Applied Water Science (2024) 14:170
https://doi.org/10.1007/s13201-024-02233-z

ORIGINAL ARTICLE

# Assessing the groundwater quality of El Fahs aquifer (NE Tunisia) using multivariate statistical techniques and geostatistical modeling

**Constantinos F. Panagiotou[1] · Anis Chekirbane[2] · Marinos Eliades[1] · Christiana Papoutsa[1] · Evangelos Akylas[1] · Marinos Stylianou[3] · Nikolaos Stathopoulos[4]**

Received: 10 August 2023 / Accepted: 24 June 2024 / Published online: 9 July 2024


## Abstract

This study is the first attempt to characterize the quality status of El Fahs aquifer by combining graphical tools, multivariate statistical techniques and traditional geostatistical methods. Water samples are collected from thirty-six observation wells during April 2016 to characterize the physicochemical properties of the aquifer. Subsequently, these samples are partitioned into three hydrochemically distinct water classes (i.e., C1, C2, and C3) using the *K*-means clustering method. Principal Component Analysis is used to reduce the dimensionality of the dataset prior performing the clustering computations, resulting in clusters of higher quality than the non-reduced case in terms of Silhouette coefficient. Piper diagram is used to display the chemical composition of the samples, revealing the dominant role of Mg–Ca–Cl water type for all three classes, whereas Sodium and Sulfate were found to be the second most important cations and anions respectively. Indicator kriging (IK) is used to identify the probability of occurrence of the hydrochemical classes beyond the sampling locations. It is found that Class 1, associated with fresh groundwater component, is most probable to occur at the central part of the plain, mainly due to the presence of a dense hydrological network, whereas Classes 2 (agricultural activities) and 3 (dissolution of evaporate geological formations) are expected to occur at the southern and northern regions respectively. IK also identified the regions associated with high levels of uncertainty, mostly occurring in a large portion of the northern area due to the absence of available hydrochemical information. The results showed that integration of graphical methods, multivariate statistical techniques and geostatistical modeling, is an efficient approach for characterizing the hydrochemical status of the aquifer system, to spatially optimize the groundwater monitoring well networks and quantify the uncertainty levels of the water classes in a systematic way.

**Keywords** Groundwater quality · Clustering analysis · Principal component analysis · Indicator kriging · Tunisia

✉ Constantinos F. Panagiotou
constantinos.panagiotou@eratosthenes.org.cy

Anis Chekirbane
anis.chkirbene@inat.ucar.tn

Marinos Eliades
marinos.eliades@eratosthenes.org.cy

Christiana Papoutsa
christiana.papoutsa@eratosthenes.org.cy

Evangelos Akylas
evangelos.akylas@cut.ac.cy

Marinos Stylianou
marinos.stylianou@ouc.ac.cy

Nikolaos Stathopoulos
n.stathopoulos@noa.gr

[1] ERATOSTHENES Centre of Excellence, Limassol, Cyprus

[2] Department of Rural Engineering, Water and Forests, INAT, University of Carthage, Tunis, Tunisia

[3] Laboratory of Chemical Engineering and Engineering Sustainability, Faculty of Pure and Applied Sciences, Open University of Cyprus, Nicosia, Cyprus

[4] Operational Unit BEYOND "Centre for Earth Observation Research and Satellite Remote Sensing", Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens, Athens, Greece

## Introduction

Aquifer systems provide a significant portion of water for human consumption, as well as agricultural and industrial use. However, available water resources are limited and diminishing due to human activities such as population growth, urbanization, and improved living standards. Groundwater quality is becoming increasingly concerning and has been the subject of numerous studies in various areas of Tunisia (Chekirbane et al. 2022; Aouiti et al. 2021; Houatmia et al. 2016). It represents a constant risk to the availability of these water sources.

Currently, the water crisis is worsening due to the unsustainable use of water resources and inadequate sanitation. Consumption is accompanied by a decline in the quality of these resources, usually associated with overexploitation and aquifer pollution (Mastrocicco and Colombani 2021; Eliades et al. 2023). This has become a major topic in geochemistry, including the salinization of groundwater, as various factors can contribute to the emergence of excessive dissolved salts from different sources. This issue is also observed in the El Fahs plain aquifer system in the Zaghouan Governorate (northeast of Tunisia), where water resources are of paramount importance in satisfying irrigation and drinking water needs. The poor quality and vulnerability of the water in this system currently poses problems for its use in agriculture and drinking water supply.

Groundwater systems support a diversity of subsurface microbiota and fauna, which are the major drivers of an enormous number of (bio)geochemical processes (Hancock et al. 2005). Additionally, the groundwater chemical status is known to play a vital role in many significant ecological processes (Moore 1999). Therefore, it is essential to acquire a good understanding of the groundwater processes that affect the hydrochemical state of the aquifer, allowing the design of sustainable water strategies. To achieve that, information regarding various properties of the subsurface system is required.

Since the early 1920s, a plethora of graphical and multivariate statistical methods has been proposed to assist the classification of water samples in terms of hydrochemical considerations. Most of the graphical methods are designed to represent the relative proportions of certain major ionic species (Hem 1989; Güler et al. 2002; Piper 1944; Collins 1923; Stiff 1951; Schoeller 1962). Among these methods, the Piper diagram (Piper 1944) is the most commonly used graphical technique (Güler et al. 2002), displaying the major anions and cations on two separate ternary diagrams, along with a central diamond-shaped diagram that contains the projected values from the two separate ones. Other graphical tools, for example Collins (1923) pie and Stiff (1951) diagrams, generate single graphs for each sample, thus being impractical to sort and classify large datasets.

A major drawback of these methods is the use of a limited number of parameters, contrary to multivariate statistical techniques that can utilize all the available datasets. Multivariate statistical techniques are commonly used for understanding the physicochemical status of water samples by identifying statistical associations among dissolved constituents and environmental parameters (Drever 1997), assisting in the interpretation of hydrochemical processes. For example, clustering analysis is often used for classifying water samples into hydrochemical groups according to their similarity (Abu-alnaeem et al. 2018; Güler et al. 2012; Acikel and Ekmekci 2018). A clustering technique often used in groundwater applications is the *K*-means algorithm due to its simplicity, ease of implementation and computational efficiency (Güler et al. 2002; Masoud 2014; Javadi et al. 2017). Another popular multivariate approach is Principal Component Analysis (PCA), which is proven to be an effective tool for characterizing the status of aquifer systems, to identifying major descriptors of groundwater processes (Abu-alnaeem et al. 2018; Javadi et al. 2017) and reducing the dimensionality of the available datasets prior performing clustering computations (Güler et al. 2002). Guler et al. (2002) compared different graphical and statistical multivariate techniques in terms of their ease of use and ability to partition water samples into groups. They concluded that the combination of both types of methods provides an efficient methodology that retains its advantages while minimizing the limitations of each approach.

Geostatistics is another branch of statistics that uses spatial information to describe the spatial continuity of groundwater systems, proven to be an effective tool in the characterization of heterogeneities in such complex media (Issaks and Srivastava 1989; Delfiner and Chilès 2012). It provides a comprehensive framework for combining different types of datasets to provide estimates of attributes/parameter values at unsampled locations, and for building local models of spatial uncertainty. Kriging is a geostatistical interpolation method that is extensively used in groundwater contamination problems for mapping the spatio-temporal evolution of groundwater quality parameters (Adhikary et al. 2011; Yimit et al. 2011; Deepika et al. 2020).

There is an extensive number of studies in the literature that use and/or combine these techniques to assess the status of groundwater systems located in the Eastern Mediterranean, Middle East, and North Africa (EMMENA) region (M'nassri et al. 2019; Mejri et al. 2018; Hajji et al. 2021; Makni et al. 2013; Benmarce et al. 2023; Panagiotou et al. 2023). For example, Benmarce et al. (2023) conducted hydrochemical, statistical, and isotopic analyses to characterize the hydrochemical status of groundwater systems in

the Guelma Basin (North-Eastern Algeria). Principal Component Analysis and correlation matrix revealed the strong influence of water-rock interactions and various exchange mechanisms on the mineral content, whereas the dominant ions were found to be chloride, bicarbonate, calcium, sodium and magnesium. The Schoeller, Berkalof, and Piper diagrams were used to visualize the hydrochemical facies of three groundwater systems in a graphical form. The presence of high concentration levels of magnesium and calcium ions was associated with the dissolution of carbonate minerals, which also contributed to the alkalinity status of the groundwater system. A recent study conducted by Ncibi et al. (2022) focused on the identification of the origins of nitrate pollution in the Sidi Bouzid semi-arid basin (Tunisia) via the integration of physical models (MODFLOW) and multivariate statistical tools. Principal Component Analysis was used to correlate nitrates with other dissolved species, providing insights into the role of the groundwater processes. MODFLOW and particle tracking predictions suggested that groundwater recharge and abstraction from wells as the main drivers of the groundwater evolution. M'nassri et al. (2019) combined multivariate statistical tools with a traditional geostatistical interpolation to estimate the spatial heterogeneity of major ionic species that are present in a shallow unconfined aquifer located in the Ouled Chamekh plain in central-eastern Tunisia. The results of factor and principal component analyses revealed that the deterioration of the quality status is mainly attributed to natural processes, such as rock weathering, whereas human-induced activities are marginal, associated with artificial groundwater recharge

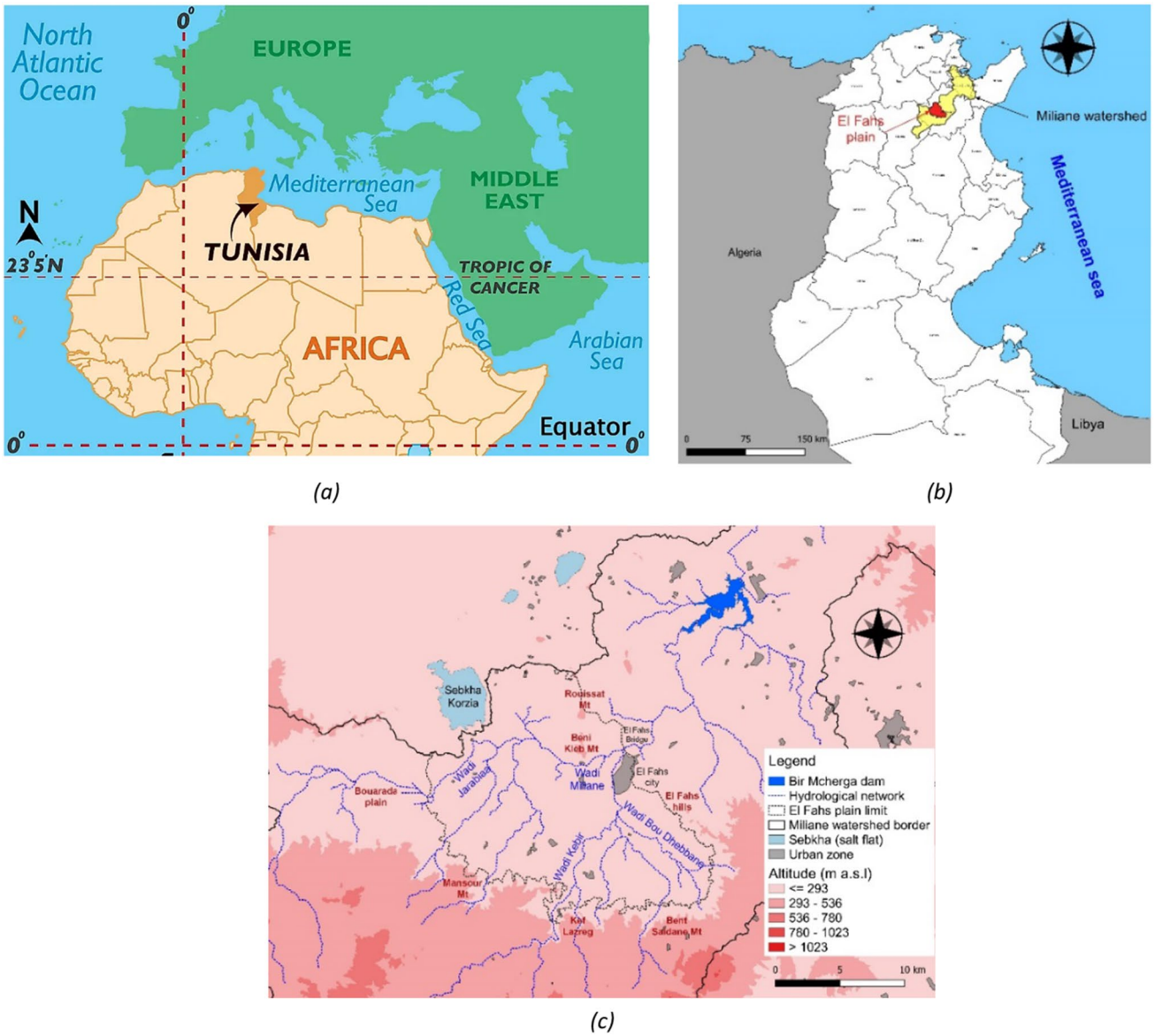
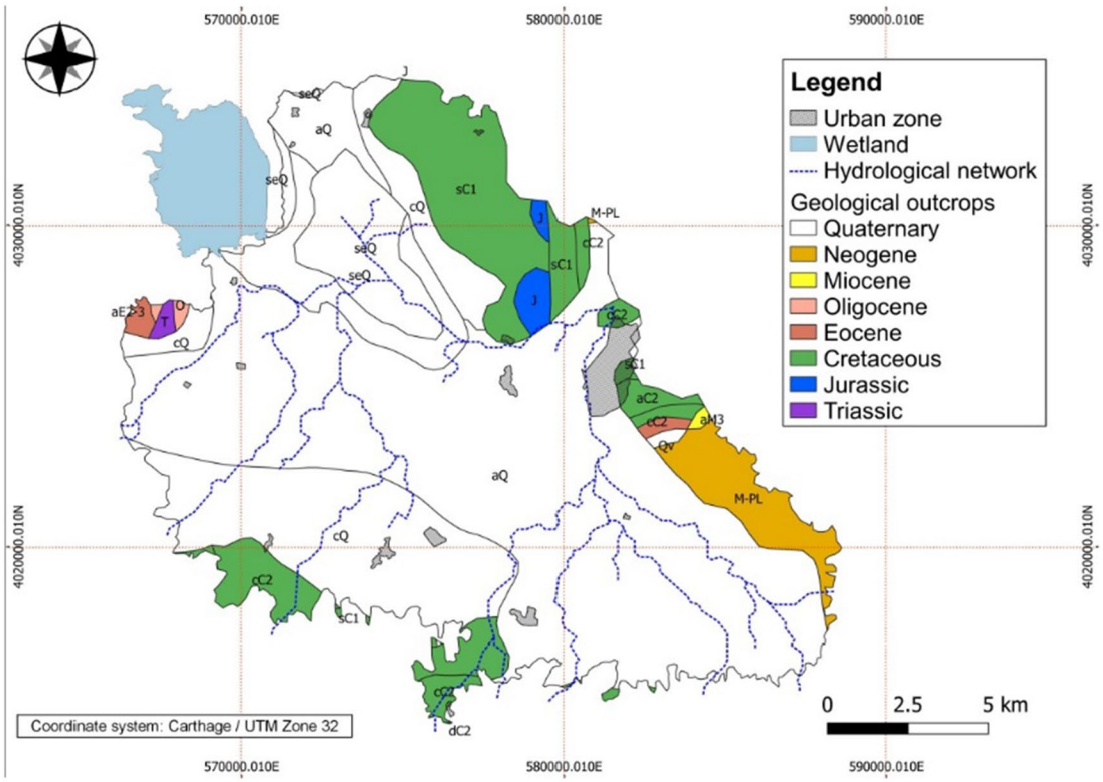
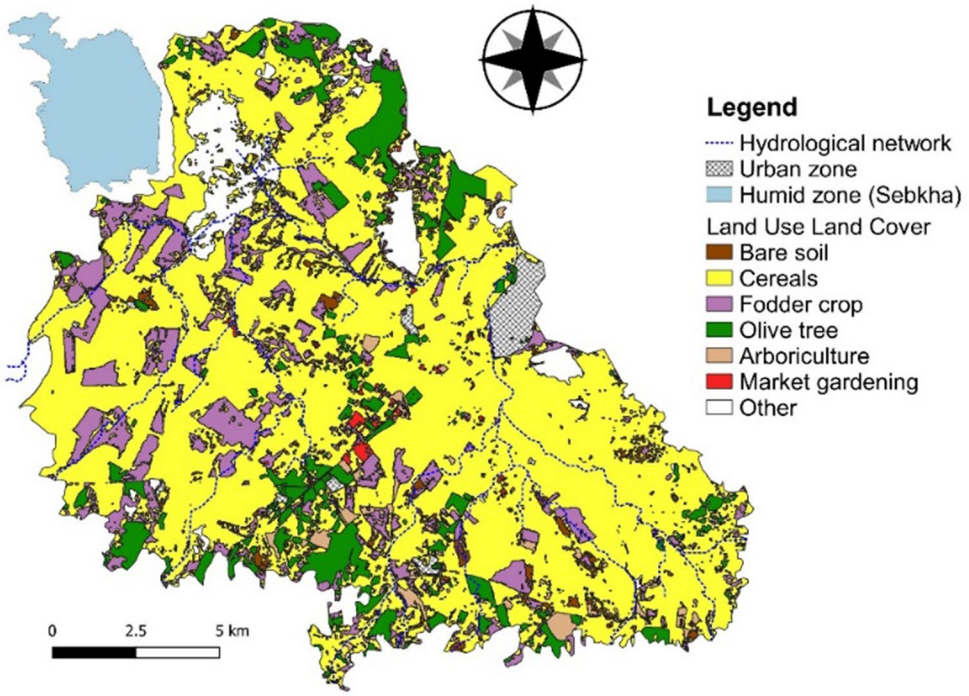



**Fig. 1** **a** Location of Tunisia, **b** El Fahs plain and **c** hydrological network and altitude of El Fahs plain


570000.010E
580000.010E
590000.010E
4030000.010N
4020000.010N
Legend
Urban zone
Wetland
Hydrological network
Geological outcrops
Quaternary
Neogene
Miocene
Oligocene
Eocene
Cretaceous
Jurassic
Triassic
Coordinate system: Carthage / UTM Zone 32
0 2.5 5 km

(a)


Legend
Hydrological network
Urban zone
Humid zone (Sebkha)
Land Use Land Cover
Bare soil
Cereals
Fodder crop
Olive tree
Arboriculture
Market gardening
Other
0 2.5 5 km

(b)

◂**Fig. 2** **a** Geological map of El Fahs plain (from 1:500 000 scale) (Ali et al. 1985), overlayed with the location of the cross-sections and **b** land-use map of El Fahs plain aquifer

stemming from irrigation return flow. Multivariate statistical analyses have been also used by Hajji et al. (2021) to decipher the role of natural and anthropogenic activities on the groundwater quality of Sfax coastal aquifer, whereas deterministic interpolation methods were used to map the spatial extent of the seawater intrusion. Their results showed that the groundwater processes are controlled either by rock-water interactions or agricultural and domestic activities, whereas spatial mapping of the seawater intrusion indicated high salinity hazards for local communities.

The objective of this work is to assess the hydrochemical status of El Fahs aquifer via the integration of different multivariate statistical techniques. To the authors' knowledge, this study is the first attempt to integrate graphical tools, multivariate statistical techniques and geostatistical interpolation methods to classify the quality status of El Fahs aquifer in the presence of limited amount of groundwater samples. In addition, the spatial variability of the uncertainty levels of the classification results is quantified, providing insights for future research and assisting the regional water authorities in the design of cost-effective monitoring practices. PCA is used to reduce the dimensionality of the sampling dataset prior performing the *K*-means clustering algorithm, used to partition the dataset into similar hydrochemical groups. Additionally, a geostatistical interpolation method that is often used in groundwater applications, namely indicator kriging (Panagiotou et al. 2022; Bradaï et al. 2016; Makkawi 2014) is used to estimate the probability of occurrence of the water classes beyond the sampling locations.

## Description of the study area

### General settings

The El Fahs aquifer is located in the mean valley of Miliane watershed in Zaghouan governorate and at a distance of 60 km in the south of Tunis (Fig. 1). The aquifer is located in El Fahs plain, which covers an area of 619 km$^2$. It is geographically bordered by geomorphological features formed by Beni Kleb and Rouissat mountains as well as the wetland of Korzia (Sebkha) in the North, by Bouarada plain and Mansour mountain in the west, by El Kef Lazreg and Bent Saidane mountains in the South, and by El Fahs hills in the East.

The plain is characterized by a semi-arid climate with an average annual precipitation of 325 mm and an average annual evapotranspiration of 1450 mm (INM 2022), which highlights the hydrological deficit of the area. The landscape is shaped by a well-developed hydrological network mainly formed by Miliane wadi and its tributaries, wadi Kebir, wadi Jarabiaa and wadi Boudhebbane (Fig. 1). The El Fahs aquifer plays a vital role in the region's water supply and socio-economic development. It serves as a critical water source for agriculture, supporting the irrigation needs of the fertile plains in the surrounding regions. The aquifer's water is used to cultivate various crops, including cereals, vegetables, and fruits, contributing to the country's region productivity and food security.

### Geology and hydrogeology

The study area belongs to the Tunisian Atlas, which is characterized by E–W and NW–SE grabens directed by a series of NE–SW and E–W faults and the apparition of Triassic outcrops (Belguith et al. 2011; Hachani et al. 2020). The geological outcrops extend from the Triassic to the Quaternary (Fig. 2). Triassic formations are formed by sedimentary rocks (e.g., gypsum, limestone and dolomite) and they are outcropping NW of El Fahs city. It's characterized by an irregular stratigraphy due to the presence of salt diapirs. Jurassic formations are composed of limestone and marls, and they are located in the southwestern part of the plain in Bent Saidane mountain. The cretaceous outcrops are mainly surrounding the plain, and they are formed by limestone, marly limestone, and marls. Marly and calcareous deposits of Paleocene and Eocene are found in the southwestern part of the plain near Mansour Mountain. The Oligocene outcrops which are formed by limestone and alternation of marls and sandstone beds are mainly located in the west and south of the plain. The Miocene and Pliocene which are mainly composed of continental sediments of clays, sandstone and conglomerate are outcropping in the north and south-eastern part of the plain. The central part of the plain is dominantly formed by Quaternary deposits which are mainly formed by red silt, limestone crusts and alluvia neighboring the wadis.

Considering the subsurface geology, previous investigations reveal formations ranging from marl-limestone to Upper Cretaceous marls, including limestone and clayey limestone near El Fahs Bridge. Quaternary formations, such as alluvium and terraces with pebble components, dominate the plain, with scree slopes at mountain bases. Triassic outcrops are visible at the northeastern (Mejri et al. 2018; Hachani et al. 2020; Ferjani et al. 2020).

From a hydrogeological viewpoint, El Fahs plain hosts an unconfined aquifer with around 60 m thickness, composed by porous Mio-Plio-Quaternary deposits (Fig. 3). The piezometric map shows a significant influence of topographic conditions on the groundwater flow in the water table of the phreatic aquifer (Fig. 4). The flow directions are mainly

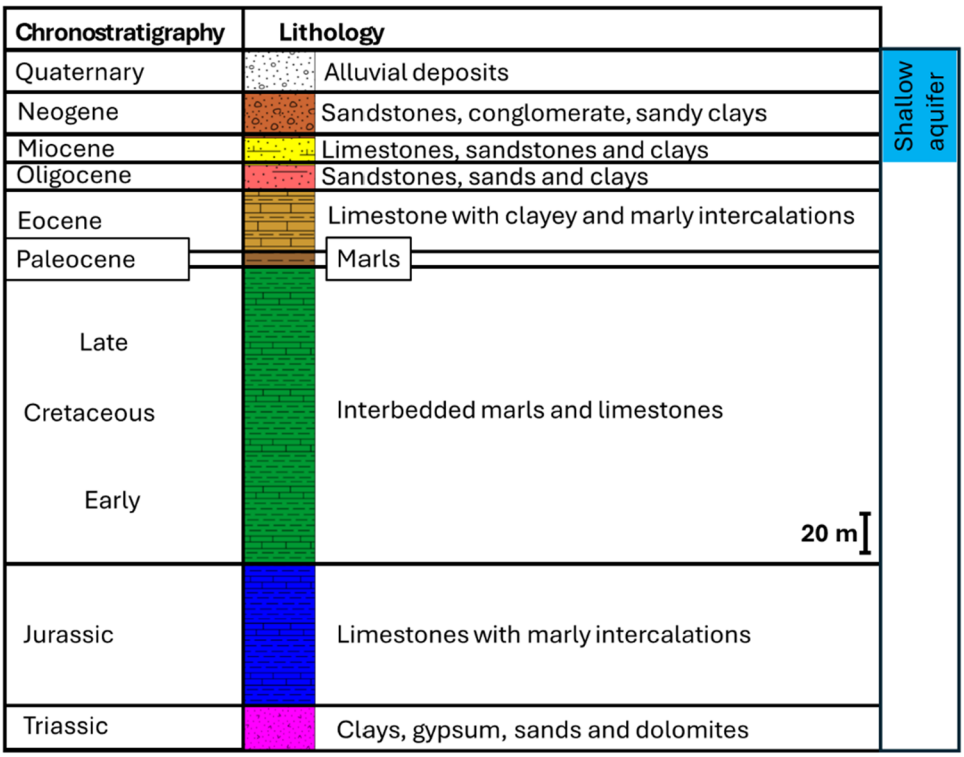



**Fig. 3** Synthetic hydro-stratigraphic log (after Bajanik et al. (1977))

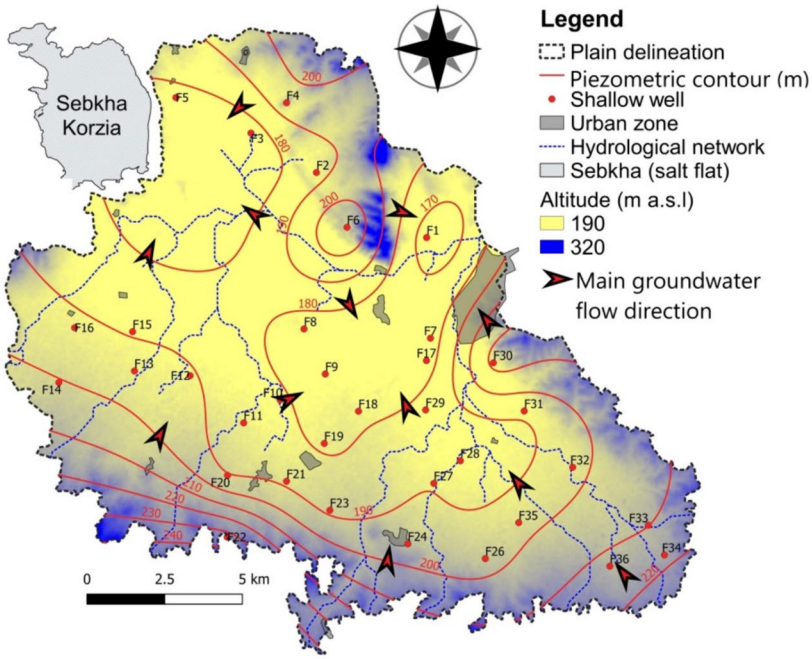



**Fig. 4** Piezometric map of El Fahs aquifer in April 2016

converging toward the center of the plain and Sebkha Korzia, which are considered as the aquifer outlets and where the piezometric contour is around 180 m. On the other hand, the highlands surrounding the plain, with a piezometric contour raging between 240 and 200 m, are the main recharge zones of the aquifer.

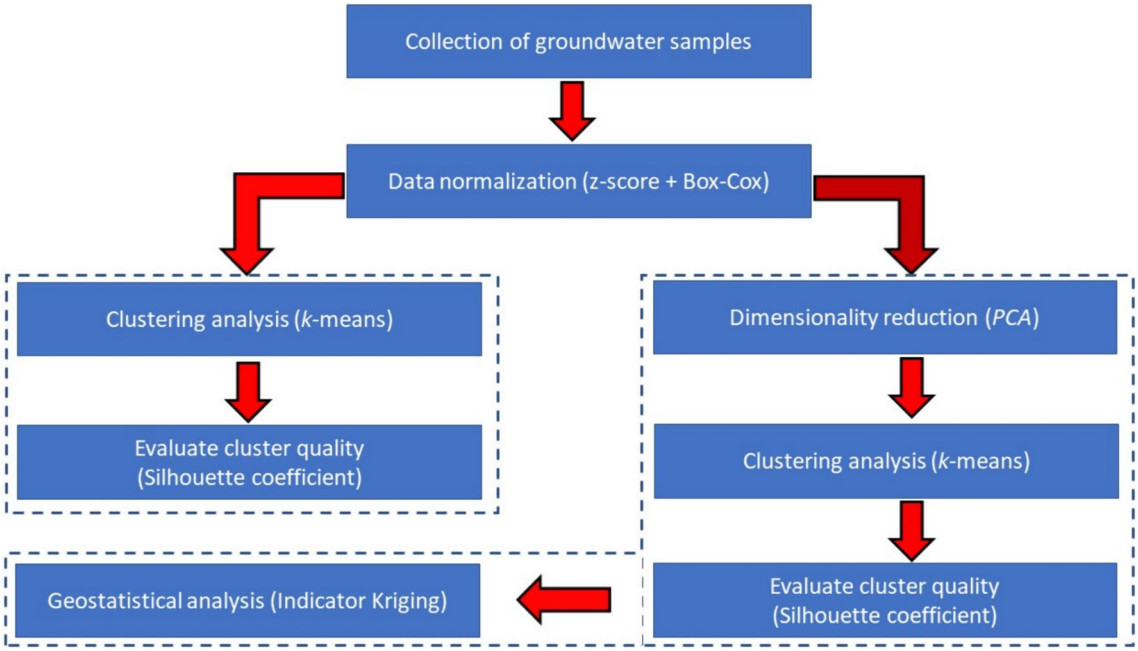



**Fig. 5** Flow chart of the proposed methodology

**Table 1** Descriptive statistics of the eight groundwater variables of the sampling dataset, and sub-samples belonging to the three water classes obtained via PCA + *K*-means clustering analysis

| Parameters | Unit | All samples ($n=36$) | | | Class 1 ($n=10$) | | | Class 2 ($n=13$) | | | Class 3 ($n=13$) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Min* | Median | Max* | Min* | Median | Max* | Min* | Median | Max* | Min* | Median | Max* |
| EC | µS/cm | 1.94 | 4.22 | 9.80 | 1.94 | 2.47 | 4.14 | 2.65 | 4.21 | 9.80 | 2.81 | 5.23 | 6.88 |
| $Ca^{2+}$ | mg/L | 28.1 | 208 | 432 | 28.1 | 136 | 321 | 76.2 | 321 | 433 | 40.1 | 120 | 369 |
| $Mg^{2+}$ | mg/L | 139 | 764 | 1709 | 139 | 635 | 830 | 679 | 802 | 1116 | 321 | 790 | 1710 |
| $Na^{+}$ | mg/L | 220 | 880 | 2535 | 220 | 819 | 2535 | 257 | 521 | 1818 | 401 | 950 | 1403 |
| $K^{+}$ | mg/L | 2.33 | 6.97 | 213 | 5.25 | 16.4 | 213 | 2.33 | 6.22 | 30.6 | 2.33 | 7.54 | 150 |
| $HCO_3^-$ | mg/L | 115 | 272 | 500 | 115 | 240 | 300 | 195 | 250 | 315 | 235 | 325 | 500 |
| $Cl^-$ | mg/L | 355 | 1118 | 4579 | 355 | 781 | 1207 | 781 | 1136 | 4579 | 816 | 1242 | 1917 |
| $SO_4^{2-}$ | mg/L | 236 | 763 | 1897 | 236 | 473 | 631 | 468 | 712 | 1109 | 1110 | 1336 | 1897 |

**Min:* minimum, *Max:* maximum

## Sample collection and analytical procedures

A field survey was carried out in El Fahs plain including in-situ measurement of physicochemical parameters, as well as surface water and groundwater sampling. All analyses were carried out at the Laboratory of Georesources at the Water Research and Technologies Center in Borj Cedria Technopark, Tunisia. Field measurements were carried out using portable instruments to assess parameters such as pH, temperature and electrical conductivity (EC). These measurements provided valuable insights into the water's overall condition and helped identify any potential anomalies that may require further investigation. A total of 36 groundwater samples were taken from shallow monitoring wells, ranging from 20 to 25 m below the ground surface, mainly used for irrigation water supply. Geographic coordinates were recorded using a handheld GPS unit (Garmin ETrex 32X). The representativeness of the taken samples was ensured since their collection was only performed after sufficient pumping time and stabilization of the EC. After that, the filled polyethylene bottles with groundwater samples were stored in an esky containing ice packs and transported to the laboratory of Georesources in the Borj Cedria Technopark in Tunisia where they were refrigerated at 4 °C until analysis.

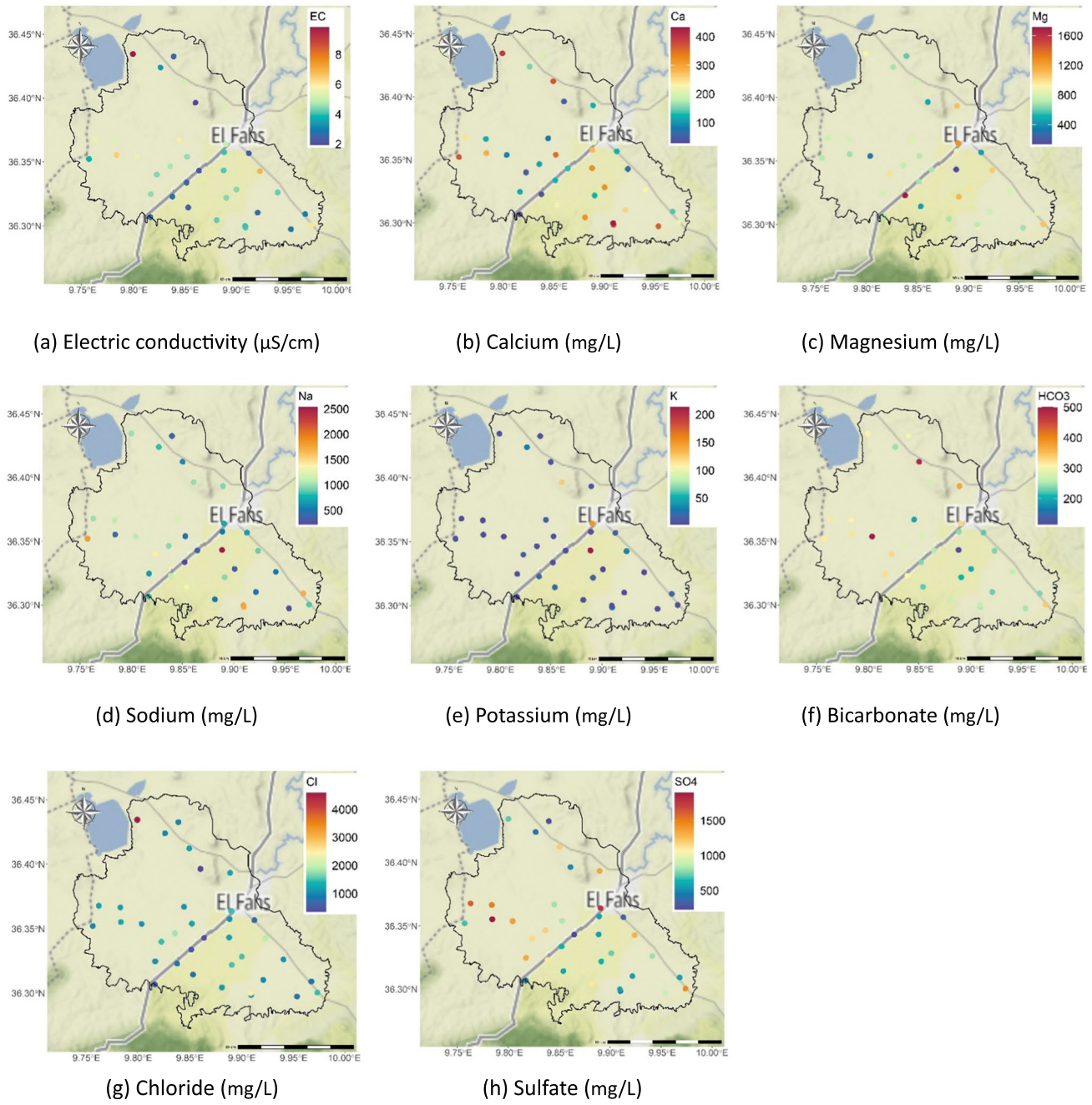



**Fig. 6** Spatial distribution of the groundwater parameters at the sampling locations

Bicarbonate ($HCO_3^-$) concentrations were measured using titration method with sulfuric acid. Major anion ($Cl^-, SO_4^{2-}$) and cations ($Na^+, Ca^{2+}, Mg^{2+}, K^+$) concentrations were measured via atomic absorption spectrometry (PerkinElmer Analyst 200) and ion liquid chromatography (IC 732) at the Georesources Lab of the Water Research and Technologies Center (CERTE, Tunisia).

## Methodology

An overview of the statistical methodology adopted in this study is given in Fig. 5. The R-4.3.2 open-source software, together with R-Studio 2023.09.1 integrated development software (IDE), were used for performing all statistical and graphical computations.

The original dataset can be expressed as:

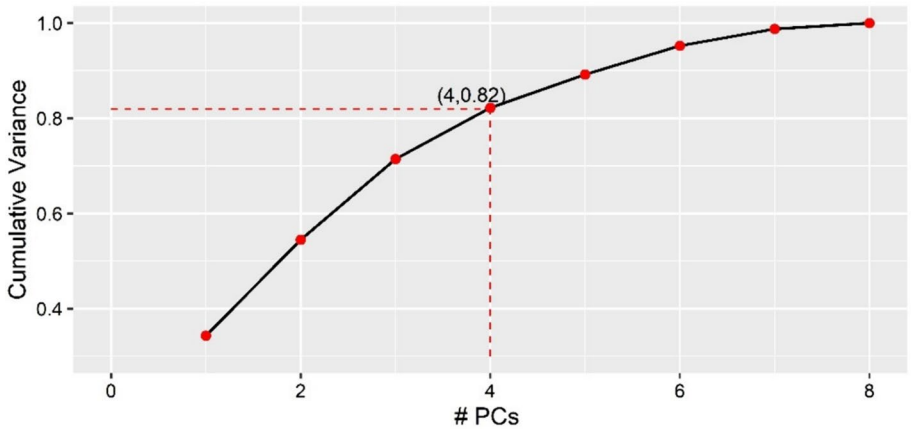



**Fig. 7** Cumulative variance as a function of the number of principal components

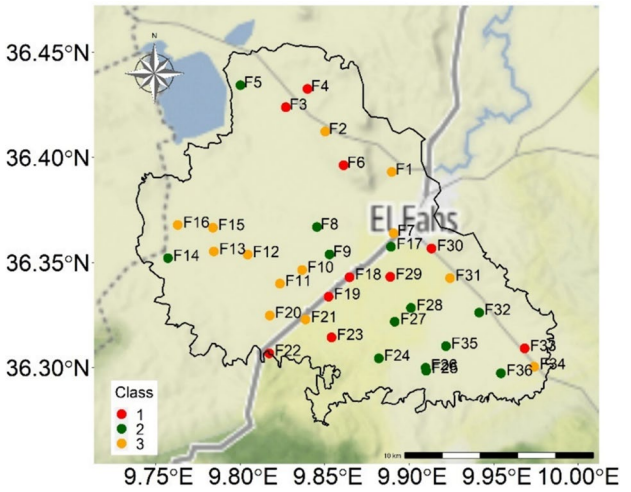



**Fig. 8** Spatial distribution of water classes at well locations via *K*-means clustering

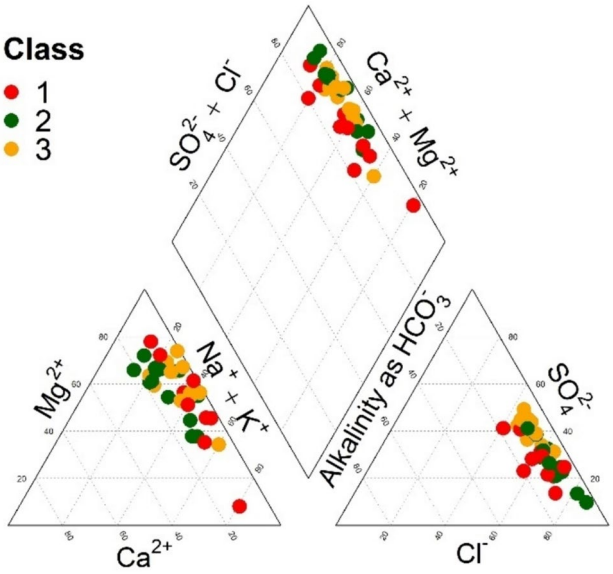



**Fig. 9** Piper diagram showing the relative percentages of major ions for each class

$$\boldsymbol{x}(\mathbf{u}_\alpha) = \{x_1, x_2, \ldots, x_p\}_a,\ \alpha = 1, \ldots, n, \tag{1}$$

where $\mathbf{u}_\alpha$ denotes the spatial coordinates of the $\alpha$-th sampling location, $n$ denotes the total number of sampling points and $p$ denotes the total number of hydrochemical parameters. To contribute equally to the clustering analysis, all variables were transformed to their $z$-scores so that their mean value becomes zero and their standard deviation becomes equal to one:

$$\boldsymbol{x}^z(\mathbf{u}_\alpha) = \left\{ \frac{x_1 - \mu_{x_1}}{\sigma_{x_1}}, \frac{x_2 - \mu_{x_2}}{\sigma_{x_2}}, \ldots, \frac{x_p - \mu_{x_p}}{\sigma_{x_p}} \right\}_a,\ \alpha = 1, \ldots, n, \tag{2}$$

where the superscript $^z$ denotes $z$-score variables, whereas $\mu$ and $\sigma$ denote the mean and standard deviation of a hydrochemical parameter respectively. A Box-Cox transformation is then applied to mitigate the impact of the outliers and skewness during the clustering process:

$$\boldsymbol{y}(\mathbf{u}_\alpha) = \left\{ \frac{(x_1^z)^{\lambda_1} - 1}{\lambda_1}, \frac{(x_2^z)^{\lambda_2} - 1}{\lambda_2}, \ldots, \frac{(x_p^z)^{\lambda_p} - 1}{\lambda_p} \right\}_a,\ \alpha = 1, \ldots, n, \tag{3}$$

where $\lambda$ denotes the power coefficient, estimated using the profile likelihood function via goodness-of-fit computations (Box and Cox 1964).

The next step is to decorrelate the Box-Cox variables via PCA, which is necessary to determine the number of major principal components that will be selected to reduce the dimensionality of the dataset. This is done by multiplying them with eigenvector matrix $\mathbf{V}$:

**Table 2** Average percentages values of the ions and cations which are displayed in Piper diagram

| Class | $Ca^{2+}$ (%) | $Mg^{2+}$ (%) | $Na^{+}$ (%) | $K^{+}$ (%) | $Cl^{-}$ (%) | $SO_4^{2-}$ (%) | $HO_3^{-}$ (%) |
|---|---|---|---|---|---|---|---|
| 1 | 7.7 | 51.9 | 39.4 | 1.1 | 60.4 | 27.5 | 12.1 |
| 2 | 12.7 | 58 | 29.2 | 0.2 | 66.0 | 26.8 | 7.3 |
| 3 | 6.9 | 60.3 | 32.4 | 0.4 | 51.7 | 40.4 | 7.9 |

**Table 3** Summary of fitting values for indicator variogram model of each water class

| Class | Variogram model | Nugget | Sill | Range (m) |
|---|---|---|---|---|
| 1 | Spherical | 0.0 | 0.21 | 2880 |
| 2 | Exponential | 0.0 | 0.29 | 2576 |
| 3 | Spherical | 0.0 | 0.26 | 4442 |

$$q(\boldsymbol{u}_a) = \left\{ \sum_{I=1}^{P} y_i(\boldsymbol{u}_a) V_{i,1}, \ldots, \sum_{I=1}^{P} y_i(\boldsymbol{u}_a) V_{i,p} \right\}_a, \; a = 1, \ldots n. \tag{4}$$

Subsequently, a subset of the decorrelated dataset will be chosen as input to the *K*-means clustering algorithm for the classification of the water samples.

## *K*-means clustering algorithm

*K*-means is one of the most commonly used clustering algorithms (Mohammadrezapour et al. 2020), first introduced in 1967 by MacQueen (1967), to partition a dataset into distinct populations (hydrochemical groups in this study). The assumptions of clustering algorithms, including *K*-means, include homoscedasticity (equal variance) and normal distribution of the variables (Alther 1979). First, the dataset is randomly distributed into a pre-selected number of groups (clusters). Several metrics are commonly used to determine the optimal number of clusters in terms of cluster quality (Charrad et al. 2014), together with experts' opinions, in terms of hydrochemical considerations. Then, each each data point iteratively moves among these clusters aiming at: (a) minimizing the variability within the clusters and (b) maximizing the variability among the clusters (Mohammadrezapour et al. 2020). This is done through an iterative process that aims to minimize the following objective function:

$$J(Q;V) = \sum_{i=1}^{c} \sum_{k \in A_i} \left|\left| \boldsymbol{q}^{(k)} - \boldsymbol{v}(\boldsymbol{A_i}) \right|\right|^2, \tag{5}$$

where $c$ is the total number of clusters, $A_i$ is the set of data points belonging to the $i$-th cluster and $\boldsymbol{v}(\boldsymbol{A_i})$ denotes the average coordinates of the the $i$-th cluster (cluster centroid):

$$\boldsymbol{v}(\boldsymbol{A_i}) = \left\{ \frac{1}{N} \sum_{j=1}^{N} q_{1,j}, \frac{1}{N} \sum_{j=1}^{N} q_{2,j}, , \ldots, \frac{1}{N} \sum_{j=1}^{N} q_{P,j}, \right\}_i, \; i = 1, \ldots c, \tag{6}$$

where $N$ is the number of data points belonging to a cluster. Euclidean metric is used to estimate the distance (dissimilarity), denoted by$\|.\|$, between pairs of data points within the variable space. The entire process is repeated multiple times (25 in the present study), each time using randomly selected initial cluster centers to reduce the dependence of the final solution to the initial conditions.

Several metrics can be used to assess the quality of the resulting clusters in terms of their cohesiveness and separation. The Silhouette coefficient (Rousseeuw 1987), ranging

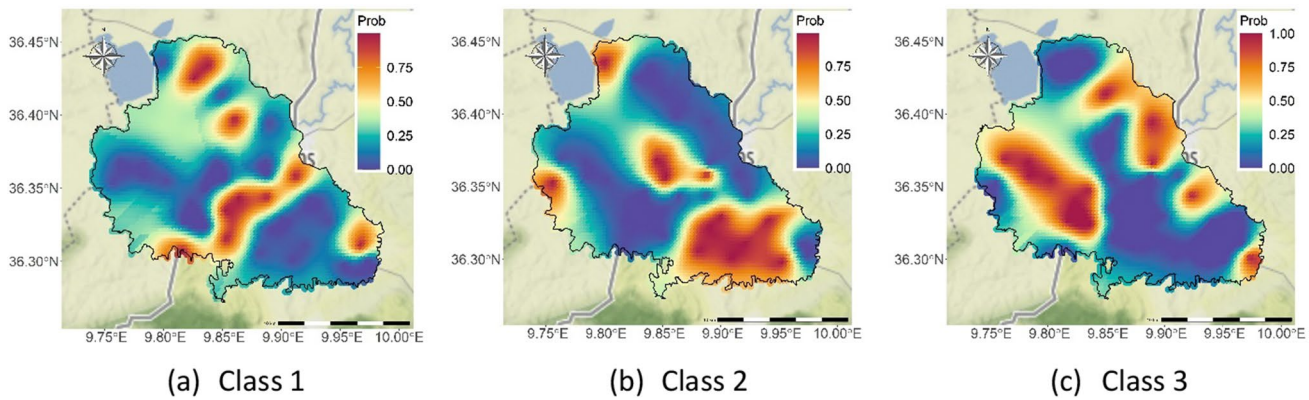



**Fig. 10** Spatial distribution of the probability values for each water class via indicator kriging

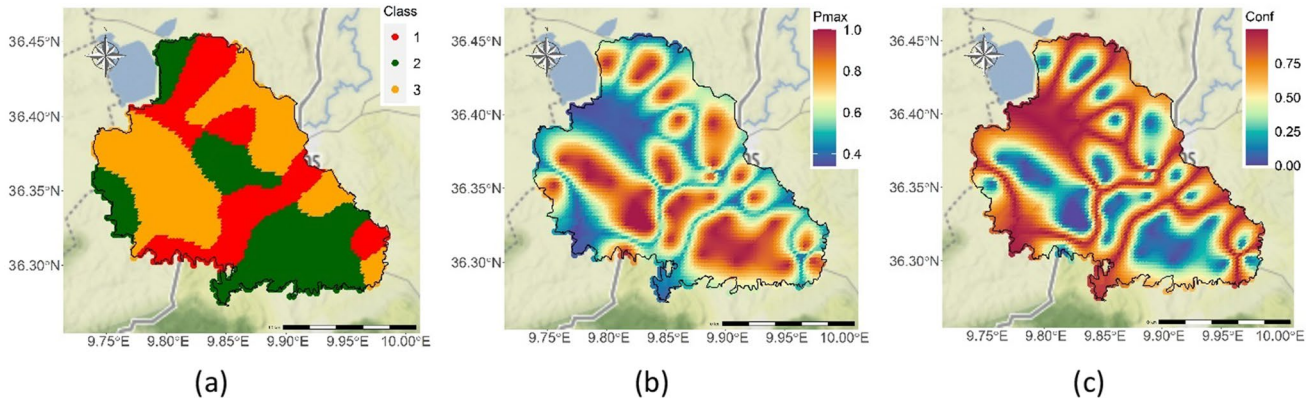



**Fig. 11** Spatial distribution of **a)** the most probable water classes, **b)** maximum probability among water classes, and **c)** difference between two most probable classes with the use of indicator kriging

from −1 (bad quality) to 1 (perfect quality), is a widely used method to evaluate the degree of cohesion and separation of the partitions, hence it is adopted in the present study. In brief, for each data point, the average of the distances to all other data points that belong in the same cluster is calculated and stored, along with the minimum distance. These two values are used to estimate the Silhouette coefficient. The average of all the Silhouette coefficients is then calculated to evaluate the quality of the resulting clusters.

## Indicator kriging

The next step is to provide the basic equations for a kriging variant, namely indicator kriging (IK), to identify the regions of high probability of occurrence of the water classes. IK is a non-parametric kriging variant since it does not aim at predicting the actual attribute values but rather convert these into binary data. The transformed dataset is then used to estimate the cumulative distribution function of an attribute at unknown locations, conditioned to the attribute values of the neighboring sampling points. First, the indicator variable, $I$, is defined via the following equation:

$$I\left(x_i;z_m\right) = \begin{cases} 1, & Z\left(x_i\right) \le z_k, \\ 0, & \text{otherwise} \end{cases} \quad m = 1, \ldots, M, \tag{7}$$

where $M$ refers to the total number of cut-offs and $z_m$ denotes the $m$-th cut-off value. The spatial association of the indicator variables for each threshold value is quantified via empirical semi-variograms, defined as:

$$\gamma_I(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left[I\left(x_i;z_k\right) - I\left(x_i + h;z_k\right)\right]^2, \tag{8}$$

where $h$ denotes the lag distance. To ensure the spatial continuity of the indicator variable, semi-definite functions are used to model the empirical semi-variograms, which are prerequisite of the kriging computations. Two widely used variogram models, are considered in the present study: the exponential:

$$\begin{matrix} s\left[1 - exp\left(\frac{h}{r}\right)\right] + n, & h \le r, \\ s + n, & h > r, \end{matrix} \tag{9}$$

and the spherical one:

$$\begin{matrix} s\left[\frac{3h}{2r} - \frac{1}{2}\left(\frac{h}{r}\right)^3\right] + n, & h \le r, \\ s + n, & h > r, \end{matrix} \tag{10}$$

where $s$, $n$ and $r$ denote the sill, nugget and range of the variogram model respectively. The kriging estimator of the indicator variable at a location $x_0$ is expressed as a linear combination of the known indicator values $I\left(x_i;z_k\right)$ at $n$ neighboring sampling locations $x_i$:

$$Z_{\mathrm{IK}}\left(x_o;z_k\right) = \sum_{i=1}^{n} \lambda_{i,\mathrm{IK}} I\left(x_i;z_k\right), \ \sum_{i=1}^{n} \lambda_{i,\mathrm{IK}} = 1, \tag{11}$$

where $\lambda_{i,\mathrm{IK}}$ denotes the weighting coefficient assigned to $I\left(x_i;z_k\right)$, summing to one to ensure unbiasedness of the estimated value. More details regarding the calculation of these weights and the kriging equations can be found in relevant works (Delfiner and Chilès 2012; Issaks and Srivastava 1989; Delbari et al. 2016; Goovaerts 1997). To represent probabilities, the kriging estimations of the water classes are normalized to ensure that their sum equals to one at each grid cell.

## Results and discussion

### Descriptive statistics

Table 1 shows the minimum, maximum and median values for all groundwater quality parameters considered in this study, whereas Fig. 6 shows their spatial distribution at the sampling locations.

The lowest EC values are measured along a narrow region that crosses the central part of the study area where a major river network is present (Fig. 1), whereas an isolated hot spot is observed close to the lagoon located at the northern part of the study area. A strong spatial heterogeneity of $Cl^-$ concentration values within the entire study area is observed, with the maximum value being present close to the lagoon, whereas the majority of the high values are located at the central part of the domain.

The central-western part of the case site is characterized by intermediate values of $Mg^{2+}$, except from an isolated hot spot close to the western border. An isolated hot spot is observed at the central region for $Na^+$ and $K^+$, where both parameters exhibit strong heterogeneity within the study area. Regarding $HCO_3^-$, high values are observed at the eastern and western parts of the central region due to the presence of Triassic formations composed of evaporites, whereas the majority of the north-western part is associated with high $SO_4^{2-}$ levels.

### Classification of the groundwater samples at sampling locations via *K*-means

Next, the sampling locations are partitioned into discrete classes based on their (dis)similarity in terms of the hydrochemical dataset.

For all computations, parameters are standardized to their z-scores (Eq. 2), and then normalized via Box-Cox transformation (Eq. 3) prior entering the clustering process. Twenty-seven quality indices are used to determine the number of clusters with the use of *Nbclust* package from *R* programming language, suggesting an optimal value of 3. Therefore, the optimum number of classes were chosen to be 3 for all clustering computations.

Principal Component Analysis was also used to reduce the dimensionality of the transformed dataset due to the relatively small number of sampling locations compared to the number of variables. The results revealed that four major principal components contain 82% of the total variance of the sampling data (Fig. 7), which were used as input data to the *K*-means clustering process.

Figure 8 shows the spatial variability of the three classes at the well locations.

According to Table 1, Class 3 exhibits the highest median value for $HCO_3^-$ (325 mg/l). This class is dominant in the western part of the study area where calcareous formations are dominant near Mansour mountain are present (Fig. 3), which can be attributed to the dissolution of carbonates.

Class 1 is characterized by the lowest EC values in terms of median and maximum (Table 1). Most of the wells belonging to this class are clustered within a thin zone aligned along the SW–NE direction, located in the central part of the study area. This class is associated with the fresh groundwater component due to the presence of a dense river network mainly Kebir wadi that is contributing to aquifer recharge with low salinity water (Fig. 2).

Class 2 dominates almost all sampling points that are present in the southern regions of the study area, except for two isolated observations wells in the south-eastern parts where Classes 1 and 3 are predicted. According to Table 1, this class exhibits the highest concentration values in terms of Calcium and Chloride, along with the relatively high Potassium values with respect to standard values for agriculture use (Smith et al. 2015). Subsequently, it can be associated with the extensive use of fertilizers in agriculture areas (Fig. 2b).

Piper diagram is also used to display the chemical composition of each cluster in terms of major anions and cations (Fig. 9), whereas Table 2 shows the average percentage values of the ion species for each class (cluster centroids).

For all classes, it is observed that Magnesium and Chloride are the dominant cations and ions respectively, illustrated by the fact that Mg–Ca–Cl is the dominant type for most samples. Among the three classes, Class 3 exhibits the highest average values with respect to Magnesium (60%) among the three classes, whereas Class 2 exhibits the highest Chloride percentage values for Chloride (66%).

Piper diagram results, along with the spatial distribution of sample classes, confirm that the groundwater mineralization in El Fahs aquifer could be explained by a natural process coming from rock-water interaction, especially in the case of salinization (Class 3) which is related to the presence of evaporate deposits (gypsum and halite of the Triassic) in subsurface geological materials in the eastern and western parts of the plain. However, the mineralization of Class 2 samples could be explained by an anthropogenic source, such as the irrigation return flow, mainly in the south-eastern part of the plain where chloride and nitrate concentrations exhibit high concentrations.

*K*-means method is also applied to the original dataset to assess the impact of dimensionality reduction on the classification process. As before, the original dataset has been subjected to *z*-score and Box-Cox transformations prior to entering the clustering method. Compared to the previous results, significant discrepancies are observed

(see Fig. 12 in Appendix) in the southern part of the study area due to the notable presence of Class 3, an observation however that is not supported by the local hydrogeological information. Additionally, the Silhouette score for the clusters generated via the PCA-reduced dataset was found to be 0.23, revealing an improvement of the cluster quality compared to the non-reduced case ($S = 0.20$).

### Spatial distribution of the water classes beyond the sampling locations

IK is applied to identify regions of high probability of occurrence for the water classes based on the classification of the PCA-reduced dataset via *K*-means clustering. Table 3 provides details of the best-fit parameters of the indicator variogram model for each water class.

Different numbers of neighboring data points are considered during the interpolation process to decide the spatial extent of the search neighborhood. Comparing the leave-one-out cross-validation errors, a minimum of 10 and a maximum of 26 nearby samples are adopted for all classes. Figure 10 shows the spatial variability of the probabilities for each water class to occur beyond the sampling locations. Regions of high probability for Class 1 to occur are identified along a narrow region in the central part of the study area, which coincides with a major branch of the hydrological network. Additionally, a large-scale coherent region exists close to the northern border, whereas a small section is also present southwards. Class 2 is expected to be dominant in the study area's southern regions, whereas low probabilities are predicted in almost the entire northern part of the study area. Regarding Class 3, it is most likely to occur in the eastern and western regions of the domain, where evaporate geological formations (Triassic age) are present.

Figure 11 reveals the spatial pattern of the maximum probabilities' values, along with their confusion levels, allowing the identification of regions of high and low uncertainty for classes to occur. The level of confusion is estimated by:

$$\mathrm{Conf}\left(\boldsymbol{u}_i\right) = 1 - \Delta P_{\max}\left(\boldsymbol{u}_i\right), i = 1, \ldots, N_c, \tag{12}$$

where $N_c$ denotes the total number of grid cells and $\Delta P_{\max}$ denotes the difference between the probabilities of the two most probable classes at the location of the $i$-th grid cell, leading to almost identical spatial patterns with the maximum probabilities.

Classes 1 (C1) and 3 (C3) are more probable than Class 2 (C2) to occur in most of the northern region, whereas Class 1 is expected to be present in the eastern region, close to the lagoon. However, high levels of uncertainty are observed in a large portion of the northern-eastern area, attributed to the lack of sampling data in that region. Class 2 is expected to be dominant in the southern part, which is characterized by low levels of uncertainty, except for some sporadic regions in the eastern part where Classes 1 and 3 are most probable to occur.

These results are in accordance with the overall hydrogeological regime, as we expect the class associated with the fresh groundwater component (C1) to be located along major branches of the hydrological network. Also, the presence of Class 3 in the eastern and western parts is justified by the presence of high bicarbonate concentrations, attributed to the dissolution of carbonate due to the evaporate formations.

## Summary and conclusions

Characterizing and diagnosing the condition of groundwater environments is a very challenging task due to their complexity in terms of hydrogeology, geology and land-use practices. El Fahs plain aquifer lies in an economically and ecologically important area, suffering from overexploitation of wells that are not optimally managed. The present study is the first attempt to investigate the hydrochemical state of El Fahs aquifer via the integration of graphical methods, multivariate statistical techniques and geostatistical modeling. Groundwater samples were collected from thirty-six observation wells, and eight physicochemical properties were analyzed in a single campaign during April 2016. The groundwater data were subjected to $z$-score and Box-Cox transformation, whereas Principal Component Analysis was used to reduce the dimensionality of the dataset before applying the cluster computations. A popular clustering approach, called *K*-means, is then used to partition the observation wells into groups based on their hydrochemical (dis)similarities.

The clustering results are shown to be consistent with the land-use and hydrogeological historic data of the study area, revealing the dominant presence of Class 2 at the eastern and western parts of the domain, which is related to high $HCO_3^-$ concentrations due to the rock–water interaction and the resulting dissolution of the evaporitic formations. A Piper diagram is constructed to display the relative proportions of the ionic species for the three water classes, revealing that Mg–Ca–Cl is the dominant type for most of the samples. *K*-means computations are also applied to the non-reduced dataset, resulting in a deterioration of the cluster quality compared to the PCA-reduced case in terms of the Silhouette coefficient.

Indicator kriging is also used to estimate the probability of occurrence of the water classes beyond the sampling locations. Class 1 is expected to occur in regions where dense river networks are present, especially in the central part of

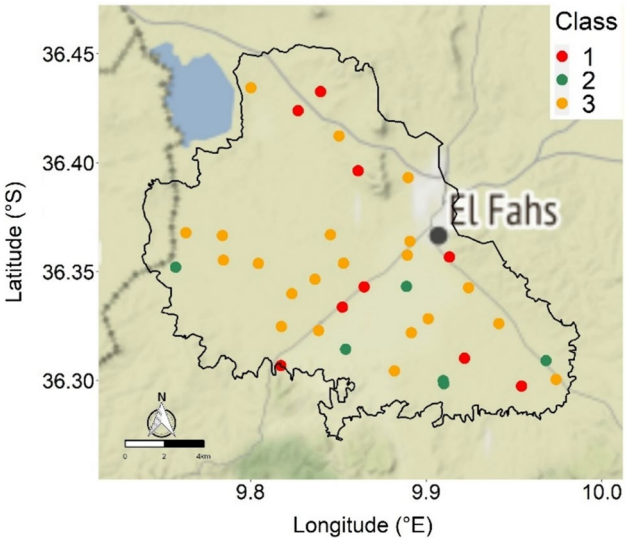



**Fig. 12** Spatial distribution of the water classes at the sampling locations via *K*-means with the use of the non-reduced dataset

the study area. Class 2 is expected to dominate the southern part, which is characterized by low confusion levels, whereas Class 1 and, especially, Class 3 are more probable to occur in the northern regions. However, a significant portion of the northern region is characterized by high levels of uncertainty due to the lack of a sampling network. Nevertheless, the ability of the integrated method to identify regions of high probability of occurrence for the water classes while quantifying the levels of uncertainty based on the spatial pattern of the hydrochemical parameters, could be used by the local policymakers to design/revise their water policies.

## Appendix

Figure 12 shows the spatial variability of the water classes generated by applying *K*-means on the complete dataset, mentioned in Sect. "Classification of the groundwater samples at sampling locations via K-means".

**Author contributions** Constantinos F. Panagiotou contributed to conceptualization, investigation, methodology, software, writing—review and editing, supervision. Anis Chkribene contributed to conceptualization, investigation, validation, writing—review and editing, supervision. Marinos Eliades contributed to validation, writing—review and editing. Christiana Papoutsa contributed to review. Evangelos Akylas contributed to review, writing—review and editing. Marinos Stylianou contributed to writing—review and editing. Nikos Stathopoulos contributed to writing—review and editing.

**Funding** This work was funded through the EXCELSIOR Teaming project (Grant Agreement No. 857510, www.excelsior2020.eu, accessed on 28 June 2024) that has received funding from the European Union's Horizon 2020 research and innovation programme and from the Government of the Republic of Cyprus through the Directorate General for the European Programmes, Coordination and Development.

**Availability of data and materials** Available upon request.

**Code availability** Available upon request.

## Declarations

**Conflict of interest** On behalf of all authors, the corresponding author states that there is no known conflict of interests, no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Ethical approval** We verify that the current manuscript presents part of the work done within the frame of EXCELSIOR project.

**Consent to participate and publish** We provide consent to participate and publish our work, which is jointly contributed by all authors.



## References

Abu-alnaeem A, Yusoff I, Ng T, Alias Y, Raksmey M (2018) Assessment of groundwater salinity and quality in Gaza coastal aquifer, Gaza Strip, Palestine: an integrated statistical, geostatistical and hydrogeochemical approaches study. Sci Total Environ 615:972–989

Acikel S, Ekmekci M (2018) Assessment of groundwater quality using multivariate statistical techniques in the Azmak Spring Zone, Mugla, Turkey. Environ Earth Sci 77(22):753

Adhikary P, Dash C, Chandrasekharan H, Rajput T, Dubey S (2011) Evaluation of groundwater quality for irrigation and drinking using GIS and geostatistics in a peri-urban area of Delhi, India. Arab J Geosci 5:1423–1434

Ali B, Jedoui Y, Dali T, Ben Salem H, Memmi L (1985) Geological map of Tunisia at the scale 1/500 000, Serv Géol, Tunisia

Alther G (1979) A simplified statistical sequence applied to routine water quality analysis: a case history. Ground Water 17(6):556–561

Aouiti S, HamzaouiAzaza F, El Melki F, Hamdi M, Celico F, Zammouri M (2021) Groundwater quality assessment for different uses using various water quality indices in semi-arid region of central Tunisia. Environ Sci Pollut Res 28:46669–46691

Bajanik S, Biely A, Mencik E, Salaj J, Stranik Z (1977) Notice explicative de la carte géologique 1/50.000 de Zaghouan. Service géologique de la Tunisie

Belguith Y, Geoffroy L, Rigane A, Gourmelen C, Dhia H (2011) Neogene extensional deformation and related stress regimes in central Tunisia. Tectonophysics 509(3–4):198–207

Benmarce K, Hadji R, Hamed Y, Zahri F, Zighmi K, Hamad A, Gentilucci M, Ncibi K, Besser H (2023) Hydrogeological and water quality analysis of thermal springs in the Guelma region of North-Eastern Algeria: a study using hydrochemical, statistical, and isotopic approaches. J Afr Earth Sci 205:105011

Box G, Cox D (1964) An analysis of transformations. J R Stat Soc B 26(2):211–252

Bradaï A, Douaoui A, Bettahar N, Yahiaoui I (2016) Improving the prediction accuracy of groundwater salinity mapping using indicator kriging method. J Irrig Drain Eng 142(7):04016023

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust: an R package for determining the relevant number of clusters in a data set. J Stat Softw 61(6):1–36

Chekirbane A, Gasmi O, Mlayah A, Gabtni H, Khadhar S, Lachaal F, Taupin JD (2022) Anthropogenic aquifer recharge effect on groundwater resources in an agricultural floodplain in northeastern Tunisia: insights from geochemical tracers and geophysical methods. Nat Resour Res 31:315–334

Collins W (1923) Graphic representation of water analyses. Ind Eng Chem Res 15(4):394

Deepika B, Ramakrishnaiah C, Naganna S (2020) Spatial variability of ground water quality: a case study of Udupi district, Karnataka State, India. J Earth Syst Sci 129:221

Delbari M, Amiri M, Motlagh M (2016) Assessing groundwater quality for irrigation using indicator kriging method. Appl Water Sci 6:371–381

Delfiner J, Chilès P (2012) Geostatistics: modeling spatial uncertainty. John Wiley & Sons Inc, New Jersey, Unites States

Drever J (1997) The geochemistry of natural waters, 3rd edn. Prentice-Hall, Upper Saddle River, New Jersey

Eliades M, Michaelides S, Evagorou E, Fotiou K, Fragkos K, Leventis G, Theocharidis C, Panagiotou C, Mavrovouniotis M, Neophytides S et al (2023) Earth observation in the EMMENA region: scoping review of current applications and knowledge gaps. Remote Sens 15:4202

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Güler C, Thyne G, McCray J, Turner A (2002) Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol J 10:455–474

Güler C, Kurt M, Alpaslan M, Akbulut C (2012) Assessment of the impact of anthropogenic activities on the groundwater hydrology and chemistry in Tarsus coastal plain (Mersin, SE Turkey) using fuzzy clustering, multivariate statistics and GIS techniques. J Hydrol 414–415:435–451

Hachani F, Balti H, Montassar M, Kadri A, Chkirbene A, Mlayah A, Gasmi M (2020) Contribution of geophysical methods in characterizing the structure of El Fahs plain: hydrogeological implications. J Afr Earth Sci 172:103984

Hajji S, Allouche N, Bouri S, Aljuaid AM, Hachicha W (2021) Assessment of seawater intrusion in coastal aquifers using multivariate statistical analyses and hydrochemical facies evolution-based model. Int J Environ Res Public Health 19(1):155

Ferjani H. A, Guellala R, Gannouni S, Inoubli M (2020) Enhanced characterization of water resource potential in Zaghouan region, Northeast Tunisia. Nat Resour Res 29:3253–3274

Hancock P, Boulton A, Humphreys W (2005) Aquifers and hyporheic zones: towards an ecological understanding of groundwater. Hydrogeol J 13:98–111

Hem J (1989) Study and interpretation of the chemical characteristics of natural water, vol 2254, 3rd edn. US geological survey water-supply, p 263

Houatmia F, Azouzi R, Charef A, Bédir M (2016) Assessment of groundwater quality for irrigation and drinking purposes and identification of hydrogeochemical mechanisms evolution in Northeastern, Tunisia. Environ Earth Sci 75(9):746

INM (2022) Climatic data of Zaghouan governorate. Institut National de la Météorologie, Tunis, Tunisia [online]

Issaks E, Srivastava R (1989) an introduction to applied geostatistics. Oxford University Press, Oxford, p 592

Javadi S, Hashemy S, Mohammadi K (2017) Classification of aquifer vulnerability using K-means cluster analysis. J Hydrol 549:27–37

M'nassri S, Dridi L, Schäfer G et al (2019) Groundwater salinity in a semi-arid region of central-eastern Tunisia: insights from multivariate statistical techniques and geostatistical modelling. Environ Earth Sci 78:288

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability

Makkawi M (2014) Geostatistics as a groundwater exploration planning tool: case of a brackish-saline aquifer. Arab J Geosci 8(5):3311–3319

Makni J, Bouri S, Ben Dhia H (2013) Hydrochemistry and geothermometry of thermal groundwater of southeastern Tunisia (Gabes region). Arab J Geosci 6:2673–2683

Masoud A (2014) Groundwater quality assessment of the shallow aquifers west of the Nile Delta (Egypt) using multivariate statistical and geostatistical techniques. J Afr Earth Sci 95:123–137

Mastrocicco M, Colombani N (2021) The issue of groundwater salinization in coastal areas of the Mediterranean region: a review. Water 13:90

Mejri S, Chekirbene A, Tsujimura M, Boughdiri M, Mlayah A (2018) Tracing groundwater salinization processes in an inland aquifer: a hydrogeochemical and isotopic approach in Sminja aquifer (Zaghouan, northeast of Tunisia). J Afr Earth Sci 147:511–522

Mohammadrezapour O, Kisi O, Pourahmad F (2020) Fuzzy *c*-means and *K*-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality. Neural Comput Appl 32:3763–3775

Moore W (1999) The subterranean estuary: a reaction zone of ground water and sea water. Mar Chem 65:111–125

Ncibi K, Mastrocicco M, Colombani N, Busico G, Hadji R, Hamed Y, Shuhab K (2022) Differentiating nitrate origins and fate in a semi-arid basin (Tunisia) via geostatistical analyses and groundwater modelling. Water 14:4124

Panagiotou C, Kyriakidis P, Tziritis E (2022) Application of geostatistical methods to groundwater salinization problems: a review. J Hydrol 615:128566

Panagiotou C, Stefan C, Papanastasiou P et al (2023) Quantitative microbial risk assessment (QMRA) for setting health-based performance targets during soil aquifer treatment. Environ Sci Pollut Res 30:14424–14438

Piper A (1944) A graphic procedure in the geochemical interpretation of water-analyses. Trans Am Geophys Union 25:914–923

Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65

Schoeller H (1962) Les eaux souterraines: Hydrologie dynamique et statique. Comptes rendus critiques. In: Hydrogéologie en chambre Paris, Masson, in-8°, Paris

Smith CJ, Oster JD, Sposito G (2015) Potassium and magnesium in irrigation water quality assessment. Agric Water Manag 157:59–64

Stiff HJ (1951) The interpretation of chemical water analysis by means of patterns. J Petrol Technol 3(10):15

Yimit H, Eziz M, Mamat M, Tohti G (2011) Variations in groundwater levels and salinity in the Ili river irrigation area, Xinjiang, northwest China: a geostatistical approach. Int J Sust Dev World Ecol 18:55–64