterior evaluation of the homogeneity of the derived regions using the same or related flood statistics suggests that the obtained regions may appear homogeneous but may not be effective for RFFA. Besides, physiographic watershed attributes are due to the activity of water in these watersheds, so it is logical to consider a substantial relationship between the physical and geographical attributes of the watershed and the variables that describe the hydrological nature of the watershed; though, according to Rao and Srinivas (2008), the formation of regions should not be based exclusively on physiographic attributes because affinity in only physiographic characteristics does not necessarily imply affinity in watershed hydrologic response.

Principal component analysis (PCA) was employed to examine dominant patterns of intercorrelation among the attributes and identify subsets of attributes that describe the major sources of variation while minimizing redundancy. The PCA was conducted using the following attributes: mean basin elevation, drainage density, modified Fournier index, curve number, 24-hr rainfall having a recurrence interval of 2 years, average monthly temperature and average basin slope. The attributes are reduced using the PCA, which resulted that first four attributes explained 99.60% of the variance. Therefore, the features extracted for cluster analysis are one location attribute (mean basin elevation), one physiographic attribute (drainage density), one meteorological attribute (Modified Fournier index) and one attribute related to soil type (curve number). This way, every attribute group is represented in the fuzzy clustering analysis. Information related to rainfall and flow data is extracted from the national water information system website (MAGRAMA, 2016). To enhance the acceptability of the research findings, missing data points not more than 10% were used. The block maxima method was used to derive annual maximum flow data. There is an available observation period of 70 years (between 1 October 1940 and 30 September 2010). Topographic maps at different scales, Digital Elevation Model (DEM), land use and soil hydrologic groups were also collected from various sources. Mean basin elevation (Rao and Srinivas, 2006; Basu and Srinivas, 2015) and drainage density (Srinivasa Raju and Nagesh Kumar, 2011) have been used before as attributes for cluster analysis; however, the use of Modified Fournier Index (MFI) and curve number (CN) is new for cluster analysis purposes.

Fournier developed an erosivity index (FI) using monthly rainfall data (Fournier, 1960) for correlation with sediment loads in rivers. This index is universally used due to the availability of monthly rainfall.

$$FI = \frac{p_{max}^2}{P} \tag{11}$$

where p_{max} is the mean monthly rainfall of the wettest month of the year and P is the mean annual rainfall. This index has some restrictions for the evaluation of the rainfall erosivity. As low amounts of rainfall also have erosive power, an increase in total rainfall amount should result in an increase of erosivity. If p_{max} remains the same with P increasing, FI decreases, and this is unreasonable (Gabriels, 2006). That is why Arnoldus (1980) modified this index and called it 'MFI'. It is defined by:

$$MFI = \frac{\sum p_m^2}{P} \tag{12}$$

where p_m is the monthly rainfall for each month (m = 1,2,...,12) and P is the corresponding total annual precipitation. The MFI indicates a degree of rainfall aggressivity. The MFI has erosivity classes that are: very low (MFI = 0–60), moderate (MFI = 90–120), high (MFI = 120–160) and very high (MFI > 160).

The CN is an empirical parameter used in hydrology for calculating direct runoff or infiltration from rainfall excess (Hawkins et al., 2009). It depends on three basic variables: soil group that defines the potentiality of runoff based on the hydraulic conductivity of the soil, soil cover and its condition. CN has a range from 30 to 100; lower numbers indicate low runoff potential, while larger numbers are for increasing runoff potential.

RESULTS AND DISCUSSION

The mean basin elevation, drainage density, MFI and CN were included in the feature vector to identify the regions that are hydrologically homogeneous. The range of each of these attributes is presented in Table 2.

Table 2. Attributes considered in the study.

Attribute	Minimum	Mean	Maximum
Mean basin elevation (m a.s.l.)	615.15	1117.54	2129.16
Drainage density (km/km^2)	3.50	26.43	50.91
Modified Fournier Index	59.64	90.08	123.46
Curve number	43.60	57.04	67.90

Since variables with different units generally influence the clustering results, they have to be rescaled before entering the cluster analysis in order to have the final results influenced in an equal way (Dikbas et al., 2012). The transformation function used to normalise data was:

$$X_{ij}^N = \frac{X_{ij} - X_{i,min}}{X_{i,max} - X_{i,min}} \tag{13}$$

where X_{ij} is the i^{th} attribute of j^{th} watershed; $X_{i,min}$ is the minimum i^{th} attribute in all watersheds; $X_{i,max}$ is the maximum i^{th} attribute in all watersheds; and X_{ij}^N is normalised i^{th}

attribute of j^{th} watershed. Equal weight was assigned to all the attributes, implying equal importance to all the attributes.

The cluster analysis using FCM algorithm was started by choosing the number of clusters as at least 2, and the optimum number of clusters were sought by increasing the number of clusters to 10. Related to FCM algorithm fuzzifier, the value of fuzzifier was set to 2. Pal and Bezdek (1995) mentioned that the FCM provides better performance for a fuzzifier in the range 1.5–2.5. The optimal number of clusters for a data set was determined by applying various fuzzy cluster validation measures such as V_{PC}, V_{PE}, V_{FS} and V_{XB}. The corresponding results of applying these measures are shown in Table 3.

Table 3. Comparison of different cluster validity measures for varying number of clusters.

Number of Clusters	V_{PC}	V_{PE}	V_{XB}	V_{FS}
2	**0.843**	**0.268**	0.071	**107482.40**
3	0.803	0.367	0.082	164236.28
4	0.796	0.414	**0.065**	185455.45
5	0.769	0.493	0.142	184826.76
6	0.775	0.494	0.100	200775.14
7	0.765	0.522	0.084	200057.68
8	0.729	0.616	0.313	180094.50
9	0.767	0.547	0.333	198375.76
10	0.748	0.588	0.183	203933.44

V_{PC}: Partition Coefficient; V_{PE}: Partition Entropy; V_{XB}: Xie-Beni Index; V_{FS}: Fukuyama-Sugeno Index. The maximum value of V_{PC} suggests optimal clustering. On the contrary, the minimum value of V_{PE}, V_{XB} and V_{FS} corresponds to an optimal partition.

V_{PC}, V_{PE} and V_{FS} clearly suggest two clusters as the best partition, irrespective of the structure in the data being analysed. The V_{XB} measure weakly suggests four clusters as the best partition; this result is very similar for two clusters. According to these results, the delineation of homogeneous regions using the FCM algorithm was computed for two different cases. In the first case, the number of predefined clusters was two, taking into account the results obtained from cluster validity measures V_{PC}, V_{PE} and V_{FS}. In the second case, four was the number of predefined clusters, according to the indications of V_{XB} index.The obtained regions in both cases are shown in Fig. 2 (a) and (b).

FCM clustering results using two clusters are clearly influenced by elevation attributes. One of the clusters (FCM2_2) is formed by headwaters, while the other cluster (FCM2_1) is formed by watersheds whose elevation is lower. When results using four clusters are analysed, we realise that cluster FCM2_2 is divided in two groups in a logical way due to the fact that watersheds that formed cluster FCM4_4 are those with a higher amount of precipitation in the form of snow. FCM2_1 cluster is also divided in two different clusters (FCM4_2 and FCM4_3). Overall, both kinds of clustering are consistent.

As apparent in Fig. 2 (c), FM algorithm identifies three different clusters from the analysis of membership values. The membership value in each group indicates the probability for the watershed to be included in that specific group (Rao and Srivinas, 2008). Membership values of the 20 watersheds under each of the 15 groups are presented in Table 4. The group which is having the highest membership value among the 15 groups is the representative group for that watershed. The sum of the membership values per watershed should be equal to 1 (Ross, 1995). For instance, the representative group for watershed 9 is group number 4 (having the maximum membership value of 0.5485). Similarly, all other watersheds were analysed. Twelve out of 15 groups are formed by only one watershed, so it can be interpreted that those watersheds do not present similarities with other watersheds in Alto Genil. Therefore, the FM algorithm identifies 8 watersheds forming three different regions (FM_4, FM_6 and FM_7). These regions are not geographically contiguous, but according to Nathan and McMahon (1990), subregions defined on the basis of similarity of hydrologic or physiographic characteristics may not have geographical significance.

As can be seen in Table 5, homogeneity test proposed by Hosking and Wallis is applied to the clusters generated by FCM algorithm and FM algorithm. When FCM algorithm is applied, assuming that 2 is the optimal number of clusters, one of those regions is heterogeneous ($H = 3.64$) and the other one is acceptably homogeneous ($H = 0.28$). If it is considered that 4 is the optimal number of clusters, two of those regions (FCM4_2 and FCM4_3) are highly heterogeneous: $H = 6.63$ and $H = 9.23$, respectively. However, FCM4_1 and FCM4_4 are acceptably homogeneous. According to Rao and Srinivas (2006),

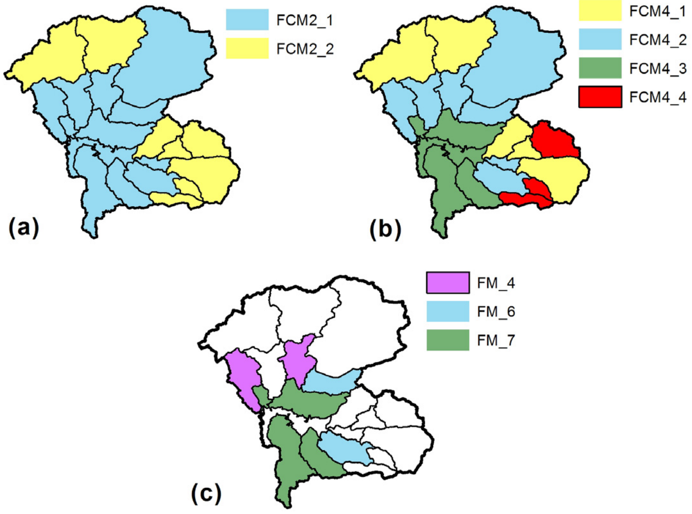

Fig. 2. Spatial distribution of fuzzy clusters in Alto Genil obtained from fuzzy cluster analysis. (a) FCM clustering results using two clusters; (b) FCM clustering results using four clusters; (c) Fuzzy Minimals results.

Table 4. Membership values of the watersheds under each group showing the representative group of each watershed (in boldface).

Watershed	Group														
	FM_1	FM_2	FM_3	FM_4	FM_5	FM_6	FM_7	FM_8	FM_9	FM_10	FM_11	FM_12	FM_13	FM_14	FM_15
1	**0.6633**	0.0099	0.0591	0.0174	0.0647	0.0449	0.0043	0.0075	0.0688	0.0199	0.0119	0.0050	0.0022	0.0047	0.0164
2	0.0087	**0.5731**	0.0047	0.0953	0.0176	0.0254	0.0331	0.0020	0.0056	0.0033	0.2178	0.0016	0.0071	0.0015	0.0032
3	0.0542	0.0048	**0.6136**	0.0072	0.0166	0.0134	0.0025	0.0158	0.1266	0.0685	0.0056	0.0089	0.0015	0.0082	0.0526
4	0.0183	0.0992	0.0079	**0.5660**	0.0499	0.0872	0.0145	0.0028	0.0098	0.0051	0.1256	0.0022	0.0048	0.0021	0.0046
5	0.0656	0.0197	0.0182	0.0455	**0.6424**	0.1211	0.0066	0.0047	0.0236	0.0101	0.0241	0.0034	0.0029	0.0032	0.0089
6	0.0463	0.0368	0.0162	0.0955	0.1223	**0.5639**	0.0097	0.0047	0.0229	0.0095	0.0529	0.0035	0.0040	0.0033	0.0085
7	0.0097	0.0892	0.0060	0.0401	0.0156	0.0194	**0.7008**	0.0030	0.0069	0.0046	0.0582	0.0025	0.0372	0.0024	0.0044
8	0.0126	0.0665	0.0084	0.0381	0.0183	0.0221	**0.5976**	0.0045	0.0094	0.0065	0.0508	0.0038	0.1515	0.0037	0.0062
9	0.0152	0.1221	0.0068	**0.5485**	0.0425	0.0667	0.0148	0.0025	0.0085	0.0045	0.1551	0.0020	0.0047	0.0019	0.0042
10	0.0075	0.0022	0.0170	0.0028	0.0046	0.0041	0.0014	**0.6587**	0.0130	0.0361	0.0025	0.1118	0.0009	0.0832	0.0542
11	0.0645	0.0061	0.1284	0.0092	0.0224	0.0195	0.0029	0.0121	**0.6234**	0.0497	0.0073	0.0073	0.0016	0.0066	0.0390
12	0.0186	0.0035	0.0682	0.0048	0.0094	0.0079	0.0020	0.0329	0.0495	**0.6170**	0.0040	0.0162	0.0012	0.0137	0.1511
13	0.0105	0.2214	0.0054	0.1161	0.0216	0.0377	0.0188	0.0022	0.0068	0.0038	**0.5438**	0.0017	0.0053	0.0016	0.0033
14	0.0044	0.0016	0.0084	0.0019	0.0030	0.0027	0.0010	0.1018	0.0069	0.0154	0.0017	**0.5634**	0.0007	0.2682	0.0189
15	0.0055	0.0389	0.0036	0.0197	0.0086	0.0104	**0.8358**	0.0019	0.0040	0.0028	0.0274	0.0015	0.0359	0.0015	0.0025
16	0.0030	0.0098	0.0021	0.0069	0.0041	0.0046	0.0338	0.0013	0.0024	0.0018	0.0082	0.0011	**0.9182**	0.0011	0.0016
17	0.0060	0.0439	0.0038	0.0221	0.0095	0.0113	**0.8234**	0.0020	0.0043	0.0030	0.0301	0.0016	0.0346	0.0016	0.0028
18	0.0045	0.0016	0.0083	0.0020	0.0030	0.0028	0.0011	0.0739	0.0067	0.0135	0.0018	0.2654	0.0008	**0.5978**	0.0168
19	0.0512	0.0287	0.0156	0.0741	0.1581	**0.5716**	0.0079	0.0043	0.0217	0.0089	0.0409	0.0031	0.0033	0.0030	0.0076
20	0.0153	0.0033	0.0521	0.0043	0.0081	0.0070	0.0019	0.0513	0.0387	0.1517	0.0036	0.0206	0.0011	0.0174	**0.6236**

regions given by clustering algorithms are, in general, not statistically homogeneous. Consequently, further adjustment is required to form the homogeneous region. There are different options for adjusting the regions formed by cluster analysis (Hosking and Wallis, 1997), such as deleting or shifting some sites of any region to some other region or dividing one region into two or more regions. Adjustment of cluster is found to be very useful in improving the results found via FCM (Goyal and Gupta, 2014). Nevertheless, the application of the regional homogeneity test to the regions formed by FM algorithm shows that all the regions can be define as acceptably homogeneous, so further adjustment is not needed.

Table 5. Results of regional homogeneity test.

Algorithm	CC	NW	Heterogeneity Measure (H)	Region type
FCM (2 Clusters)	FCM2_1	12	3.64	Definitely Heterogeneous
	FCM2_2	8	0.28	Acceptably Homogeneous
FCM (4 Clusters)	FCM4_1	5	–0.51	Acceptably Homogeneous
	FCM4_2	7	6.63	Definitely Heterogeneous
	FCM4_3	5	9.23	Definitely Heterogeneous
	FCM4_4	3	0.91	Acceptably Homogeneous
FM	FM_4	2	0.46	Acceptably Homogeneous
	FM_6	2	0.98	Acceptably Homogeneous
	FM_7	4	–0.26	Acceptably Homogeneous

CC: Cluster Code; NW: Number of Watersheds.

CONCLUSIONS

In this paper, Fuzzy C-Means and Fuzzy Minimals algorithms were applied to check the effectiveness of both algorithms in the identification of hydrologically homogeneous regions for flood frequency analysis. This was illustrated through its application to annual maximum flow data from the watersheds in Alto Genil (South Spain). For Fuzzy C-Means Approach, optimum numbers of clusters were analysed by using four fuzzy cluster validity indices, namely partition coefficient, partition entropy, Xie-Beni index and Fukuyama-Sugeno on 20 stations with 4 attributes. The regional homogeneity of the regions identified by FCM was tested using Hosking and Wallis homogeneity test. It was found from the regional homogeneity test that further adjustment is required to form homogeneous regions. FM algorithm was also applied, and the homogeneity of the regions formed was also tested. Formulation of FM algorithm, proposed by Flores-Sintas et al. (1998), uses an objective function (shown in Eq. (8)) different from the one that is used in the FCM algorithm (shown in Eq. (1)). In contrast to FCM objective function, FM objective function does not need to know the number of prototypes, as is evident. In this case, there is no need to use cluster validity indices, and regions formed by this algorithm were acceptably homogeneous. When the performances of Fuzzy C-Means and Fuzzy Minimals method are compared for the case study presented here, it was seen that the regions identified by Fuzzy Minimals method are more homogeneous than those identified by Fuzzy C-Means method.

REFERENCES

Agarwal, A., Maheswaran, R., Sehgal, V., Khosa, R., Sivakumar, B., Bernhofer, C., 2016. Hydrologic regionalization using wavelet-based multiscale entropy method. J. Hydrol., 538, 22–32.

Arnoldus, H.B.J., 1980. An approximation of the rainfall in the universal soil loss equation. In: De Boodt, M., Gabriels, D. (Eds.), Assesment of Erosion. John Wiley & Sons, Chichester, pp. 127–132

Bargaoui, Z.K., Fortin, V., Bobée, B., Duckstein, L., 1998. A fuzzy approach to the delineation of region of influence for hydrometric stations. Revue des sciences de l'eau 11, 2, 255–282. (In French.)

Basu, B., Srinivas, V.V., 2014. Regional flood frequency analysis using kernel-based fuzzy clustering approach. Water Resour. Res., 50, 4, 3295–3316.

Basu, B., Srinivas, V.V., 2015. Analytical approach to quantile estimation in regional frequency analysis based on fuzzy framework. J. Hydrol., 524, 30–43.

Bellman, R., Kalaba, R., Zadeh, L.A., 1966. Abstraction and pattern classification. J. Math. Anal. Appl., 2, 581-585.

Bezdek, J.C., 1974. Cluster validity with fuzzy sets. J. Cybernet., 3, 3, 58–73.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, 266 p.

Burn, D.H., Zrinji, Z., Kowalchuk, M., 1997. Regionalization of catchments for regional flood frequency analysis. J. Hydrol. Eng., 2, 2, 76–82.

Dikbas, F., Mahmut, F., Cem, K., Gungor, M., 2012. Classification of precipitation series using fuzzy cluster method. Int. J. Climatol., 32, 1596–1603.

Dunn, J.C., 1974. A fuzzy relative of the ISODATA process and its use in detecting compact well separated clusters. J. Cybernet., 3, 3, 32–57.

Flores-Sintas, A., Cadenas, J.M., Martin, F., 1998. A local geometrical properties application to fuzzy clustering. Fuzzy Sets and Systems, 100, 237–248.

Flores-Sintas, A., Cadenas, J.M., Martin, F., 1999. Membership functions in the Fuzzy C-Means algorithm. Fuzzy Sets and Systems, 101, 49–58.

Flores-Sintas, A., Cadenas, J.M., Martin, F., 2000. Partition validity and defuzzification. Fuzzy Sets and Systems, 112, 433–447.

Fournier, F., 1960. Climat et érosion. La relation entre l´érosion du sol par l´eau et les précipitations atmosphériques. [Relationship between soil erosion by water and rainfall]. Presse Universitaire de France, Paris. (In French.)

Fukuyama, Y., Sugeno, M., 1989. A new method of choosing the number of clusters for the fuzzy c-means method. In: Proc. 5th Fuzzy Syst. Symp., pp. 247–250 (In Japanese.)

Gaál, L., Szolgay, J., Lapin, M., Fasko, P., 2009. Hybrid approach to delineation of homogeneous regions for regional precipitation frequency analysis. J. Hydrol. Hydromech., 57, 4, 226–249.

Gabriels, D., 2006. Assessing the modified Fournier Index and the Precipitation Concentration Index for some European countries. In: Boardman, J., Poesen, J. (Eds.): Soil Erosion in Europe. John Wiley & Sons, Chichester, pp. 675–684.

Goyal, M.K., Gupta, V., 2014. Identification of homogeneous rainfall regimes in northeast region of India using Fuzzy Cluster Analysis. Water Resour. Manag., 28, 4491–4511.

Goyal, M.K., Sharma, A. 2016. A fuzzy c-means approach regionalization for analysis of meteorological drought homogeneous regions in western India. Nat. Hazards. DOI: 10.1007/s11069-016-2520-9.

Hall, M.J., Minns, A.W., 1999. The classification of hydrologically homogeneous regions. Hydrol. Sci. J., 44, 5, 693–704.

Hawkins, R.H., Ward, T.J., Woodward, D.E., Van Mullem, J.A., 2009. Curve number hydrology: state of the practice, Report of ASCE/EWRI Task Committee, American Society of Civil Engineers, Reston, Virginia, USA.

Hosking, J.R.M., Wallis, J.R., 1993. Some statistics useful in regional frequency-analisis. Water Resour. Res., 29, 2, 271–281.

Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis: An Approach based on L-Moments. Cambridge University Press, New York.

Isik, S., Singh, V.P., 2008. Hydrologic regionalization of watersheds in Turkey. J. Hydrol. Eng., 13, 824-834.

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

Kumar, R., Goel, N.K., Chatterjee, C., Nayak, P.C., 2015. Regional flood frequency analysis using soft computing techniques. Water Resour. Manage., 29, 1965.

MAGRAMA, 2016. Ministerio de Agricultura, Alimentación y Medio Ambiente. Sistema de Información del Agua. Retrieved from http://www.magrama.gob.es/es/agua/temas/planificacion-hidrologica/sia-/ (In Spanish)

Nathan, R.J., McMahon, T.A., 1990. Identification of homogeneous regions for the purposes of regionalization. J. Hydrol., 121, 217–238.

Pal, N.R., Bezdek, J.C., 1995. On cluster validity for the fuzzy c-means model. IEEE Trans. Fuzzy Syst., 3, 3, 370–379.

Raju, K.S., Nagesh Kumar, D., 2011. Classification of microwatersheds based on morphological characteristics. J. Hydro-Environ. Res., 5, 101–109.

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

Rao, A.R., Srinivas, V.V., 2008. Regionalization of Watersheds: An Approach Based on Cluster Analysis. Water Science and Technology Library Vol. 58. Springer Science & Business Media.

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

Ruspini, E.H., 1969. A new approach to clustering. Inform. and Control, 15, 22–32.

Smithers, J.C., Schulze, R.E., 2001. A methodology for the estimation of short duration design storms in South Africa using a regional approach based on L-moments. J. Hydrol., 24, 42–52.

Soto, J., Flores-Sintas, A., Paralea-Albaladejo, J., 2008. Improving probabilities in a fuzzy clustering partition. Fuzzy Sets and Systems, 159, 406–421.

Srinivas, V.V., Tripathi, S., Rao, A.R., Govindaraju, R.S., 2008. Regional flood frequency analysis by combining self-organizing feature maps and fuzzy clustering. J. Hydrol., 348, 148–166.

Timón, I., Soto, J., Pérez-Sánchez, H., Cecilia, J.M., 2016. Parallel implementation of fuzzy minimals clustering algorithm. Expert Systems with Applications, 48, 35–41.

Xie, X.L., Beni, G., 1991. A validity measure for fuzzy clustering. IEEE Trans. Pattern Anal. Mach. Intell., 13, 8, 841–847.

Zadeh, L.A., 1965. Fuzzy sets. Information and Control, 8, 3, 338–353.

Received 13 July 2016
Accepted 2 December 2016