https://doi.org/10.3311/PPci.13885
Creative Commons Attribution ⓑ

Periodica Polytechnica Civil Engineering, 63(2), pp. 674–686, 2019

Optimized Fuzzy Cmeans – Fuzzy Covariance – Fuzzy Maximum Likelihood Estimation Clustering Method Based on Deferential Evolutionary Optimization Algorithm for Identification of Rock Mass Discontinuities Sets

Akbar Esmaeilzadeh[1], Kurosh Shahriar[1*]

[1] Mining Department,
Faculty of Mining and Metallurgical Engineering,
Amir Kabir University of Technology
424 Hafez Ave, Tehran, Iran, P.O.B. 159163-4311
[*] Corresponding author, e-mail: k.shahriar@aut.ac.ir

Received: 14 February 2019, Accepted: 01 April 2019, Published online: 25 April 2019

Abstract

Detecting of joint sets (clusters) is one of the most important processes in determining properties of fractures. Joints clustering and consequently, determination of the mean value representing each cluster is applicable to most rock mass studies. It is clear that the accuracy of the clustering process plays a key role in analyzing stability of infrastructures such as dams and tunnels and so on. Hence, in this paper, by reviewing several methods proposed for clustering fractures and considering their advantages and disadvantages, a three-stage hybrid method is developed which contains Fuzzy c-means, Fuzzy covariance and Fuzzy maximum likelihood estimation that by utilizing the modified orientation matrix had been optimized. This method is optimized by the Differential Evolutionary algorithm using a new and strong cost function which is defined as the computation core. In addition, using three clustering quality comparing criteria, the new developed method of differential evolutionary optimized of fuzzy cmeans - fuzzy covariance - fuzzy maximum likelihood estimation clustering method (DEF3) is compared with other base and common methods using field data. After doing the calculations, the developed method by giving the best values for all the criteria provided the best results and good stability in meeting different criteria. The DEF3 method was validated using actual field data which mapped in Rudbar Lorestan dam site. The results revealed that DEF3 acquired the best rank among the other method by getting the value of 0.5721 of Davis-Bouldin criterion, 1403.1 of Calinski-Harabasz criterion, and 0.83482 of Silihotte as comparing criteria of clustering methods.

Keywords

joint sets, optimized clustering, fuzzy logic, differential evolutionary, DEF3

1 Introduction

The existence of fractures in rock masses has made rock mechanics a unique science. Fractures are the main determinants of rock mass behavior. Mechanical and hydraulic behavior of rock mass is mainly controlled by fractures. Therefore, the stability of structures involved in or on rock mass such as rock slopes, tunnels and caverns are directly related to fractures states. Hence, the recognition and study of fractures is critical in rock mechanics.

Due to the statistical nature of the fractures, they are studied and analyzed through field surveying of their properties using standard mapping techniques. Orientation is among the most important properties of fractures, which is presented in terms of dip and dip direction. The importance of dips and dip directions of fractures arises from the key role which they play in fractures clustering. Identification of fractures clusters and access to mean dip and dip direction of clusters is considered as the most important step in statistical analysis of field data. Grouping surveyed data of joints is key part of identification of preferred plane direction which is used in statically and dynamically analysis and hydraulic behavior and other engineering characterizations of rock mass. Hence cluster analyses in order to identification of joint sets is the main part of statistical analysis of field data [1–6].

Cite this article as: Esmaeilzadeh, A., Shahriar, K. "Optimized Fuzzy Cmeans – Fuzzy Covariance – Fuzzy Maximum Likelihood Estimation Clustering Method Based on Deferential Evolutionary Optimization Algorithm for Identification of Rock Mass Discontinuities Sets", Periodica Polytechnica Civil Engineering, 63(2), pp. 674–686, 2019. https://doi.org/10.3311/PPci.13885

Due to the importance of the subject, many efforts have been made to achieve the best clustering method and consequently present the most accurate representative dip and dip direction of the joint sets. Generally, fracture clustering methods are divided into two groups. The first category consists of manual or counting techniques by which, dependent on the fracture poles location on the lower hemisphere equal area, the clustering takes place based on the proximity of poles and personal recognition. Manually fracture clustering is completely rest on experience and personal accuracy and It is not also free from human errors since it is based on the user's preference and diagnosis. In addition, in the case of more scattered data, the level of human error in detecting the boundaries of clusters is not likely negligible [7].

The limitations of manual recognition of joint sets and the importance of accurate clustering, have led researchers to focus their studies on non-manual and automatic diagnostic techniques, which in fact develops the second category of fracture clustering methods [8]. Automatic techniques are classified in two categories: the methods that assume the statistical structure of fractures orientation data [5, 9, 10], and the methods that classify them without regard to the default statistical model of data [11–15].

The methods of first category attempt to detect joint sets by using automatic technique based on orientational properties of joints was made by Shanely and Mahtab [5]. The basis of this method is to select the appropriate radius of the small search circle to find the high density that will be used later in the process to calculate the core of the joint set. The efficiency of this automatic method is highly dependent on the search circle radius chosen by the user, which shows the high effect of individual judgment in this method [5].

Mahtab and Yegulalp introduced a new algorithm in 1982 to cluster joints considering the rejection criterion in the definition of fracture clusters based on random test steps, using the Poisson distribution function, detection of high-density clusters, and the estimation of the clusters statistical structure by Bingham distribution. Clearly, it seems that imposing Bingham's distribution on the statistical structure of the clusters is not always a logical assumption and is not necessarily always true [9].

Dershowitz et al. proposed an iterative-based process for joint clustering by assumptions about the fractures statistical structures so that they had a direct effect on the performance of the method [16]. However, access to the actual statistical model of the fractures is very difficult in some cases. The above methods, operating based on compression density of poles, by counting the number of poles placed within a reference circle. The main problem of these methods, however, was that the clustering result was strongly dependent on the size of the reference circle and the user's preference [11, 12, 17, 18].

In brief, disadvantages of methods based on counting poles have led to the development of alternate methods for automatic detection of joint sets. Marcotte and Henry assuming that the orientation of each joint sets can be modeled using a truncated integrated bivariate normal distribution function, proposed an algorithm for identifying joint sets. However, it is difficult and sometimes false to consider a definitive default in relation to the statistical structure of the fractures [10].

Given the above deficiency, the researchers' attempted to introduce the clustering methods without requiring the statistical structure of fracture data. In the process of identifying joint sets, the spatial distance of fractures orientation vectors is the most distinctive parameter [11]. Harrison used a fuzzy logic to analyze the orientational data of fractures. Based on fuzzy theory, Harrison presented a three-stage technique for clustering fractures using the methods proposed by Bezdek [19], Gustafson and Kessel [20], Gath and Geva [21]; which is called FCM, F-Co, and FMLE methods, respectively [13]. Harrison's three-step algorithm, despite its capabilities, has two major flaws: first the lack of good compatibility with the vectoral nature of fractures statistical data and poor performance in handling data with a spherical cluster structure [22]. To fix mismatch of FCM method with directional data of fractures orientation, Hammah and Curran suggested a modified orientation matrix that improved the automatic detection of fractures by FCM. However, in his later work, he discussed issues related to the operation of different distance-meters and validation indicators of fuzzy clustering methods [11, 12, 17, 18].

Sirat and Talbot proposed a new technique for clustering fracture data by using the neural network method. They applied the capability of neural networks self-organizing maps for clustering real data [23]. Meanwhile, a group of researchers, in addition to orientational data of joints in the clustering process, introduced other properties of fractures.

Zhou and Maerz [15] and Tokhmechi et al. [24, 25], using other properties of fractures, along with dip and dip direction, and based on the K-means method presented other algorithms for fracture clustering. Zhou and Maerz developed a computer program to detect and cluster fractures by several multivariate clustering algorithms. It should be noted that FCM and K-means were dynamic clustering algorithms that required the initial proposal of number and center of clusters to start the search process, though both methods are not guaranteed to reach the global optimum points and likely they would get stuck at the optimal local points. The results of the clustering process also depend on the number of clusters and cluster nuclei location. Without a clear boundary between clusters, choosing the appropriate cluster center is difficult and the results are untrustworthy [13]. In order to avoid problems due to the false selection of cluster nuclei, Jimenez-Rodriguez and Sitar [1], proposed a spectral clustering method for identification of fracture clusters using a proximity matrix to determine the similarity between two fractures. In this method, by transferring data to the secondary space using the values and the proximity matrix, the boundary between fracture clusters changes to a clear mode from a non-transparent space. Xu et al., employed the chaos theory optimization method to cluster fractures [26].

Klose et al. developed a fracture clustering technique based on quantization of vector and minimization of cost function, which was defined by the acute angle between fractures poles and the mean of poles. Their technique operated by the arc length between the poles of fractures on single unit spheres [14]. Based on fuzzy theory, Hammah and Curran provided fuzzy K-means method. Due to the improved distance meters and the new fuzzy clustering evaluator, their method classified fractures [11, 12, 17, 18]. Based on the genetic optimization algorithm, Cai et al. presented algorithm to obtain cluster nuclei and assess clustering efficiency [27]. Ma et al. proposed a reinforced K-Means clustering technique by which the clusters' cores were determined more precisely based on the meliorated initial centers and hierarchical clustering method [28].

In this paper, considering the strengths and deficiencies of the previous methods, a new method is developed based on the Harrison's three-step technique, optimized by differential evolutionary algorithm [13]. It has a good speed of convergence and global extremum point search, along with simplicity and compatibility with the spatial and spherical nature of rock engineering data. The developed method utilizes strengths of the Harrison method, which in fact includes advantages of the Bezdek's FCM [19, 29, 30], Gustafson-Kessel' FCo [20, 31–33] and Gath-Geva's FMLE [21] methods, and integrates the concept of orientation matrix proposed by Hammah and Curran in the clustering process to further adapt the method to the nature of the orientation data. Accordingly, new proposed technique performs more accurate than the methods developed by Shanely and Mahtab, Yanan Li et al, Jimenez and Sitar, and Xu et al., which were selected because of their high efficiency [1, 5, 22, 26]. In the following, the steps of the proposed new method of differential evolutionary optimized fuzzy cmeans- fuzzy covariance- fuzzy maximum likelihood estimation (DEF3) is presented and then, using the mapped actual data and comparing criteria, its efficiency is examined versus common clustering methods.

2 Materials and methods

2.1 Assessing clusterability of the data

Relating to the data clustering process, ignoring quality and quantity of process, first step includes study of data clusterability which evaluates whether data possesses such structure, is an integral part of cluster analysis. However, methods for evaluating clusterability vary radically, making it challenging to select a suitable measure. In this method, we use a most common statistics parameter which called Hopkins statistic. The Hopkins statistic is a way of measuring the cluster tendency of a data set [34]. It belongs to the family of sparse sampling tests. It acts as a statistical hypothesis test where the null hypothesis is that the data is generated by a Poisson point process and are thus uniformly randomly distributed. A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0. A typical formulation of the Hopkins statistic follows: Let X be the set of n data points. Consider a random sample (without replacement) of m << n data points with members x_i. Generate a set Y of m uniformly randomly distributed data points. Define two distance measures: u_i the distance of $y_i \in Y$ from its nearest neighbour in X, w_i and the distance of $x_i \in Y$ from its nearest neighbour in X. With the above notation, if the data is dimensional, then the Hopkins statistic is defined as [35]:

$$H = \frac{\sum_{i=1}^{m} u_i^d}{\sum_{i=1}^{m} u_i^d + \sum_{i=1}^{m} w_i^d} . \quad (1)$$

2.2 Determination of optimum clusters number

After ensure data clusterability, one of the most important steps in data clustering process is determination of proper cluster number which is as clustering process input. In this paper for achieving proper cluster number, we apply one of the most useful technique which is called Elbow method.

The Elbow method is a method of interpretation and validation of consistency within cluster analysis designed to help finding the appropriate number of clusters in a dataset. This method looks at the percentage of variance explained as a function of the number of clusters. One should choose a number of clusters so that adding another cluster doesn't give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information (explain a lot of variance), but at some point, the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point, hence the “elbow criterion”. This “elbow criterion” cannot always be unambiguously identified [36]. Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test. A slight variation of this method plots the curvature of the within group variance [37–39] (Eq. 2):

$$F = \frac{explained\ variance}{unexplained\ variance} = \frac{\sum_{i=1}^{K}\left(n_i\left(\bar{Y}_i - \bar{Y}\right)\right)/\left(K-1\right)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(\left(\bar{Y}_{ij} - \bar{Y}_i\right)\right)/\left(N-K\right)}. \quad (2)$$

Where $\bar{Y}_i$ denotes the sample mean in the i-th group, n_i is the number of observations in the i-th group, $\bar{Y}$ denotes the overall mean of the data, and K denotes the number of groups, $\bar{Y}_{ij}$ is the j-th observation in the i-th out of K groups and N is the overall sample size. The statistic will be large if the between-group variability is large relative to the within-group variability, which is unlikely to happen if the population means of the groups all have the same value.

2.3 Differential Evolutionary algorithm

DE is Stochastic, population-based optimization algorithm [40]. DE is comparatively a recent addition to class of population-based search heuristics. Nevertheless, it has emerged as one of the techniques most favored by engineers for solving continuous optimization problems. DE has several attractive features. Besides being an exceptionally simple evolutionary strategy, it is significantly faster and robust for solving numerical optimization problems and is more likely to find the function's true global optimum. Also, it is worth mentioning that DE has a compact structure with a small computer code and has fewer control parameters in comparison to other evolutionary algorithms. Originally Price and Storn proposed a single strategy for DE, which they later extended to ten different strategies [40, 41]. DE has been successfully applied to a wide range of engineering problems [42].

A basic variant of the DE algorithm works by having a population of candidate solutions (called agents). These agents are moved around in the search-space by using simple mathematical formulae to combine the positions of existing agents from the population. If the new position of an agent is an improvement it is accepted and forms part of the population, otherwise the new position is simply discarded. The process is repeated and by doing so it is hoped, but not guaranteed, that a satisfactory solution will eventually be discovered. Formally, let $f: R^n \rightarrow R$ be the cost function which must be minimized or fitness function which must be maximized. The function takes a candidate solution as argument in the form of a vector of real numbers and produces a real number as output which indicates the fitness of the given candidate solution. The gradient of f is not known. The goal is to find a solution m for which $f(m) \leq f(p)$ for all p in the search-space, which would mean m is the global minimum. Maximization can be performed by considering the function $h := -f$ instead. Let $x \in R^n$ designate a candidate solution (agent) in the population. $CR \in [0.1]$ which is called the crossover probability. Let $F \in [0.2]$ be called the differential weight. Both these parameters are chosen by the practitioner along with the population size $NP \geq 4$ (see below). The basic DE algorithm can then be described as follows:

1. Initialize all agents x with random positions in the search-space.
2. Until a termination criterion is met (e.g. number of iterations performed, or adequate fitness reached), repeat the following:
   - 2.1 For each agent x in the population do:
     - 2.1.1 Pick three agents a, b and c from the population at random, they must be distinct from each other as well as from agent x.
     - 2.1.2 Pick a random index $R \in \{1, \ldots, n\}$ (n being the dimensionality of the problem to be optimized).
     - 2.1.3 Compute the agent's potentially new position $y = [y_1, \ldots, y_n]$ as follows:
       - 2.1.3.1 For each $i = [1, \ldots, n]$, pick a uniformly distributed number $r_i \equiv U(0.1)$
       - 2.1.3.2 If $r_i < CR$ or $i = R$ then set $y_i = a_i + F \times (b_i - c_i)$ otherwise set $y_i = x_i$.
       - 2.1.3.3 (In essence, the new position is the outcome of the binary crossover of agent x with the intermediate agent $z = a + F \times (b - c)$.
     - 2.1.4 If $f(y) < f(x)$ then replace the agent in the population with the improved candidate solution, that is, replace x with y in the population.

2.2 Pick the agent from the population that has the highest fitness or lowest cost and return it as the best-found candidate solution [43].

2.4 New developed method (DEF3 method)

A rock discontinuity is often assumed to have a planar structure and its spatial orientation is expressed in terms of dip direction $\alpha(0 \le \alpha \le 360°)$ and dip angle $\beta(0 \le \beta \le 90°)$. The orientation of a discontinuity can also be represented by a unit normal vector e_i (Fig. 1), which is often described by its direction cosines $e_i = (x_i\, y_i\, z_i)$ in a Cartesian coordinate system [44, 45]:

$$\begin{cases} x_i = \cos\alpha \sin\beta \\ y_i = \sin\alpha \sin\beta \\ z_i = \cos\beta \end{cases} . \tag{3}$$

It is in fact a mixture of the triple method introduced by Harrison [13] and Hammah and Curren [11], optimized by differential evolutionary method. Harrison's three-step method combines the fuzzy clustering method, fuzzy clustering by fuzzy covariance matrix of membership function, and fuzzy clustering based on the fuzzy maximum likelihood method by applying the methods suggested by Bezdek, Gustafson and Kessel, Gath and Geva. Using the capabilities of the above methods, it has gained a special ability in data clustering. It is also uniquely and clearly capable to identify clusters of different shapes like spherical and non-spherical, and to detect cluster boundaries. The only weakness of Harrison's method is the lack of good match to and intrinsic design for orientation data structure. Harrison's method is presented in Table 1.

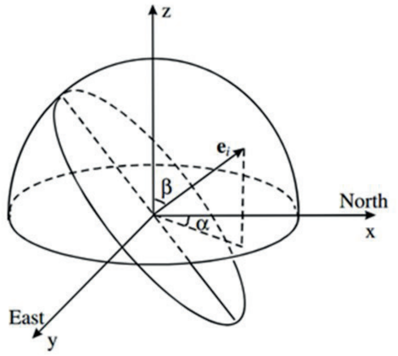


Fig. 1 directional data representation in spherical space

Table 1 The FCM, FCV and FMLE fuzzy clustering algorithms [12]

step	Fuzzy cmeans	Fuzzy covariance	Fuzzy maximum likelihood estimation
1	Adopt *C*, number of cluster and initialize *U*	Adopt final *U* from FCM as new initialization of U	Adopt final U from FCV as new initialization of *U*
2	Compute centroids, V_i $\begin{cases} V_i = \dfrac{\sum_{j=0}^{N} (u_{ij})^q X_j}{\sum_{j=0}^{N} (u_{ij})^q} \forall i \end{cases}$	Compute centroids, V_i, fuzzy SSP matrices, *SF*, and norm inducing matrices *A*: $\begin{cases} V_i = \dfrac{\sum_{j=0}^{N} (u_{ij})^q X_j}{\sum_{j=0}^{N} (u_{ij})^q} \forall i \end{cases}$ $SF_i = \sum_{j=0}^{N} (u_{ij})^q (X_j - V_i)(X_j - V_i)^T \,\forall i$ $A_i = (\rho_i \lvert SF_i)^{1/m} (SF_i)^{-1} \,\forall i$	Compute centroids, V_i, a priori probabilities, P_i, and fuzzy covariance matrices, *F*: $\begin{cases} V_i = \dfrac{\sum_{j=0}^{N} (u_{ij})^q X_j}{\sum_{j=0}^{N} (u_{ij})^q} \forall i \end{cases}$ $P_i = \dfrac{1}{N} \sum_{j=0}^{N} h(i \mid X_j) \forall i$ $F_i = \dfrac{\sum_{j=0}^{N} h(i \mid X_j)(X_j - V_i)(X_j - V_i)^T}{\sum_{j=0}^{N} h(i \mid X_j)} \forall i$
3	Compute membership values, *u*: $u_{ij} = \dfrac{\left[1/d^2(X_j.V_i) \right]^{1/(q-1)}}{\sum_{K=1}^{C} \left[1/d^2(X_j.V_i) \right]^{1/(q-1)}} \forall i.j$ Where $d^2(X_j.V_i) = (X_j - V_i)^T I (X_j - V_i)$	Compute membership values, *u*: $u_{ij} = \dfrac{\left[1/d^2(X_j.V_i) \right]^{1/(q-1)}}{\sum_{K=1}^{C} \left[1/d^2(X_j.V_i) \right]^{1/(q-1)}} \forall i.j$ Where $d^2(X_j.V_i) = (X_j - V_i)^T I (X_j - V_i)$	Compute membership values, *h* based on maximum likelihood estimation: $h(i \mid X_j) = \dfrac{\left[1/d^2(X_j.V_i) \right]^{1/(q-1)}}{\sum_{K=1}^{C} \left[1/d^2(X_j.V_i) \right]^{1/(q-1)}} \forall i.j$ Where $d^2(X_j.V_i) = \dfrac{\sqrt{\lvert F_i \rvert}}{P_i} exp\left[\dfrac{1}{2} (X_j - V_i)^T F_i^{-1} (X_j - V_i) \right]$
4	If $\max_{ij} \lVert u_{ij} - \check{u}_{ij} \rVert > \varepsilon$, $0 \le \varepsilon \le 1$, where $\check{u}_{ij}$, corresponding value from previous iteration, repeat from step 2.	If $\max_{ij} \lVert u_{ij} - \check{u}_{ij} \rVert > \varepsilon$, $0 \le \varepsilon \le 1$, where $\check{u}_{ij}$, corresponding value from previous iteration, repeat from step 2.	If $\max_{ij} \lVert u_{ij} - \check{u}_{ij} \rVert > \varepsilon$, $0 \le \varepsilon \le 1$, where $\check{u}_{ij}$, corresponding value from previous iteration, repeat from step 2.

Where in Table 1, N denotes all data Number, C, Cluster number, u_{ij} degree of membership of j-th data in the i-th cluster, V_i, Center of i-th cluster, $d^2(X_j,V_i)$, Cosine distance of j-th from i-th cluster center, T, transpose matrix, q, fuzziness of resulting partitions, SF_i, the fuzzy sums of squares and products(SSP) matrix of i-th cluster, A_i the norm-inducing matrix of i-th cluster, ρ_i, hyper volume of i-th cluster, P_i, the a priori probability of selecting the i-th cluster, F_i, the fuzzy covariance matrix of the i-th cluster. Hammah and Curran [11] in order to achieving best compatibility between vectoral feature of fracture orientation data and fuzzy clustering methods, proposed modified orientation matrix for i-th cluster, S_i^*, which is computed using formula (Eq. 4):

$$S_i^* = \frac{1}{\sum_{j=1}^{N}(ui_j)^m}\begin{bmatrix} \sum_{j=1}^{N}(u_{ij})^m x_jx_j & \sum_{j=1}^{N}(u_{ij})^m x_jy_j & \sum_{j=1}^{N}(u_{ij})^m x_jz_j \\ \sum_{j=1}^{N}(u_{ij})^m x_jy_j & \sum_{j=1}^{N}(u_{ij})^m y_jy_j & \sum_{j=1}^{N}(u_{ij})^m y_jz_j \\ \sum_{j=1}^{N}(u_{ij})^m x_jz_j & \sum_{j=1}^{N}(u_{ij})^m y_jz_j & \sum_{j=1}^{N}(u_{ij})^m z_jz_j \end{bmatrix}.(4)$$

It was proven and demonstrated in Hammah and Curran that this approach for computing prototypes correctly deals with clusters that contain antipodal vectors (sets which wrap between upper and lower hemispheres), because it always determines cluster centroids to lie within the acute angles between vectors. In both of these methods, with the definition of cost functions, it is attempted to reach the optimal value through steps; this function is given in relations [11, 12]:

$$J_m\left(U_i.V_i\right) = \sum_{j=1}^{N}\sum_{i=1}^{K}\left(u_{ij}\right)^m d^2\left(X_j.V_i\right);\ K \le N\ . \tag{5}$$

The results of the studies show that, while having high strengths in calculation, the two methods needs for a more powerful and effective tool to achieve optimal clustering due to the inefficiency of the optimizer section, which operates by maximization of the membership function difference in successive stages (Eq. 5). Therefore, in order to solve the problem, the current paper uses differential evolutionary optimization algorithm, based on the special capabilities of this algorithm in searching for functions global extremum points. The proposed method, unlike the Harrison method, instead of random generation of membership function values for any data in each cluster which produce a significant number of clusters and thus increased volume of computation, reduces the volume of computing and increases its speed through a new and effective initiation and attempts to generate random cluster centers and thus calculates the value of the membership function. In order to further adapt the new method to the vector space of the data, the orientation matrix proposed by Hammah and Curran is used to obtain cluster centers during the calculation steps.

It is need to be mentioned that because of more compatibility of cosine distance measure with spherical data, in new proposed method Euclidian distance measure replaced by cosine distance measure which used to measure distance between a pattern vector and the prototype or centroid. According to the above-mentioned reasons, proposed new method because of incorporating aforementioned strong clustering methods abilities, is able to yield reliable and acceptable results in clustering directional data such as dip and dip direction data of fractures. Fig. 2 depicted flowchart of various steps of new proposed method.

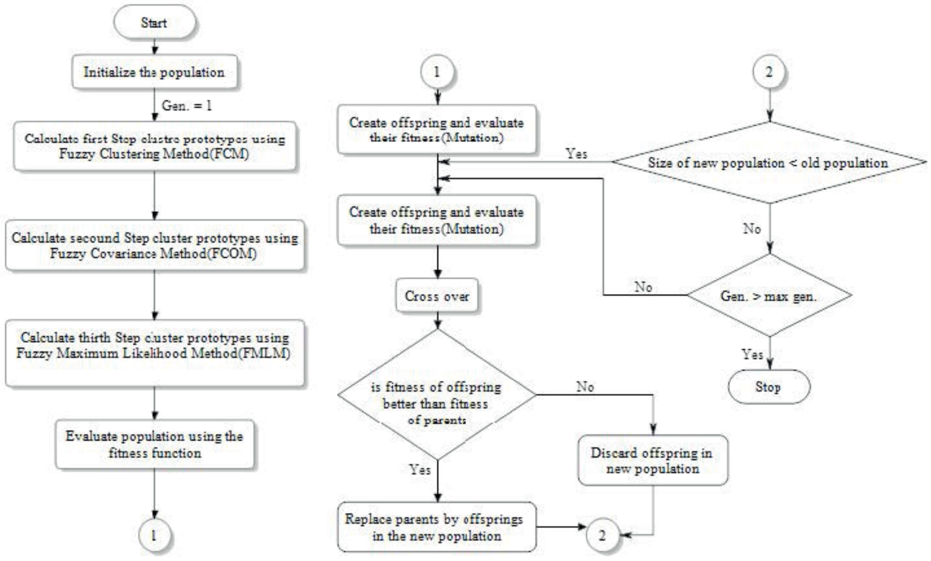

Start
Initialize the population
Gen. = 1
Calculate first Step clustre prototypes using Fuzzy Clustering Method(FCM)
Calculate secound Step cluster prototypes using Fuzzy Covariance Method(FCOM)
Calculate thirth Step cluster prototypes using Fuzzy Maximum Likelihood Method(FMLM)
Evaluate population using the fitness function
1
1
Create offspring and evaluate their fitness(Mutation)
Create offspring and evaluate their fitness(Mutation)
Cross over
is fitness of offspring better than fitness of parents
Yes
No
Replace parents by offsprings in the new population
Discard offspring in new population
2
2
Size of new population < old population
Yes
No
Gen. > max gen.
No
Yes
Stop

Fig. 2 flowchart of DEF3 Method

After proposing any new method, it is needed to evaluate performance of the method. The evaluation process requires the using of criterions which could be introduced effectiveness and accuracy and precision of new proposed method. As first time in this paper, performance of proposed new method is evaluated by applying three key criterions which is called Silihotte [46], Davis-Boulden [47] and Calinski-Harabasz [48]. In following section these three criterions will be presented. The Calinski-Harabasz criterion is sometimes called the variance ratio criterion (VRC). The Calinski-Harabasz index is defined as:

$$VRC_k = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)} . \tag{6}$$

Where SS_B is the overall between-cluster variance, SS_W is the overall within-cluster variance, k is the number of clusters, and N is the number of observations. The overall between-cluster variance SS_B is defined as:

$$SS_B = \sum_{i=1}^{k} n_i m_i - m^2 . \tag{7}$$

Where k is the number of clusters, m_i is the centroid of cluster i, m is the overall mean of the sample data, and $\|m_i - m\|$ is the L^2 norm (Euclidean distance) between the two vectors. The overall within-cluster variance SS_W is defined as:

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} x - m_i^2 . \tag{8}$$

Where k is the number of clusters, x is a data point, c_i is the ith cluster, m_i is the centroid of cluster i, and $\|x - m_i\|$ is the L^2 norm (Euclidean distance) between the two vectors. Well-defined clusters have a large between-cluster variance (SS_B) and a small within-cluster variance (SS_W). The larger the VRC_K ratio, the better the data partition. To determine the optimal number of clusters, maximize VRC_K with respect to k. The optimal number of clusters is the solution with the highest Calinski-Harabasz index value. The Calinski-Harabasz criterion is best suited for k-means clustering solutions with squared Euclidean distances [48].

The Davies-Bouldin criterion is based on a ratio of within-cluster and between cluster distances. The Davies-Bouldin index is defined as:

$$DB = \frac{1}{k} \overset{k}{\underset{i=1}{}} max_{j \neq i} \left\{ D_{i,j} \right\} . \tag{9}$$

Where D_{ij} is the within-to-between cluster distance ratio for the ith and j-th clusters. In mathematical terms:

$$D_{i.j} = \frac{\left(\bar{d}_i + \bar{d}_j\right)}{d_{i.j}} . \tag{10}$$

Where $\bar{d}_i$ is the average distance between each point in the i-th cluster and the centroid of the i-th cluster and $\bar{d}_i$ is the average distance between each point in the j-th cluster and the centroid of the j-th cluster and d_{ij} is the Euclidean distance between the centroids of the ith and jth cluster. The maximum value of d_{ij} represents the worst-case within-to-between cluster ratio for cluster i. The optimal clustering solution has the smallest Davies-Bouldin index value [47].

The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the i-th point, S_i, is defined as:

$$S_i = (b_i - a_i)/\max(a_i \cdot b_i) . \tag{11}$$

Where a_i is the average distance from the ith point to the other points in the same cluster as i, and b_i is the minimum average distance from the i-th point to points in a different cluster, minimized over clusters. The silhouette value ranges from –1 to +1. A high silhouette value indicates that i is well-matched to its own cluster, and poorly-matched to neighboring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. If many points have a low or negative silhouette value, then the clustering solution may have either too many or too few clusters. The silhouette clustering evaluation criterion can be used with any distance metric [46].

The cost function for the differential evolutionary optimization algorithm is the sum of the variation percentage of the three above parameters multiplied by the total percentage of positive variations (Eq.12).

$$\text{CQI} = (\Delta\text{CB \%} + \Delta\text{DB \%} + \Delta\text{Sili \%}) * \text{SOPI} . \tag{12}$$

Where CQI represents the clustering quality index; ΔCB % refers to Calinski-Harabasz's standard percentage change, considering the sign; ΔDB % is Davies-Bouldin standard percentage change, considering the sign; ΔSili % is Silhouette standard percentage change, considering the sign; and SOPI represents the total percentages of improved criteria.

3 Case study

The data obtained from the cavern of pumped-storage power plant project of Rudbar, Lorestan dam and was used in validating DEF3 method (Fig. 3).

Fig. 3 Mapping of fractures using scanline method in Rudbar Lorestan dam site cavern

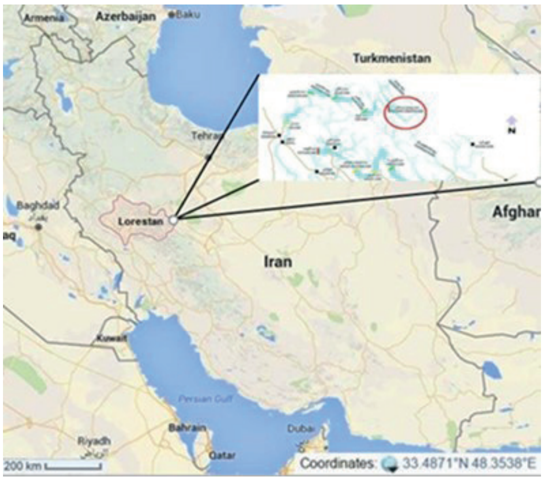


Fig. 4 Location of case study

Rudbar, Lorestan pumped- storage plan located in Zagros high zone has very complicated tectonic conditions and numerous over thrust faults have formed high areas. The active faults of Saravan Baznavid and Chaleh Hatam are the most important faults in the site range. The carven of the power plant is located almost in the center of the anticline. This project is under construction at the Rudbar, Lorestan Dam upstream, located 150 km west of Isfahan and 100 km south of Aligudarz city. The area of the project is located under the high sedimentary-structural zone (internal Zagros) (Fig. 4).

Table 2 Hopkins index value of pumped-storage power plant project of Rudbar, Lorestan dam data

Data Number	Samples Number	Hopkins index
627	63	0.81

The cavern has a length of 130 meters, a width of 26 meters and a height of 50 meters, located in the Dalan Formation (end of the first geology period), with calcareous rock masses and dolomitic limestone of average thickness.

The data is obtained from the cavern by scanline and scan window surveying method, and its number reaches 627 fractures. In the first step, the data is corrected relative to the direction of the survey line before the clustering process begins. Then, clustering capability of the data is evaluated using the Hopkins index. Table 2 shows the Hopkins index calculated for the data.

As the Hopkins statistics is 0.81, it is clear that the data obtained from the cavern site of the power plant has a high potential for clustering. It should be noted that the number of samples taken from the data to calculate the Hopkins statistics, as explained in the previous section, is equal to 10 % of the total number of data (63 samples). After ensuring the clustering capability of data, the number of optimal clusters should be determined to start the clustering. This is done using the elbow index (Fig. 5).

For data extracted from the cavern of Rudbar power plant, according to the Fig. 6 and the description given in the previous section, the best number of adaptable clusters is achieved to be 3. After assuring the data clustering capability and determining the number of optimal clusters, the data, as in the first phase of validation, are clustered based

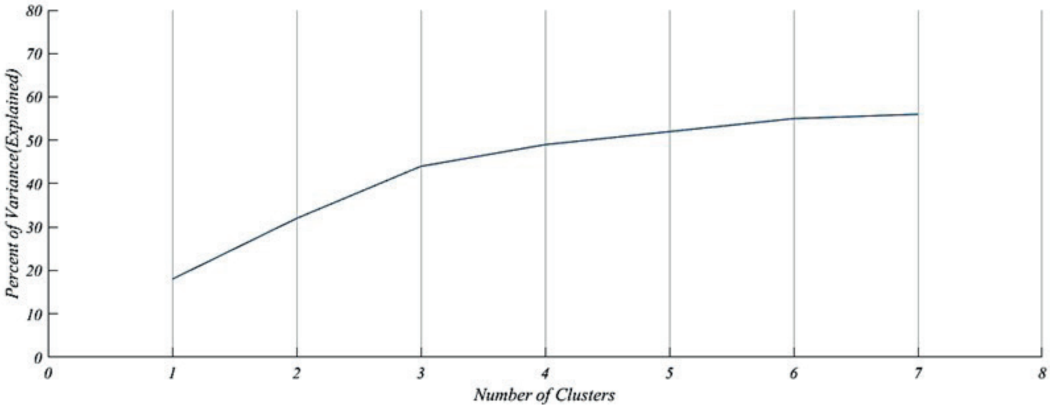


Fig. 5 Illustration of the percentage of data variance(explained) is plotted in terms of the clusters number

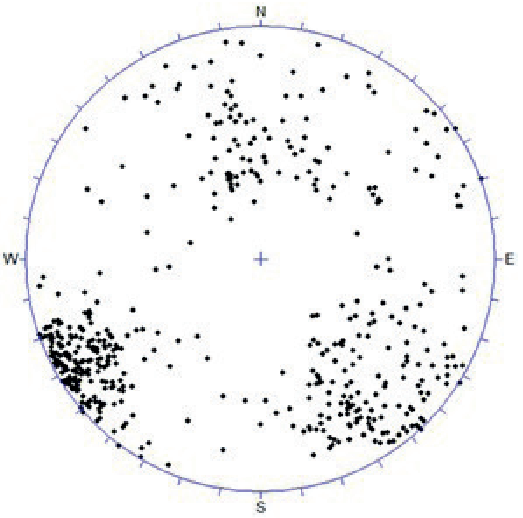
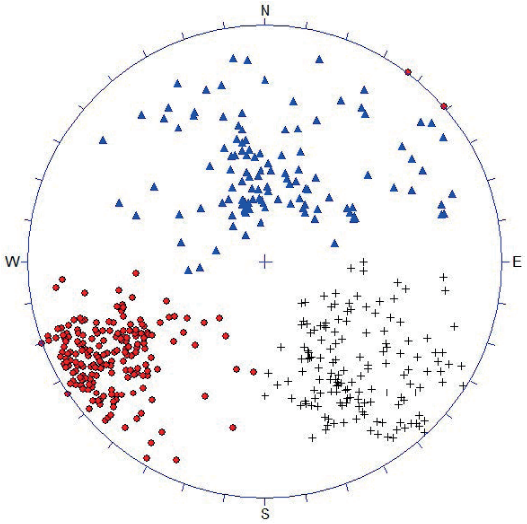
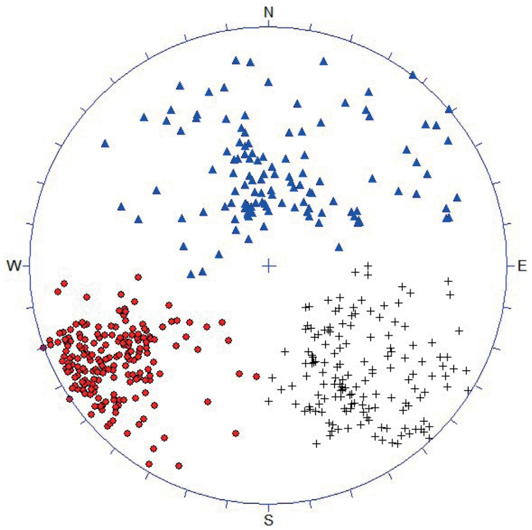


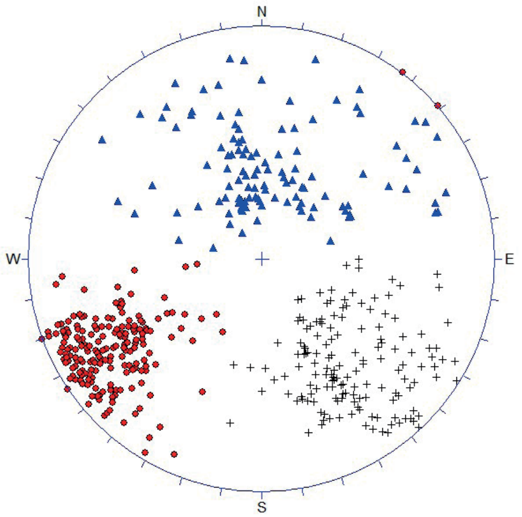
Fig. 6 Scatter diagram of sample (equal area upper hemispherical projection) of Rudbar, Lorestan dam cavern



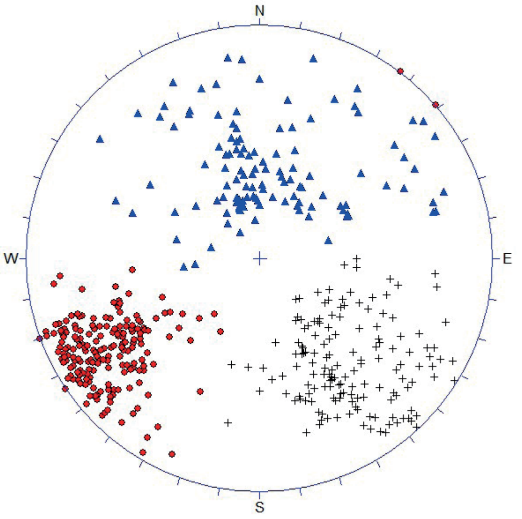
(a)



(b)



(c)



(d)

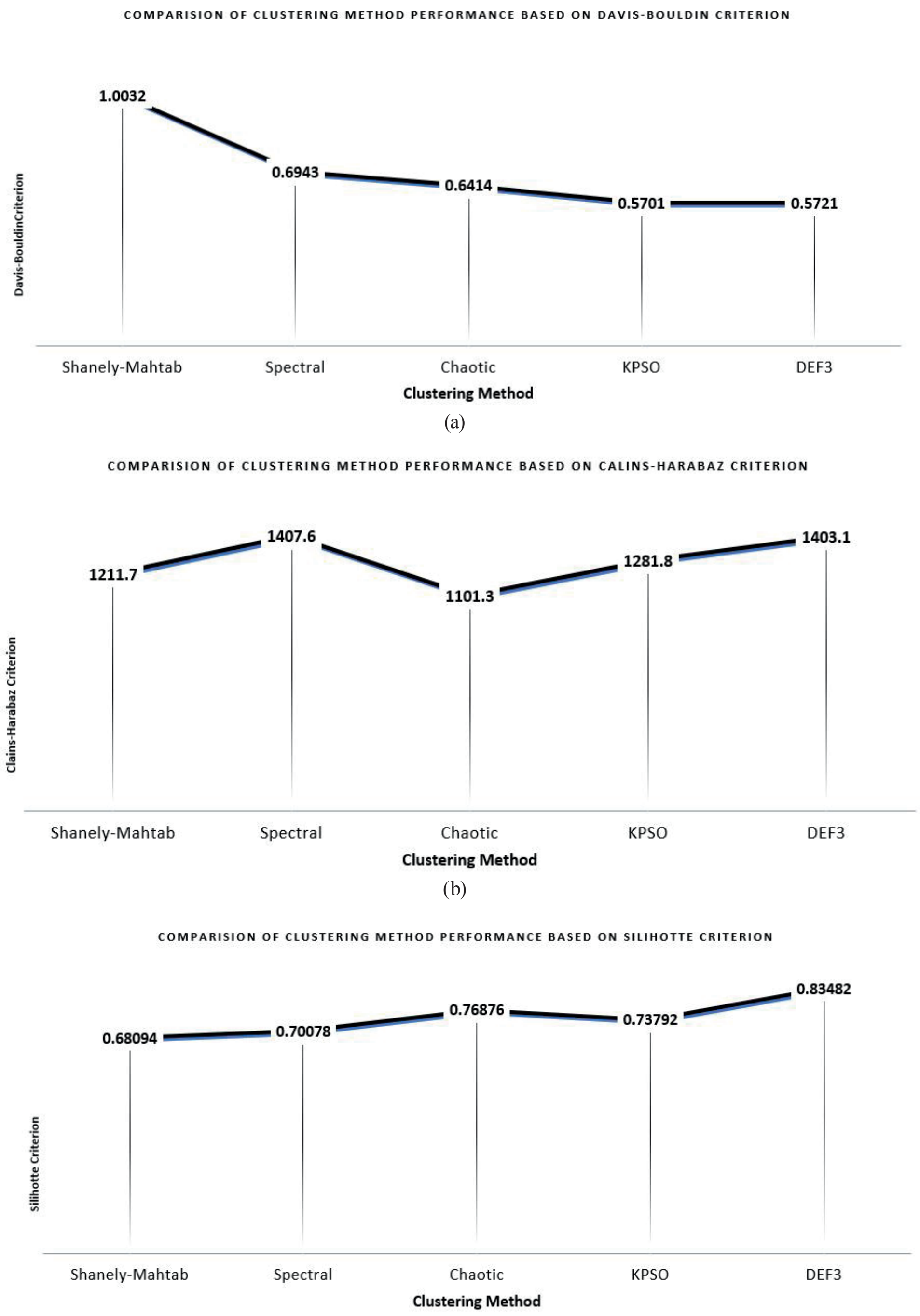
Fig. 7 Comparison of the clustering results of the New proposed method and those of the new method. A. Shanely-Mahtab method. B. Spectral method C. Chaotic Method. D. KPSO. E. DEF3

on five k-means methods: the method based on the particle swarm algorithm [22], the fuzzy method by chaos-based optimization algorithm [26], the spectral fuzzy method [7], the method proposed by Shanely and Mahtab [5], the new method proposed in this paper based on differential evolutionary algorithm. Additionally, Calinski-Harabasz, Davies-Bouldin and Silhouette criteria are calculated for every value. Fig. 7 shows stereographic plot of the clustering result of all methods.

Examining the quality of the clustering process is possible in a variety of ways due to the specific nature of this process. Under certain conditions, using the geometric structure and location of the data, it is possible to find the clustering quality and the correctness level of allocating the data to the clusters by observing the spatial layout of the data. But in most cases, due to the multiplicity of data and their complex spatial distribution, visual quality measurement methods are not sufficient. Under these conditions, indicators and criteria are presented for studying the quality of clustering. In this paper, the values of the above three criteria were calculated for the data in order to evaluate the quality and the accuracy of each method, as shown in Table 3 and Fig. 8.

Table 3 values of Calinski-Harabasz, Davies-Bouldin and Silhouette criteria of data

Method	Silihotte	Calins-Harabaz	Davis-Bouldin
Shanely-Mahtab	0.68094	1211.7	1.0032
Spectral	0.70078	1407.6	0.6943
Chaotic	0.76876	1101.3	0.6414
KPSO	0.73792	1281.8	0.5701
DEF3	0.83482	1403.1	0.5721



(a)

(b)

(c)

Fig. 8 Illustration of values of A. Calinski-Harabasz, B. Davies-Bouldin C. Silhouette criteria

4 Discussion

As in the first part of this section, according to the table 3 and the Fig. 8, it is clear that DEF3 performs better in data clustering than other methods. Davies-Bouldin criterion is the first for assessing quality of clustering processes. As explained, this criterion gives lower values for good quality clustering. In other words, the value of this criterion will be higher for low quality clustering processes. Referring to Table 3 and Fig. 8, we can see that DEF3 and KPSO method, with values of 0.5721 and 0.5701, perform better than other methods. However, the approach based on KPSO method has improved slightly compared to DEF3. In this criterion, the chaotic theory-based method with a value of 0.6414 has the third good performance. The Calinski-Harabasz criterion is considered as the second evaluator of the clustering performance. This criterion has a computational structure so that as clustering quality increases, it provides larger values. Given this description and according to Table 3 and Fig. 8, it is clear that for this criterion, DEF3 method and Jimenez spectral method with the values of 1403.1 and 1407.6, respectively, provides the highest values and thus the best performance. However, Jimenez spectral method slightly perform better than DEF3. In this criterion, after the above-mentioned methods, the KPSO method has the best performance with the value of 1281.8. Silhouette criterion is the third criterion used to evaluate the various clustering methods of data extracted from the power plant cavern. Considering the nature of the Silhouette criterion and referring to Table 3 and Fig. 8, it is clear that DEF3 with the value of 0.83482 for silhouette criterion has the best performance compared to other methods, while in the two previous criteria, the new developed method with a negligible difference showed the second-best performance. Chaotic theory-based method, with the value of 0.76876 for silhouette index, has the second-best performance after the DEF3 method. In clustering these data, stability of performance quality in the face of a variety of criteria for assessing clustering quality is another important point related to the new method. As shown by Table 3 and Fig. 8, in all three criteria, the new proposed method, has a stable performance in addition to the best performance. In other words, other methods perform differently and unstably in the face of different criteria.

5 Conclusions

Due to the importance of fractures in rock mass behavior which usually be observed as the main materials in creation of infrastructure projects, such as dams, tunnels, caverns and rock slopes, in this paper study of one of the most important aspect of fractures properties which is called joint sets identification, was carried out in a detailed way. Since fractures in most cases could be clustered based on their orientation property, hence they were classified considering their dip and dip direction. By inspecting various methods for fracture clustering proposed by different researchers, using the strengths of these methods and redefining the fractures clustering process in the form of an optimization problem, as well as providing a new and efficient cost function and its integration with the strong optimization technique of differential evolutionary algorithm a new method was proposed based on modified three-step fuzzy clustering methods. A series of field data were used to evaluate and validate the proposed new method. Four methods proposed by other researchers were selected and their performance was compared with the new method under the same conditions. The results obtained from the comparison, by considering the three criteria to evaluate clustering processes, indicate an acceptable and better performance of the method proposed in this paper. Moreover, it has a good stability in comparison with other methods. In this paper, the data was first corrected relative to the orientation of the scanline survey and then subjected to the clustering process. It should also be mentioned that before data clustering process begins, clustering capability was evaluated using the Hopkins index, and then the data entered the clustering phase. In order to overcome the problem of selecting the number of optimum clusters by using the elbow technique, the number of optimum clusters was determined and given as input into the clustering processes. Due to the capabilities of the new proposed method, it can be used to classify fractures with high confidence and the results obtained can be utilized in analyzing the behavior of structures and rock masses.

References

[1] Jimenez-Rodriguez, R., Sitar, N. "A spectral method for clustering of rock discontinuity sets", International Journal of Rock Mechanics and Mining Sciences, 43(7), pp. 1052–1061, 2006. https://doi.org/10.1016/j.ijrmms.2006.02.003

[2] Maffucci, R., Bigi, S., Corrado, S., Chiodi, A., Di Paolo, L., et al. "Quality assessment of reservoirs by means of outcrop data and "discrete fracture network" models: The case history of Rosario de La Frontera (NW Argentina) geothermal system", Tectonophysics, 647–648, pp. 112–131, 2015. https://doi.org/10.1016/j.tecto.2015.02.016

[3] Park, H.-J., West, T. R., Woo, I. "Probabilistic analysis of rock slope stability and random properties of discontinuity parameters, Interstate Highway 40, Western North Carolina, USA", Engineering Geology, 79(3–4), pp. 230–250, 2005. https://doi.org/10.1016/j.enggeo.2005.02.001

[4] Priest, S. D. "Discontinuity analysis for rock engineering", Springer, Dordrecht, Netherlands, 2012.

[5] Shanley, R. J., Mahtab, M. A. "Delineation and analysis of clusters in orientation data", Journal of the International Association for Mathematical Geology, 8(1), pp. 9–23, 1976. https://doi.org/10.1007/BF01039681

[6] Kulatilake, P. H. S. W., Fiedler, R., Panda, B. B. "Box fractal dimension as a measure of statistical homogeneity of jointed rock masses", Engineering Geology, 48(3–4), pp. 217–229, 1997. https://doi.org/10.1016/S0013-7952(97)00045-8

[7] Jimenez, R. "Fuzzy spectral clustering for identification of rock discontinuity sets", Rock Mechanics and Rock Engineering, 41(6), pp. 929–939, 2008. https://doi.org/10.1007/s00603-007-0155-6

[8] Liu, J., Zhao, X.-D., Xu, Z. "Identification of rock discontinuity sets based on a modified affinity propagation algorithm", International Journal of Rock Mechanics and Mining Sciences, 94, pp. 32–42, 2017. https://doi.org/10.1016/j.ijrmms.2017.02.012

[9] Mahtab, M. A., Yegulalp, T. M. "A rejection criterion for definition of clusters in orientation data", presented at The 23rd U.S Symposium on Rock Mechanics (USRMS), Berkeley, California, USA, Aug. 25–27, 1982.

[10] Marcotte, D., Henry, E. "Automatic joint set clustering using a mixture of bivariate normal distributions", International Journal of Rock Mechanics and Mining Sciences, 39(3), pp. 323–334, 2002. https://doi.org/10.1016/S1365-1609(02)00033-3

[11] Hammah, R. E., Curran, J. H. "Fuzzy cluster algorithm for the automatic identification of joint sets", International Journal of Rock Mechanics and Mining Sciences, 35(7), pp. 889–905, 1998. https://doi.org/10.1016/S0148-9062(98)00011-4

[12] Hammah, R. E., Curran, J. H. "On Distance Measures for the Fuzzy K-means Algorithm for Joint Data", Rock Mechanics and Rock Engineering, 32(1), pp. 1–27, 1999. https://doi.org/10.1007/s006030050041

[13] Harrison, J. P. "Fuzzy objective functions applied to the analysis of discontinuity orientation data", presented at Rock Characterization: ISRM Symposium, Eurock '92, Chester, United Kingdom, Sept. 14–17, 1992.

[14] Klose, C. D., Seo, S., Obermayer, K. "A new clustering approach for partitioning directional data", International Journal of Rock Mechanics and Mining Sciences, 42(2), pp. 315–321, 2005. https://doi.org/10.1016/j.ijrmms.2004.08.011

[15] Zhou, W., Maerz, N. H. "Implementation of multivariate clustering methods for characterizing discontinuities data from scanlines and oriented boreholes", Computers and Geosciences, 28(7), pp. 827-839, 2002. https://doi.org/10.1016/S0098-3004(01)00111-X

[16] Dershowitz, W., Busse, R., Geier, J., Uchida, M. "A stochastic approach for fracture set definition", presented at 2nd North American Rock Mechanics Symposium, Montreal, Quebec, Canada, June 19–21, 1996.

[17] Hammah, R. E., Curran, J. H. "Standardization and weighting of variables for the fuzzy K-means clustering of discontinuity data", presented at 4th North American Rock Mechanics Symposium, Seattle, Washington, USA, July 31–August 3, 2000.

[18] Hammah, R. E., Curran, J. H. "Validity measures for the fuzzy cluster analysis of orientations", IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), pp. 1467–1472, 2000. https://doi.org/10.1109/34.895981

[19] Bezdek, J. C., Pal, S. K. "Fuzzy models for pattern recognition: Methods That Search for Structures in Data", 1st ed., IEEE Press, New York, NY, United States, 1992.

[20] Gustafson, D. E., Kessel, W. C. "Fuzzy clustering with a fuzzy covariance matrix", In: IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes, San Diego, California, United States, 1979, pp. 761–766. https://doi.org/10.1109/CDC.1978.268028

[21] Gath, I., Geva, A. B. "Unsupervised optimal fuzzy clustering", IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(7), pp. 773–780, 1989. https://doi.org/10.1109/34.192473

[22] Li, Y., Wang, Q., Chen, J., Xu, L., Song, S. "K-means Algorithm Based on Particle Swarm Optimization for the Identification of Rock Discontinuity Sets", Rock Mechanics and Rock Engineering, 48(1), pp. 375–385, 2015. https://doi.org/10.1007/s00603-014-0569-x

[23] Sirat, M., Talbot, C. J. "Application of artificial neural networks to fracture analysis at the Äspö HRL, Sweden: fracture sets classification", International Journal of Rock Mechanics and Mining Sciences, 38(5), pp. 621–639, 2001. https://doi.org/10.1016/S1365-1609(01)00030-2

[24] Tokhmechi, B., Memarian, H., Ahmadi, N. H., Moshiri, B. "A new method for Joint set classification based on Bayesian optimum classifier", Geosciences, 18(71), pp. 115–122, 2009.

[25] Tokhmechi, B., Memarian, H., Moshiri, B., Rasouli, V., Noubari, H. A. "Investigating the validity of conventional joint set clustering methods", Engineering Geology, 118(3–4), pp. 75–81, 2011. https://doi.org/10.1016/j.enggeo.2011.01.002

[26] Xu, L. M., Chen, J. P., Wang, Q., Zhou, F. J. "Fuzzy C-Means Cluster Analysis Based on Mutative Scale Chaos Optimization Algorithm for the Grouping of Discontinuity Sets", Rock Mechanics and Rock Engineering, 46(1), pp. 189–198, 2013. https://doi.org/10.1007/s00603-012-0244-z

[27] Cai, M., Wang, P., Zhao, K., Zhang, D. "Fuzzy C-means cluster analysis based on genetic algorithm for automatic identification of joint sets", Chinese Journal of Rock Mechanics, 24(3), pp. 371–376, 2005. (in Chinese) https://doi.org/10.3321/j.issn:1000-6915.2005.03.002

[28] Ma, G. W., Xu, Z. H., Zhang, W., Li, S. C. "An enriched K-means clustering method for grouping fractures with meliorated initial centers", Arabian Journal of Geosciences, 8(4), pp. 1881–1893, 2015. https://doi.org/10.1007/s12517-014-1379-x

[29] Bezdek, J. C. "Objective function clustering", In: Bezdek, J. C. (ed.) Pattern Recognition with Fuzzy Objective Function Algorithms, Springer, Boston, Massachusetts, Unites States, 1981. pp. 43–93. https://doi.org/10.1007/978-1-4757-0450-1_3

[30] Bezdek, J. C., Coray, C., Gunderson, R., Watson, J. "Detection and Characterization of Cluster Substructure I. Linear structure: Fuzzy c-lines", SIAM Journal on Applied Mathematics, 40(2), pp. 339–357, 1981. https://doi.org/10.1137/0140030

[31] Babuka, R., van der Veen, P. J., Kaymak, U. "Improved covariance estimation for Gustafson-Kessel clustering", In: IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings, Honolulu, Hawaii, United States, 2002, pp. 1081–1085. https://doi.org/10.1109/FUZZ.2002.1006654

[32] Krishnapuram, R., Kim, J. "A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithms", IEEE Transactions on Fuzzy systems, 7(4), pp. 453–461, 1999. https://doi.org/10.1109/91.784208

[33] Rammal, A., Perrin, E., Vrabie, V., Bertrand, I., Chabbert, B. "Weighted-covariance factor fuzzy c-means clustering", In: Third International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), Beirut, Lebanon, 2015, pp. 144–149. https://doi.org/10.1109/TAEECE.2015.7113616

[34] Hopkins, B., Skellam, J. G. "A New Method for Determining the Type of Distribution of Plant Individuals", Annals of Botany, 18(2), pp. 213–227, 1954. https://doi.org/10.1093/oxfordjournals.aob.a083391

[35] Banerjee, A., Dave, R. N. "Validating clusters using the Hopkins statistic", In: IEEE International Conference on Fuzzy Systems, Budapest, Hungary, 2004, pp. 149–153. https://doi.org/10.1109/FUZZY.2004.1375706

[36] Ketchen, D. J. , Shook, C. L. "The application of cluster analysis in strategic management research: an analysis and critique", Strategic Management Journal, 17(6), pp. 441–458, 1996. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

[37] Adolfsson, A., Ackerman, M., Brownstein, N. C. "To Cluster, or Not to Cluster: How to Answer the estion", In: TKDD'17, Halifax, Nova Scotia, Canada, 2016, pp. 1–9.

[38] Goutte, C., Toft, P., Rostrup, E., Nielsen, F. Å., Hansen, L. K. "On Clustering fMRI Time Series", NeuroImage, 9(3), pp. 298–310, 1999. https://doi.org/10.1006/nimg.1998.0391

[39] Huh, M.-H. "Setting the Number of Clusters in K-Means Clustering", In: Baba Y., Hayter A.J., Kanefuji K., Kuriki S. (eds.) Recent Advances in Statistical Research and Data Analysis, Springer, Tokyo, Japan, 2002, pp. 115–124. https://doi.org/10.1007/978-4-431-68544-9_5

[40] Storn, R., Price, K. "Differential Evolution - A Simple and Efficient Heuristic for Global Optimization Over Continuous Spaces", Journal of Global Optimization, 11(4), pp. 341–359, 1997. https://doi.org/10.1023/A:1008202821328

[41] Stom, R., Price, K. "Differential Evolution - A Simple and Efficient Adaptive Scheme for Global Optimization Over Continuous Space", International Computer Science Institute, Berkeley, California, Unites States, Rep. TR-95-012, 1995.

[42] Ali, M., Pant, M., Abraham, A. "Simplex Differential Evolution", Acta Polytechnica Hungarica, 6(5), pp. 95–115, 2009. [online] Available at: https://uni-obuda.hu/journal/Ali_Pant_Abraham_21.pdf [Accessed: 15.04.2019]

[43] Fleetwood, K. "An introduction to differential evolution", presented at Proceedings of Mathematics and Statisctics of Complex Systems (MASCOS) One Day Symposium, Brisbane, Australia, Nov. 26, 2004.

[44] Pecher, A. "SCHMIDTMAC - A program to display and analyze directional data", Computers and Geosciences, 15(8), pp. 1315-1326, 1989. https://doi.org/10.1016/0098-3004(89)90095-2

[45] Wong, L. N. Y., Liu, G. "An improved K-means clustering method for the automatic grouping of discontinuity sets", presented at Rock Mechanics Symposium and 5th U.S. - Canada Rock Mechanics Symposium, Salt Lake City, Utah, United States, June 27–30, 2010.

[46] Rousseeuw, P. J. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", Journal of Computational and Applied Mathematics, 20, pp. 53–65, 1987. https://doi.org/10.1016/0377-0427(87)90125-7

[47] Davies, D. L., Bouldin, D. W. "A cluster separation measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), pp. 224-227, 1979. https://doi.org/10.1109/TPAMI.1979.4766909

[48] Caliński, T., Harabasz, J. "A dendrite Method for Cluster Analysis" Communication in Statistics - Theory and Methods, 3(1), pp. 1–27, 1974. https://doi.org/10.1080/03610917408548446