Shahid Beheshti University

Sustainable Earth Trends

Journal homepage: http://sustainearth.sbu.ac.ir

Spatio-temporal optimization of long-term groundwater monitoring networks using PSO algorithm

Mohammad Mirzavand[a], Seyyed Javad Sadatinejad[a*], Hamid Kardan Moghaddam[b]

[a] School of Energy Engineering and Sustainable Resources, College of Interdisciplinary Science and Technology, University of Tehran, Tehran, Iran
[b] Water Research Institute, Ministry of Energy, Tehran, Iran

ABSTRACT

Spatial and temporal variations of contamination in groundwater resources, necessitate long-term monitoring (LTM) at a given site. In this study, several groundwater quality parameters (EC, SAR, TH, TDS, pH, K, Na^{+}, Ca^{2+}, Mg^{2+}, $SO4^{2-}$, HCO^{3-}, and Cl^{-}) for 113 samples sites clustered based on the particle swarm optimization **(**PSO) algorithm to significantly decrease cost and save time in LTM. The optimization of the clustering process was carried out according to the Silhouette index. For verification and validation of the results, Geology, soil order, land use, hydrological network and, TDS maps were used. According to the results, the best number of clusters was 5. An acceptable agreement was obtained between land conditions and clusters represented by the PSO algorithm. Consequently, it can be inferred that the clustering of the groundwater quality using the PSO algorithm and the Silhouette index optimizer could 70% decrease the number of spatio-temporal sampling in LTM.

ARTICLE INFO

Keywords:
Clustering
Groundwater quality
Long-term monitoring
PSO algorithm
Silhouette index

Article history:
Received: 10 Aug 2024
Accepted: 7 Oct 2024

*Corresponding author
E-mail address:
jsadatinejad@ut.ac.ir
(S.J. Sadatinejad)

Citation:
Mirzavand, M et al., (2025). Spatio-temporal optimization of long-term groundwater monitoring networks using PSO algorithm, *Sustainable Earth Trends:* 5(1), (1-12).

DOI: 10.48308/set.2024.237141.1072

1. Introduction

Groundwater plays a significant role in the ecosystem, especially in arid and semi-arid areas (Mirzavand and Ghazavi, 2015). For water supply management, quantity and quality information are required (Dastorani et al., 2020), especially where groundwater is the major source of freshwater. But the quality is more important than quantity, because it determines the type of water usage. For effective management of groundwater quality and long-term monitoring (LTM) of contaminated groundwater sites, it is vital to have number of sampling locations and monitoring wells at a given site (Li and Chan Hilton, 2005). This LTM is necessary for human health and environmental risk. However, the LTM can be expensive and time-consuming due to the vast number of sampling points and the number of elements examined at a given site. (Li and Chan Hilton, 2005). Clustering of groundwater quality sampling sites should significantly decrease cost and save the time Groundwater site sampling clustering is a kind of optimization that it is done by many researchers (Hossain et al., 2013).

The first step in any optimization strategy is to divide the items into a predetermined number of groups. Then, items are assigned to clusters based on the objective function until a certain ending requirement is satisfied. There are differences between these methods in terms of the objective functions, reassignment procedures, terminating criteria, and beginning partitions. as opposed to hierarchical clustering

methods (Abu-khalaf et al., 2013; Hossain et al., 2013). Similarity matrices are not stored by optimization techniques. Thus, storage capacity has no bearing on the size of the data. Nevertheless, optimization techniques have a number of drawbacks: The efficiency of optimization approaches is largely dependent on the initial partition. (i) Some algorithms require the number of clusters a priori. (ii) Certain clustering criteria are biased towards particular cluster forms, and will impose these shapes on the data. Researchers employed a variety of novel methods for optimization, such as the genetic algorithm (GA) (Mirzavand and Walter, 2024) and simulated annealing (SA) (Jha and Datta, 2013; Saruhan, 2014). Recently, particle swarm optimization (PSO) (Kennedy and Eberhart, 1995) and ant- colony optimization (ACO) (Dorigo et al., 1996), shuffled complex evolution (SCE) (Duan et al., 1992), simplex simulated annealing (SIMPSA) (Cardoso et al., 1996), differential evolution (DE) (Storn and Price, 1997), artificial bee colony optimization (ABC) (Karaboga, 2005), harmony search (HS) (Geem et al., 2001), bacterial foraging optimization (Passino, 2002), invasive weed optimization (Mehrabian and Lucas, 2006), cuckoo search (Yang and Deb, 2010) have been successfully applied to a wide range of engineering and science problems (Mirzavand and Walter, 2024). PSO is a technique for computational intelligence that has previously been used for clustering (Mirzavand and Walter, 2024). PSO has been applied by many researchers. The proposed PSO-based method not only simplifies the computational process but also enhances the accuracy and cost-effectiveness of groundwater monitoring, providing a valuable advancement over previously established technique (Mirzavand and Walter, 2024). PSO computation is relatively straightforward. Unlike other optimization algorithms, it requires fewer computational resources and is easier to complete (Eskandari et al., 2022). PSO algorithm carried out by many researchers for optimization goals. For groundwater management and groundwater remediation optimization (Mirzavand and Walter, 2024). To the best of the author's knowledge, PSO hasn't been utilized to group groundwater quality test locations, though. This study's primary goal was to present a novel approach that uses the particle swarm optimization algorithm to cut down on the number of spatiotemporal samplings required for long-term groundwater quality monitoring. This work proposes a PSO-based groundwater quality sample location clustering technique.

2. Material and Methods

2.1. Study area

The study area (longitude: 48° 10' to 48° 37' E, latitude: 32° 03' to 32° 33' N) is located in Dezful-Andimeshk plain aquifer, Northwest of Khuzestan province, Iran (Fig. 1). The Dezful-Andimeshk plain aquifer has an area of 1778 km^2. The study area's geological formation is composed of the Aghajari formation (brown to grey, calcareous, feature-forming sandstone and low weathering, gypsum-veined, red marl and siltstone), the Bakhtiari formation (alternating hard of consolidated, massive, feature-forming conglomerate and low-weathering cross-bedded sandstone), and low-level pediment fan and valley terrace deposits (Fig. 2a). The study area's primary land use is farming, with the main crops grown there being wheat in the winter and spring and corn and vegetables in the summer and fall (Fig. 2b). Groundwater and soil TDS maps of the study area were shown in Fig. 2d and 2c. Summary of the chemical composition (EC, SAR, TH, TDS, pH, K, Na^+ ،Ca^{2+} ‘Mg^{2+} ،$SO4^{2-}$ ،HCO^{3-}, and Cl^-) of the groundwater samples (n = 113) from Dezful-Andimeshk plain aquifer were shown in Table 1. The groundwater depth in the north part of the study area is about 88 meters and in the southern part is about 3 meters. In most parts of the aquifer (80%), the water table is approximately 10 meters. The distribution of groundwater samples was shown in Fig. 1.

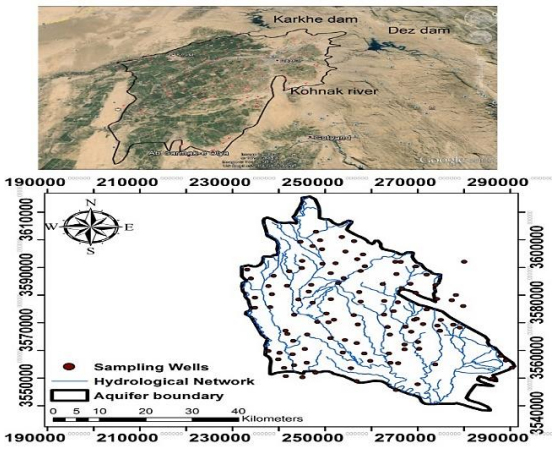


Fig. 1. Study area and sampling wells distribution

Table 1. Summary of the chemical composition of groundwater samples (n = 113) from Dezful-Andimeshk plain aquifer

Parameters	Maximum	Minimum	Average	Standard
Ca (mg/L)	485	30.8	96.60	68.83
Mg (mg/L)	304.8	8.28	44.56	47.113
Na(mg/L)	749.34	8.28	117.42	152.81
Cl(mg/L)	1117.00	17.04	152.25	216.96
SO4(mg/L)	2463.36	10.08	226.96	324.62
HCO_3(mg/L)	388.57	134.81	232.26	55.15
K(mg/L)	9.55	0.39	1.59	1.35
EC(µ mho/cm)	4660	299	1166.35	893.52
SAR(mg/L)	13.49	0.29	2.38	2.58
TH(mg/L)	2320	149.5	427.19	341.59
TDS(mg/L)	4127	10.43	754.21	695.73
pH	7.9	6.6	7.28	0.22

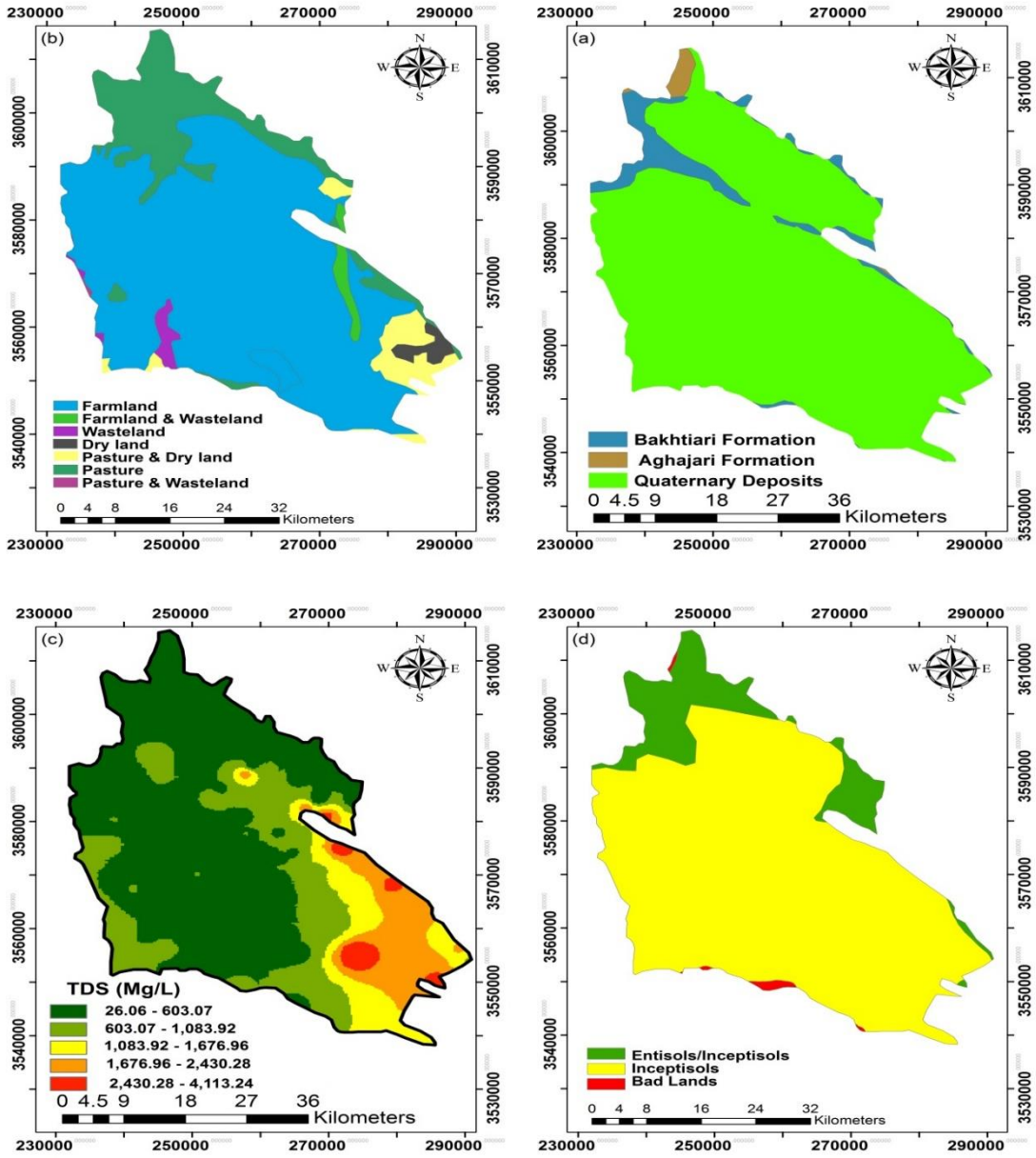


Fig. 2. Geology (a), Land use (b), TDS(c) and Soil order(d) maps in the study area

2.2. Particle Swarm Optimization Algorithm (PSO)

The particle swarm optimization algorithm is a population-based search technique that draws inspiration from animal social behavior. It considers a particle group of N particle that is exploring a global optimal solution in a D-dimension space. Vectors $X_i = (X_{i1},...,X_{iD})$, $V_i = (V_{i1},...,V_{iD})$ and $Pbest_i = (Pbest_{i1},..., Pbest_{iD})$ display each particle's position, velocity, and ideal personal position in that order. Every particle's position indicates a possible way to solve the optimization problem. By substituting a solution in an established objective function, a solution is valued. Each particle's velocity is indicated by the distance it travels in each dimension of the search space. The vector of the optimal personal position for each particle indicates the optimal position that it has visited. $gbest = (gbest_1,..., gbest_D)$ is a vector that represents the optimal global point that the entire particle swarm population has visited. Eq. (1), (2), and Fig. 3 update a particle's position and velocity, which establishes its situation.

$$v^{t+1}_{ij} = w \times v^t_{ij} + C_1 \times r_1 \times (Pbest^t_{ij} - x^t_{ij}) + C_2 \times r_2 \times (gbest^t_{ij} - x^t_{ij}) \quad (1)$$

$$x^{t+1}_{ij} = x^t_{ij} + v^{t+1}_{ij} \quad (2)$$

In this context V_i^t and V_i^{t+1} represent the velocity of particle *i* at iteration *t* and *t+1*, respectively. while X_i^t and X_i^{t+1} denote the position of particle *i* in iteration *t* and *t+1*. The term $Pbest^t_i$ refere to the best previous position of particle *i* in iteration *t*, and $gbest^t_i$ represent the best position among all particles at iteration *t*. Constants C_1 and C_2 are positive values, r_1 and r_2 are random numbers between 0 and 1 and ω is the inertia weight. The inertia weight is adjusted using Taguchi parameter tuning.

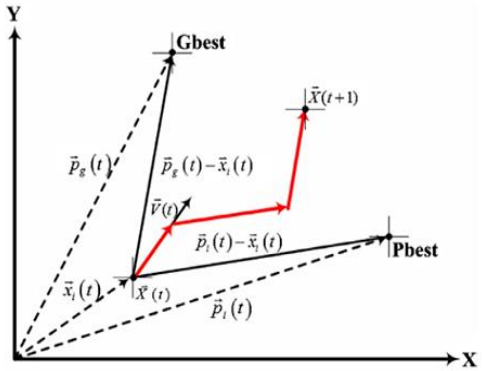


Fig. 3. Update the particle position

2.3. Combination of K-means clustering method and PSO

K-means is a basic and widely used unsupervised learning algorithm. It works by dividing a data set into a predetermined number of clusters (K clusters). To start, K centroids are selected, and their placement is crucial because different initial positions can lead to varying outcomes. Therefore, it is ideal to position the centroids as far apart as possible. The algorithm then assigns each data point to the nearest centroid. Once all points have been assigned, the initial grouping is complete. Next, the centroids are recalculated as the center points of the newly formed clusters. The process is repeated, reassigning data points to the closest updated centroids. This iterative cycle continues, causing the centroids to shift until their positions stabilize and no further movement occurs (Kuri-Morales and Rodriguez-Erazo, 2009).

In general, finding the centroid in a clustering algorithm involves first generating a random initial partition of the data into *k* non-empty clusters. Then, the mean vectors for each cluster are calculated to determine the centroids, and objects are assigned to the nearest centroid (Hatamlou, 2012). The algorithm continues by repeatedly calculating the mean vectors to update the cluster centroids and assigning objects to the corresponding clusters. This process repeats until the stopping criterion is met. The aim of the algorithm is to minimize an objective function, which is computed as follows:

An objective function is formulated using a Streamlined Silhouette Criterion Average (SSCA) (Hruschka et al., 2006; Covoes and Hruschka, 201). The silhouette criterion measures how well each object fits within its own cluster compared to objects in nearby clusters. This measure of proximity ranges from +1 to -1. A value of +1 indicates that an object

is far from any other clusters, while a value of 0 suggests the object is equally distant from neighboring clusters. A score of -1 indicates that the object has been incorrectly assigned to its cluster. Thus, the average silhouette value provides an indication of the overall separation between clusters. The Silhouette index for the i^{th} observation in the k^{th} cluster is calculated using Eq. 3:

$$s_{i,k} = \frac{b_{i,k} - a_{i,k}}{Max(b_{i,k}, a_{i,k})} \quad (3)$$

This index yields reliable K-means clustering results without requiring a predefined cluster structure (Handl et al., 2005). In this context, $a_{i,k}$ represents the average distance between observation i and the remaining (n_k -1) observations within the same k^{th} cluster. The value $b_{i,k}$ denotes the minimum average distance between observation i and all other observations in the nearest neighboring cluster, excluding the k^{th} cluster. In this study, the Euclidean distance was used to calculate these values. As a result, the most effective clustering method is the one that allows the user to maximize the SSAC, as calculated in the following manner (Brida et al., 2012):

$$\max \text{SSCA} = \frac{1}{k}\sum_{k=2}^{k}\frac{1}{n_k}\sum_{i=1}^{n_k} s(i,k) \quad (4)$$

The K-means algorithm has two key limitations: it can get stuck with suboptimal centroids, and it requires the user to specify the number of clusters (k) before starting. However, PSO-based clustering helps mitigate these issues. By updating position and velocity, PSO can identify the best particle, which allows the algorithm to select more suitable initial centroids. This reduces the likelihood of the algorithm getting stuck at suboptimal solutions. A further advantage of PSO is that it often identifies fewer clusters when using K-means within a set of k centroids. This leads to at least one cluster being more distant from the others, enabling the optimization process to focus on the best, most promising solutions, even if it results in a reduction of k.

2.4. Implementation of Hybrid K-means with PSO

The proposed hybrid algorithm combines Particle Swarm Optimization (PSO) with the K-means clustering algorithm to optimize the clustering process. The following steps outline the algorithm's implementation:

Step 1: Initialization of the Particle Population
- The process starts by randomly initializing the positions of the particles, where each particle represents a potential solution.
- Each particle's position corresponds to a set of cluster centroids.
-The number of particles in the swarm and the number of clusters (K) are predefined. The velocity of each particle is also initialized randomly.

$$population = \begin{bmatrix} particle_1 \\ particle_2 \\ ..... \\ particle_{nParticle} \end{bmatrix} \quad (5)$$

Where:
$Particle_i = [Center_1, Center_2, ..Center_K]$, $i = 1, 2, ..., nParticle$, and K is the number of clusters.

Step2: K-means Assignment
Each particle's position (i.e., the centroids) is used to assign each data point in the dataset to the nearest centroid, following the K-means approach. The objective is to minimize the intra-cluster distance, which is calculated using the Euclidean distance between data points and centroids. The objective function is calculated for each particle as follows:
Step 2-1: $i = 1$
Step 2-2: $k = 1,..,,K,,$

Step 2-2-1: If $cluster(i) = k$,
Step 2-2-1-1: $a(i) = 0;$,

Step 2-2-1-2: for all samples in the cluster (i),
Step 2-2-1-3: Calculate the distances between the i^{th} sample and i'^{th} samples, then update the objective function by adding the calculated distance to the value of the objective function.

$$a(i) = a(i) + \sqrt{\sum_{j=1}^{d}(Y_{i'}^{j} - Y_{i}^{j})^2} \quad (6)$$

End
Else
Step 2-2-2: $b(i,k) = 0;$
Step 2-2-3: for all samples in the cluster (k)

Step 2-2-3-1: calculate the distances between the i^{th} sample and i'^{th} samples, then add the value of $b(i,k)$ with the distance calculated below:

$$b(i,k) = b(i,k) + \sqrt{\sum_{j=1}^{d} (Y_{i'}^{j} - Y_{i}^{j})^2}, i = 1,2,...K \quad (7)$$

End

End

Step 2-3: Calculate the silhouette index:

$$s(i,k) = \frac{b_{i,k} - a_{i,k}}{Max\ (b_{i,k}, a_{i,k})} \quad (8)$$

End

Step 2-4: Get the SSCA:

$$S = \frac{1}{k} \sum_{k=2}^{k} \frac{1}{n_k} \sum_{i=1}^{n_k} s\ (i,k) \quad (9)$$

Step3: Objective Function Evaluation
After the assignment, the objective function (e.g., sum of squared errors or another distance-based metric) is computed for each particle. Additionally, the Silhouette index is calculated to assess cluster cohesion and separation, ensuring clusters are well-defined and distinct.

Step 4 - Loop 1: Update Personal and Global Bests
The velocity and position of each particle are calculated using equations (1) and (2). These equations are utilized to update the velocity and position of the particles during each iteration.

Step 4 - Loop 2: Particle Evaluation
Once the positions of the particles have been updated, the cost for each particle must be computed in Step 2.

Step 4 - Loop 3: Update Personal and Global Best Positions
If the current cost of a particle is lower than its previous cost, this position is recorded as pBest. Identify the particle with the best cost among all particles. If this current cost is better than the previous gBest value, then the position of this particle becomes the new gBest. Otherwise, the last recorded pBest and gBest remain as the best personal and global positions, respectively. In other words, If (f (x_i) < f (p_i)), then update the particle's best known position to ($p_i \leftarrow x_i$), and If (f (p_i) < f (g)), then update the swarm's best known position to (g ← p_i).

Step 4 - Loop 4: Check the Number of Iterations
Evaluate the number of iterations. If the maximum number of iterations has been reached, the process will stop. If not, return to Step 4 - Loop 1. The flowchart illustrating the proposed hybrid algorithm is shown in Fig. 4.

Step 5: Velocity and Position Update
The velocity and position of each particle are updated using the PSO update equation:

$$v_i^{t+1} = wv_i^t + c_1 r_1 (p_i^t - x_i^t) + c_2 r_2 (g^t - x_i^t)$$
$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

Here, v_i is the velocity, x_i is the position, p_i is the personal best, and g is the global best. Constants w, c_1, and c_2 are parameters controlling the algorithm's balance between exploration and exploitation.

Step 6: Local Search with K-means
After updating the positions, a local search is performed using the K-means algorithm. The PSO provides better initial centroids, reducing the likelihood of K-means getting stuck in suboptimal solutions. This step accelerates convergence by refining the clusters.

Step 7: Iteration and Termination
The process iterates, updating particle velocities and positions until a stopping criterion is met (e.g., a maximum number of iterations or convergence of the objective function). Once the algorithm converges, the global best particle's position provides the optimal clustering solution. This hybrid approach benefits from PSO's global search capabilities, which help overcome K-means' sensitivity to initial centroids, while K-means contributes to faster local optimization.

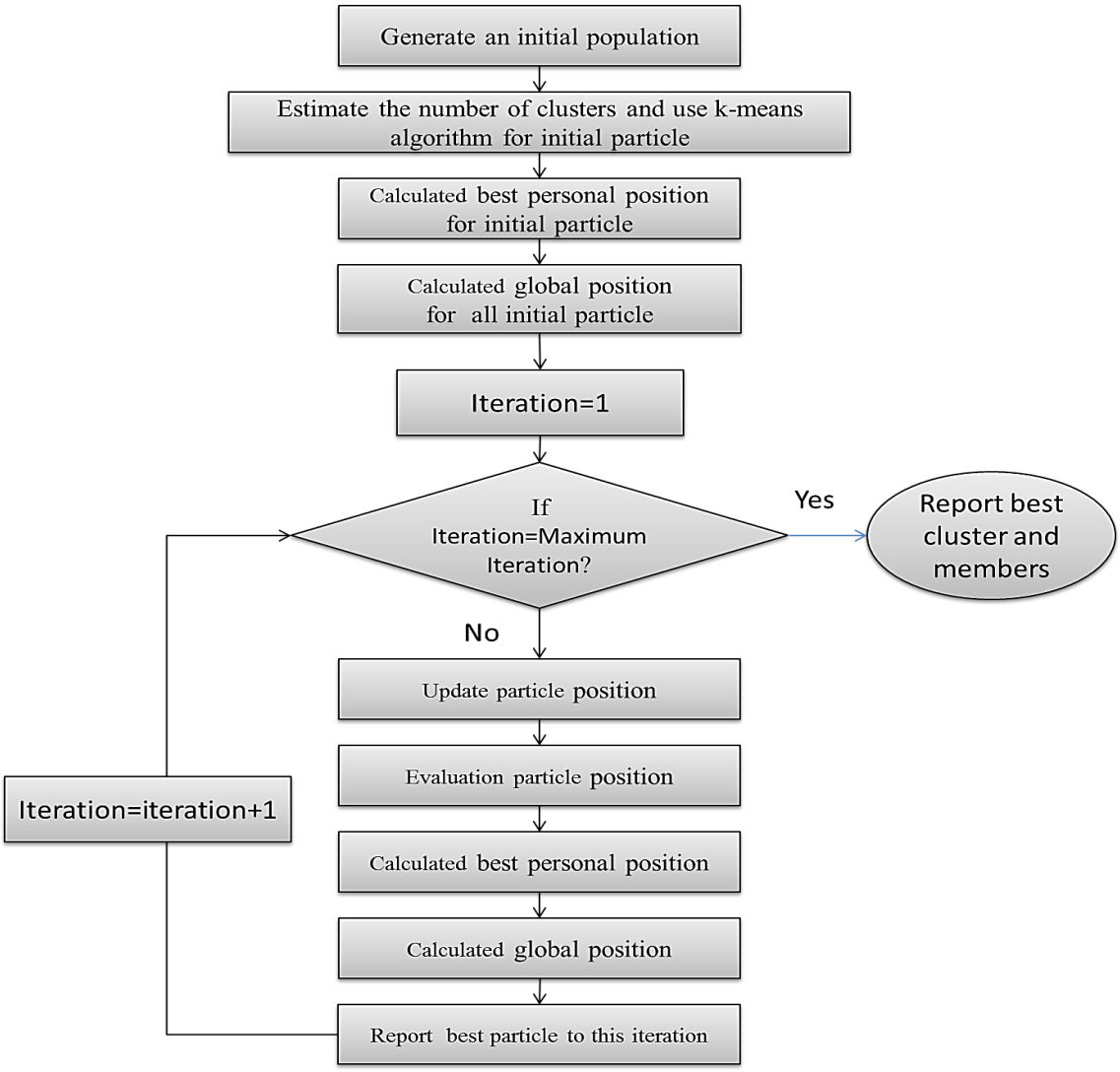


Fig. 4. Flowchart of proposed hybrid algorithm

3. Results and discussion

After different runs with different number of iterations, the optimal response for each cluster is obtained in Table 2:

3-1. Optimal number of clusters

The results indicate that the minimum value of the cost function ranges from 14.44 for 2 clusters to 8.73 for 7 clusters (Table 2). While increasing the number of clusters continues to reduce the best cost value, leading to improved outcomes, it is important to balance this with clustering principles. Specifically, both separation criteria (cost function) and cohesion must be considered to ensure meaningful cluster analysis (Khoshnevisan et al., 2014). In this research, silhouette analysis was employed to assess the level of cohesion. The number of clusters can be increased as long as the silhouette value does not decrease (Nourani et al., 2012). According to the findings presented in Table 3, a reverse trend in the silhouette criterion was noted between 5 and 6 clusters in this case study. As a result, the optimal number of clusters is determined to be 5. Since the same situation occurred between cluster 2 and 3, we can stop the process. However, the best cost amount obtains from the 2 clusters do not satisfy the problem and that is why we continue the process until receiving more reasonable responses.

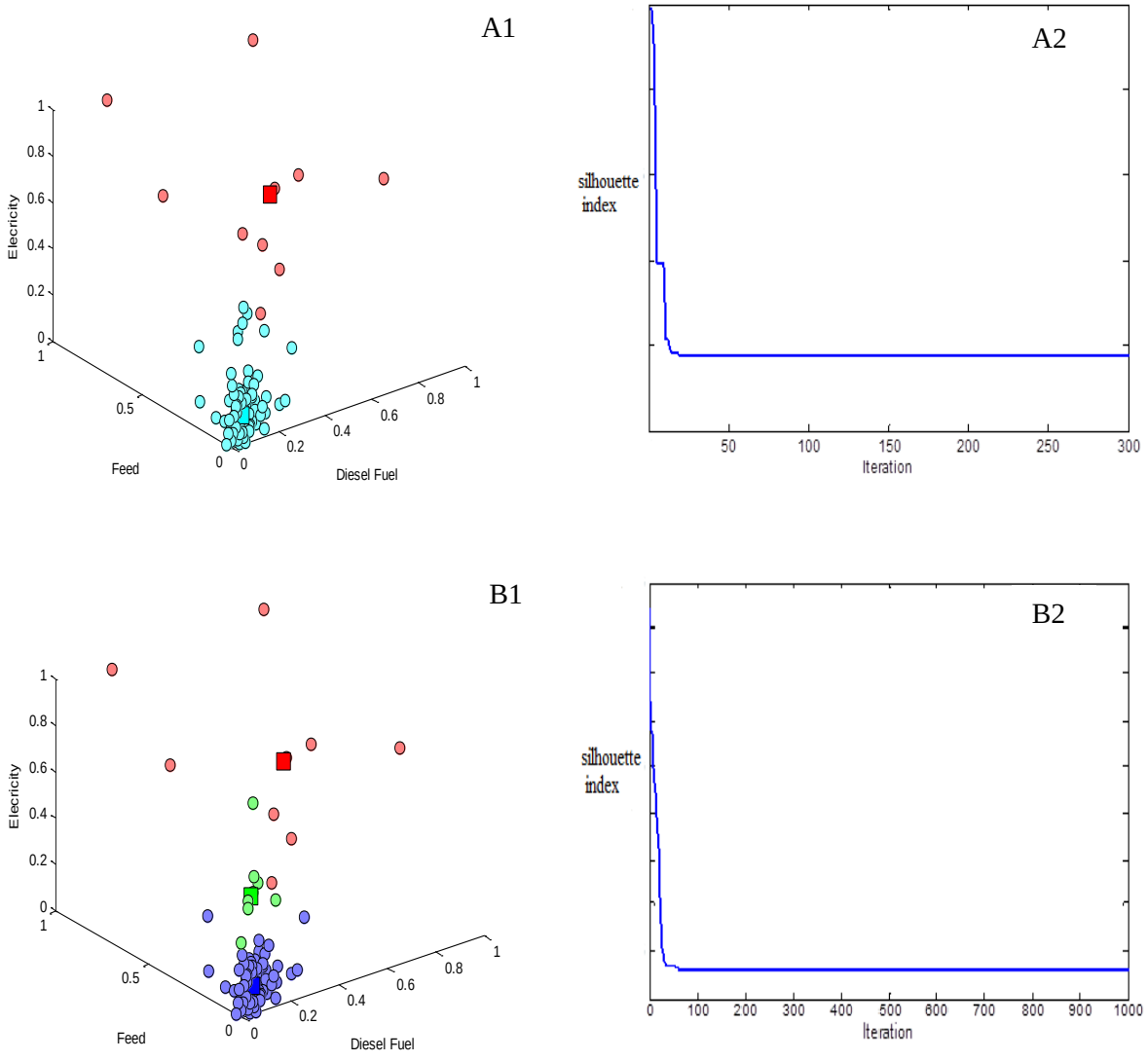
Table 2. Optimal response for each cluster

Number of Cluster	itt	Silhouette index	Best Cost
2	300	0.75	14.44
3	1000	0.72	12.29
4	1000	0.72	11.61
5	1000	0.73	10.68
6	300	0.65	9.31

Table 3. Overall clustering results

Number of clusters	2	3	4	5	6	7
Best cost	14.44	12.29	11.61	10.68	9.31	8.73
Silhouette	0.75	0.72	0.72	0.73	0.65	0.65

Fig. 5 presents the multidimensional scaling and clustering behaviors for different clustering methods, ranging from two to five clusters. The visual representation helps to illustrate how the PSO algorithm clusters the sampling wells based on groundwater quality data, emphasizing the progressive refinement in cluster separation as the number of clusters increases. The silhouette scores across the different clustering scenarios offer a measure of how well the samples fit within their respective clusters, with five clusters being the most optimal. The silhouette index is highest for the two-cluster method but drops for larger numbers of clusters, stabilizing around five clusters. This visualization underscores the balance between cluster cohesion and separation, a key factor in determining the optimal number of clusters.



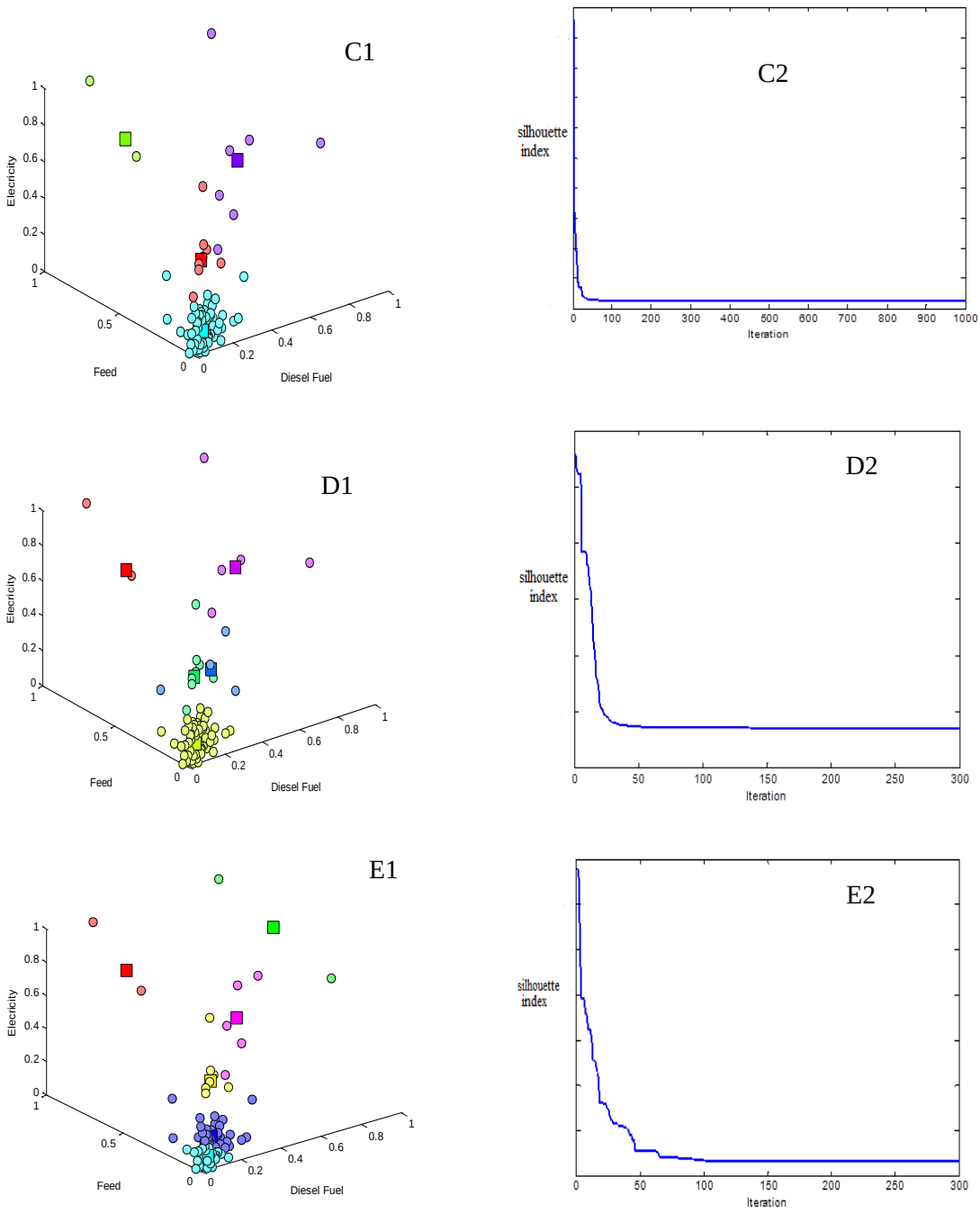


Fig. 5$_{A1\text{-}2,\ B1\text{-}2,\ C1\text{-}2,\ D1\text{-}2,\ E1\text{-}2}$**.** Multidimensional Scaling and Clustering Behaviors of Groundwater Quality Samples Using PSO Algorithm with Silhouette Method for Two to Six Clusters. A1-A2: Results of two clusters, showing the initial division of groundwater sampling sites, where a high silhouette index indicates a clear separation between the two clusters, B1-B2: Results of three clusters, illustrating an additional separation that further refines the distinction between regions with different groundwater qualities, C1-C2: Results of four clusters, continuing to break down the study area into more defined zones based on water quality parameters, D1-D2: Results of five clusters, where the clustering reaches optimal separation, as indicated by the silhouette index, highlighting regions of higher groundwater contamination and distinct water quality characteristics, and E1-E2: Results of six clusters, which show a decline in silhouette index, indicating over-clustering, with diminishing returns in the separation and interpretation of clusters.

Fig. 6 provides a spatial zonation of the clustered sampling wells. Each of the subfigures (a-e) represents the proposed clustered zonation for methods with different cluster counts. These zonations show how the sampling locations are divided across the study area, and the clustering reflects the variations in groundwater quality influenced by factors like geology, land use, and hydrological networks. For example, the figure reveals how the southeast region of the study area (with high TDS levels influenced by the Kohnak River) is consistently separated from other zones. As the clustering progresses from two to five clusters,

the spatial patterns become more distinct, and the distribution of sampling wells more accurately reflects the underlying groundwater quality variations. The combination of Figs. 5 and 6 highlights the robustness and accuracy of using PSO and the Silhouette method to optimize groundwater monitoring efforts by reducing the number of sampling wells without compromising the spatial and qualitative representation of groundwater conditions.

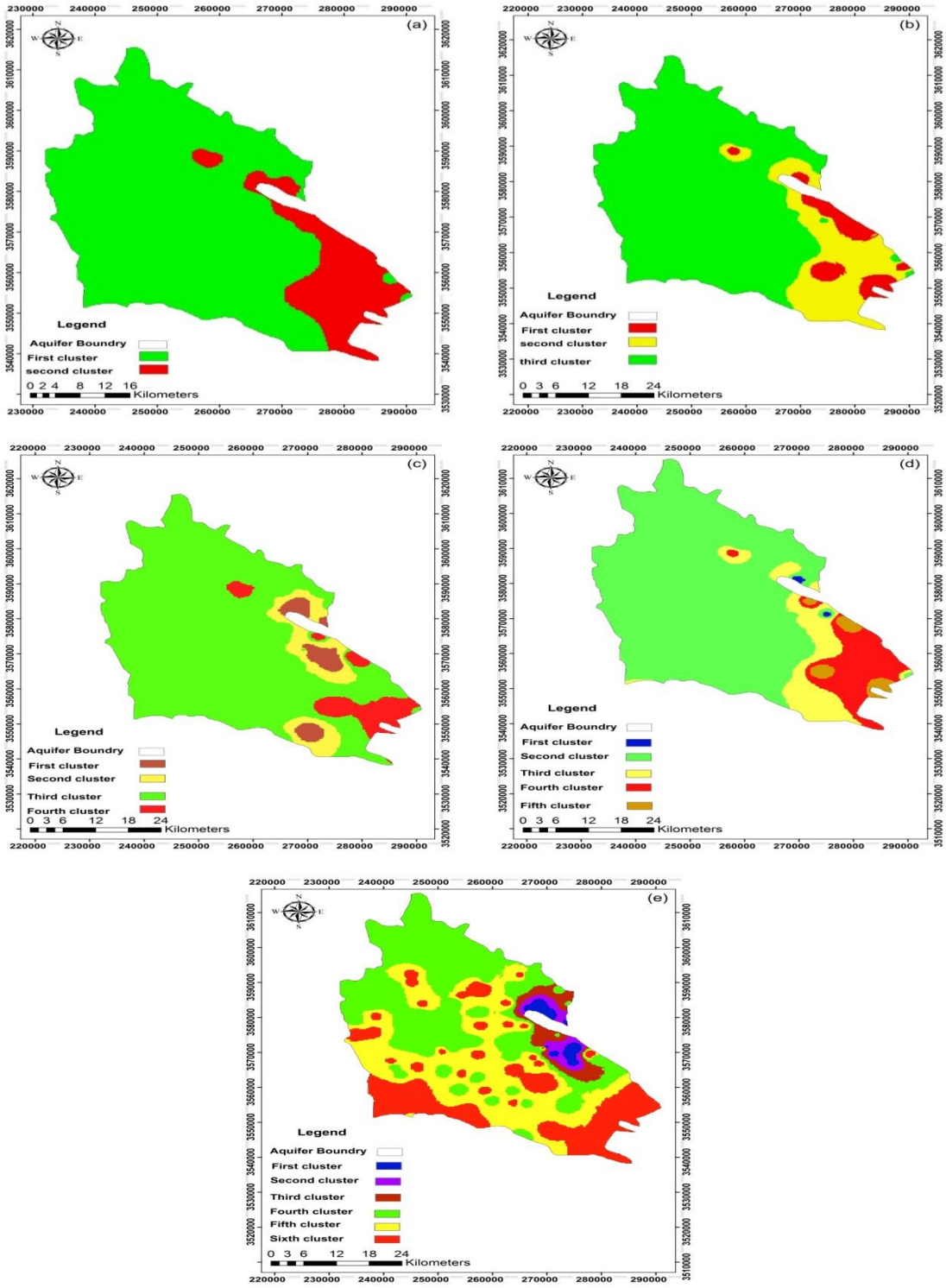


Fig. 6$_{a, b, c, d, e}$. Sampling wells clustered zonation proposed using the PSO algorithm and the Silhouette method for first to five methods respectively. a: Zonation of two clusters, representing the broadest distinction in groundwater quality distribution, b: Zonation of three clusters, providing more refined spatial clustering based on groundwater quality, c: Zonation of four clusters, further dividing the study area into zones that reflect additional groundwater quality characteristics, d: Zonation of five clusters, showing the optimal clustering result, which aligns closely with land use, geology, and hydrological factors, and e: Zonation of six clusters, where over-clustering starts to emerge, with a lower silhouette index.

3.2. clusters and land condition

Groundwater flowing beneath the Earth's surface can experience both physical and chemical alterations, which may result from natural processes or human activities. In this study parameters such as Geology, Soil, Land use and Hydrological network were used to changes in groundwater quality (Clark, 2015) will be investigated and assessing the accuracy of PSO algorithm for clustering of groundwater quality. So, we assessed the land use, geology, Soil orders (Fig. 6) and hydrological network (Fig. 1) in the study area. According to results, the main geological formation in the study area is Quaternary deposits, so this parameter is same in most parts of the aquifer and this parameter could not make difference in clustering results.

Soil order is also Inceptisols in most parts of the plain. Likewise, Land use is farmland in most parts of the study area, but it changes to pasture and dry land in southeast. So, this parameter should make change in water quality. So, at first sight, the groundwater quality in pasture land area should be better than farm land area (due to fertilizer and pesticide), but in this study, TDS in this part (Fig. 2c) is more than farmland area. The main cause of this change is likely the Kohnak River. The Kohnak River as the main hydrological network of south-east of the study area, passes from Gachsaran formation (rock salts such as Halite, Anhydrite, Red to Gray marls) and inject the natural salts to the aquifer (groundwater table in this area is around 10 meter). So, groundwater quality in these parts of the aquifer should be lower than other parts of the aquifer. It should be noted that, Karkheh and Dez dams are in the Northwest and Northeast of the aquifer and water that discharge from these big dams has good quality and recharge the Dezful-Andimeshk plain aquifer. So, According to the TDS map (Fig. 2c) and optimization methods (Figs. 4 to 6) and clustering maps (Fig 6.$_{a,b,c,d,e}$), the best number of cluster is 5.

4. Conclusion

In this study, a groundwater quality samples were clustered based on the PSO algorithm. The optimization of the clustering process carried out according to the Silhouette index. In line with the principles of clustering analysis, both separation criteria (cost function) and cohesion should be considered. In this study, silhouette analysis is used to evaluate the level of cohesion. According to the results, the optimal number of clusters is 5. For verification the results of clustering, we used some of the natural factors or human activities (such as geology, soil order, land use, hydrological network and TDS maps). Based on the results of the verification, in the study area, the number of clusters considered for clustering of the groundwater quality was confirmed. Consequently, it can be concluded that clustering groundwater quality samples using the PSO algorithm and Silhouette index optimizer yields acceptable results. Using this method, we can decrease the number of sampling locations and consequently, we can significantly decrease the costs, save the time and increase the accuracy of the data analyze in the assessment of the groundwater quality programs.

Acknowledgments

We would also like to thank the reviewers for their insightful comments, which significantly improved the quality of our work. This research received no specific financial support.

References

Abu-khalaf, N., Khayat, S. & Natsheh, B., 2013. Multivariate Data Analysis to Identify the Groundwater Pollution Sources in Tulkarm Area / Palestine. *Science and Technology*, 3(4), 99–104. doi.org/10.5923/j.scit.20130304.01.

Barak, S. & Modarres, M., 2014. Developing an approach to evaluate stocks by forecasting effective features with data mining methods. *Expert Systems with Applications,* 42, 1325–1339.

Brida, J.G., Disegna, M. & Osti, L., 2012. Segmenting visitors of cultural events by motivation: A sequential non- linear clustering analysis of Italian Christmas Market visitors. *Expert Systems with Applications,* 39, 11349–11356.

Cardoso, M.F., Salcedo, R.L. & Feyo de Azevedo, S., 1996. The simplex-simulated annealing approach to continuous non-linear optimization. *Computers & Chemical Engineering*, 20, 1065–1080. https://doi.org/10.1016/j.jmgm.2005.11.005.

Clark, I., 2015. *Groundwater geochemistry and isotopes*. Taylor & Francis.

Covoes, T.F. & Hruschka, E.R., 2011. Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences*, 181, 3766–3782. https://doi.org/10.1080/01402390.2011.569130.

Dastorani, M., Mirzavand, M., Dastorani, M.T. & Khosravi, H., 2020. Simulation and prediction of surface water quality using stochastic models. *Sustainable Water Resources Management,* 6, 1-17.

Dorigo, M., Maniezzo, V. & Colorni, A., 1996. The ant system: optimisation by a colony of cooperating agents. *IEEE Trans. Syst. Man Cybern. Part B: Cybern*, 26, 29–41.

Duan, Q., Sorooshian, S. & Gupta, V.K., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, 28, 1015–1031.

Eskandari, E., Mohammadzadeh, H., Nassery, H., Vadiati, M., Zadeh, A.M. & Kisi, O., 2022. Delineation of isotopic and hydrochemical evolution of karstic aquifers with different cluster-based (HCA, KM, FCM and GKM) methods. *Journal of Hydrology*, 609. https://doi.org/10.1016/j.jhydrol.2022.127706.

Geem, Z.W., Kim, J.H. & Loganathan, G.V., 2001. A new heuristic optimization algorithm: harmony search. *Simulation*, 76, 60–68.

Handl, J., Knowles, J. & Kell, D.B., 2005. Computational cluster validation in post-genomic data analysis. *Journal of Biological Chemistry*, 21, 3201–3212.

Hatamlou, A., 2012. In search of optimal centroids on data clustering using a binary search algorithm. *Pattern Recognition Letters*, 33, 1756–1760.

Hossain, M.G., Selim Reza, A.H.M., Lutfun-Nessa, M. & Ahmed, S.S., 2013. Factor and cluster analysis of water quality data of the groundwater wells of Kushtia, Bangladesh: Implication for arsenic enrichment and mobilization. *Journal of the Geological Society of India*, 81, 377–384. https://doi.org/10.1007/s12594-013-0048-0.

Hruschka, E.R., Campello, R.J. & De Castro, L.N., 2006. Evolving clusters in gene-expression data. *Information Sciences*, 176, 1898–1927.

Jha, M. & Datta, B., 2013. Three-Dimensional Groundwater Contamination Source Identification Using Adaptive Simulated Annealing. *Journal of Hydrologic Engineering*, 18, 307–317.

Karaboga, D., 2005. An idea based on honey bee swarm for numerical optimization. Technical Report-TR06, Department of Computer Engineering, Engineering Faculty, *Erciyes University*.

Kennedy, J. & Eberhart, R., 1995. Partic le swarm optimization, in: *The Sixth International Symposium on Micro Machine and Human Science. Nagoya, Japan*, pp. 39–43.

Khoshnevisan, B., Bolandnazar, E., Barak, S., Shamshirband, S., Maghsoudlou, H., Altameem, T.A. & Gani, A., 2014. A clustering model based on an evolutionary algorithm for better energy use in crop production. *Stochastic Environmental Research and Risk Assessment*, 1–15.

Kuri-Morales, A. & Rodriguez-Erazo, F., 2009. A search space reduction methodology for data mining in large databases. *Engineering Applications of Artificial Intelligence*, 22, 57–65.

Li, Y. & Chan Hilton, A.B., 2005. Reducing Spatial Sampling in Long-Term Groundwater Monitoring Networks Using Ant Colony Optimization. *International Journal of Computational Intelligence Research*, 1, 19–28.

Mehrabian, A.R. & Lucas, C., 2006. A novel numerical optimization algorithm inspired from weed colonization. *Ecological Informatics*, 1, 355–366.

Mirzavand, M. & Ghazavi, R., 2015. A Stochastic Modelling Technique for Groundwater Level Forecasting in an Arid Environment Using Time Series Methods. *Water Resources Management*, 29. https://doi.org/10.1007/s11269-014-0875-9.

Mirzavand, M. & Walter, J., 2024. Delineating the mechanisms controlling groundwater salinization using chemo-isotopic data and meta-heuristic clustering algorithms (case study: Saguenay-Lac-Saint-Jean region in the Canadian Shield, Quebec, Canada). *Environmental Science and Pollution Research*, https://doi.org/10.1007/s11356-024-33922-6.

Nourani, V., Hosseini Baghanam, A. & Daneshvar Vousoughi, F., Alami, M.T., 2012. Classification of Groundwater Level Data Using SOM to Develop ANN-Based Forecasting Mode. *International Journal of Soft Computing and Engineering*, 2, 464–469.

Passino, K.M., 2002. Biomimicry of bacterial foraging for distributed optimization and control. *IEEE Control Systems Magazine*, 52–67.

Saruhan, H., 2014. Differential evolution and simulated annealing algorithms for mechanical systems design. *International journal of Engineering Science and Technology*, 17, 131–1363. https://doi.org/10.1192/bjp.205.1.76a.

Shamshirband, S., Amini, A., Anuar, N.B., Kiah, M.L.M., Wah, T.Y. & Furnell, S., 2014. D-FICCA: a density-based fuzzy imperialist competitive clustering algorithm for intrusion detection in wireless sensor networks. *Measurement*, 55, 212–226. https://doi.org/https://doi.org/10.1016/j.measurement.2014.04.034.

Storn, R. & Price, K., 1997. Differential Evolution - a Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11, 341–359.

Yang, X.S. & Deb, S., 2010. Engineering optimization by cuckoo search. *International Journal of Mathematical Modelling and Numerical Optimisation*, 1, 330–343.