Seawater intrusion pattern recognition supported by unsupervised learning: A systematic review and application

Christian Narvaez-Montoya, Jürgen Mahlknecht, Juan Antonio Torres-Martínez, Abrahan Mora, Guillaume Bertrand

Science of the Total Environment 864 (2023) 160933

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Review

Seawater intrusion pattern recognition supported by unsupervised learning: A systematic review and application

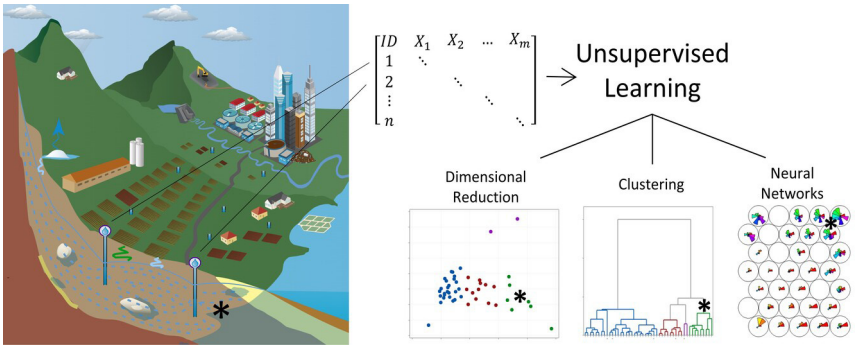
Christian Narvaez-Montoya [a], Jürgen Mahlknecht [a,*], Juan Antonio Torres-Martínez [a], Abrahan Mora [b], Guillaume Bertrand [c,d]

[a] Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Eugenio Garza Sada 2501, Monterrey 64849, Nuevo Leon, Mexico
[b] Tecnologico de Monterrey, Escuela de Ingenieria y Ciencias, Campus Puebla, Atlixcáyotl 5718, Reserva Territorial Atlixcáyotl, Puebla 72453, Mexico
[c] University of Bourgogne Franche-Comté, UMR UFC CNRS 6249 Chrono-Environnement, 16 route de Gray 25000 Besançon, 4 Place Tharradin, 25200 Montbéliard, France
[d] Federal University of Paraíba, Department of Civil and Environmental Engineering, João Pessoa 58051-900, Brazil

HIGHLIGHTS

- A database of hydrogeochemical studies in coastal aquifers was developed.
- The reviewed techniques were explained and applied to a practical case.
- R scripts for the reviewed techniques are presented.
- Sixty-two percent of the reviewed studies did not report raw data.
- TDS, EC, and TH are redundant variables if major ions are also used.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Damià Barceló

Keywords:
Coastal aquifers
Machine learning
Multivariate
Clustering
Salinization

ABSTRACT

Seawater intrusion is among the world's leading causes of groundwater contamination, as salty water can affect potable water access, food production, and ecosystem functions. To explore such contamination sources, multivariate analysis supported by unsupervised learning tools has been used for decades to aid in water resource pattern recognition, clustering, and water quality data variability characterization. This study proposes a systematic review of these techniques applied for supporting seawater intrusion identification based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and subsequent bibliometric analysis of 102 coastal hydrogeological studies. The most relevant identified methods, including principal components analysis (PCA), hierarchical clustering analysis, K-means clustering, and self-organizing maps, are explained and applied to a case study. Although 74 % of the studies that applied dimensional reduction methods, such as PCA, associated most of the database variance with the salinization process, 77 % of the studies that applied clustering methods associated at least one water sample cluster with the influence of seawater intrusion. Based on the review and a practical demonstration using the open-source R software platform, recommendations are made regarding data preprocessing, research opportunities, and publishing information necessary to replicate and validate the studies.

* Corresponding author.
E-mail address: jurgen@tec.mx (J. Mahlknecht).

http://dx.doi.org/10.1016/j.scitotenv.2022.160933
Received 25 September 2022; Received in revised form 10 December 2022; Accepted 11 December 2022
Available online 23 December 2022

Contents

1. Introduction

Anthropogenic activities and climate change have had significant negative impacts on the world's water resources over the last 200 years, and these effects are expected to intensify (Amanambu et al., 2020; Burrell et al., 2020; Ferguson and Gleeson, 2012). This situation is magnified in coastal plain areas, which are home of 70 % of the world's population (Alfarrah and Walraevens, 2018). Many of these areas are located in arid and semi-arid climates with insufficient surface water resources, leading to a critical dependence on groundwater (Mianabadi et al., 2020; Vaux, 2011). Coastal areas with Mediterranean and tropical climates tend to increase surface water and groundwater use to meet their needs (Busico et al., 2018; Yin et al., 2021). However, groundwater overexploitation reduces freshwater outflow to the sea and represents an additional adverse effect in many coastal zones. Such exploitation causes seawater to migrate towards fresh groundwater resources, and the resulting water mixture is extracted by production wells used for public water supply, irrigation, or industry (Alfarrah and Walraevens, 2018). This constitutes a severe threat to coastal water supply systems and is one of the leading causes of groundwater contamination (Michael et al., 2017; Polemio and Zuffianò, 2020; Tully et al., 2019).

Seawater intrusion has been identified in approximately 100 countries and regions around the world, and approximately 32 % of coastal metropolitan cities are estimated to be threatened by it (Cao et al., 2021). Salty water associated with this source and other salinization phenomena, such as irrigation return flow or sewage system leakages, cause health problems, such as diarrheal diseases and issues related to hypertension, like stroke, heart attack, and preeclampsia. Similarly, saline water leads to the deterioration of water quality for irrigation purposes (Damonte and Boelens, 2019; Naser et al., 2017; Rakib et al., 2020; Tully et al., 2019). Thus, monitoring and understanding the natural and human relations in groundwater systems is essential for developing appropriate and sustainable management strategies for coastal aquifers (Lall et al., 2020; Michael et al., 2017). However, investigating groundwater requires multidisciplinary approaches that incorporate environmental, geological, physical, and social aspects and analyses of normally limited physical and chemical information (Díaz-Alcaide and Martínez-Santos, 2019; Michael et al., 2017).

Analyses and models have been developed to determine the mechanisms underlying the seawater intrusion phenomenon (Cao et al., 2021; Enemark et al., 2019). However, hydrological processes are highly complex, dynamic, and non-linear at both spatial and temporal scales; therefore, local or regional studies are subject to great uncertainty (Kalteh et al., 2008; Rajabi et al., 2018). Consequently, one of the biggest challenges facing exploratory studies is to identify whether a sample of brackish water originates from seawater intrusion or another source of groundwater salinization, as shown in Fig. 1 (Abu-alnaeem et al., 2018; Mirzavand et al., 2020). From this perspective, authors have argued that a proper international tool platform based on hydrogeological data is urgently needed to verify the occurrence and influence of seawater intrusion (Cao et al., 2021).

In this context, machine learning techniques have benefitted various complex studies in hydrological research (Bertrand et al., 2022; Rajoub, 2020; Tahmasebi et al., 2020). While supervised techniques predict and optimize models based on known outputs, unsupervised techniques learn more about the internal dependencies among explanatory variables (Berry et al., 2020; Díaz-Alcaide and Martínez-Santos, 2019). Multivariate analysis supported by unsupervised machine learning tools has been used for decades to characterize the range and variability of environmental water tracers across broad temporal and spatial scales, and the number of studies using these tools is increasing exponentially (Gredilla et al., 2013; Sergeant et al., 2016). Although the unsupervised nature of these techniques does not enable direct identification of dynamic water processes, obtaining water quality patterns greatly supports the interpretation of various phenomena (Li et al., 2018; Wunderlin et al., 2001). Therefore, it is essential to review how these methods have been applied to hydrogeochemical data to identify patterns that help differentiate seawater intrusion from other sources. Such a review should permit the proposal for a typology of usable tools and their limitations based on the specific objectives and constraints of researchers.

This bibliometric review study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement (Moher et al., 2009), which allowed the review to be conducted more objectively. This systematic review was based on a flow chart, and different restrictions where applied to limit the document search, extraction, and analysis. First, we searched for papers published from 2000 to 2022 (22 years) that

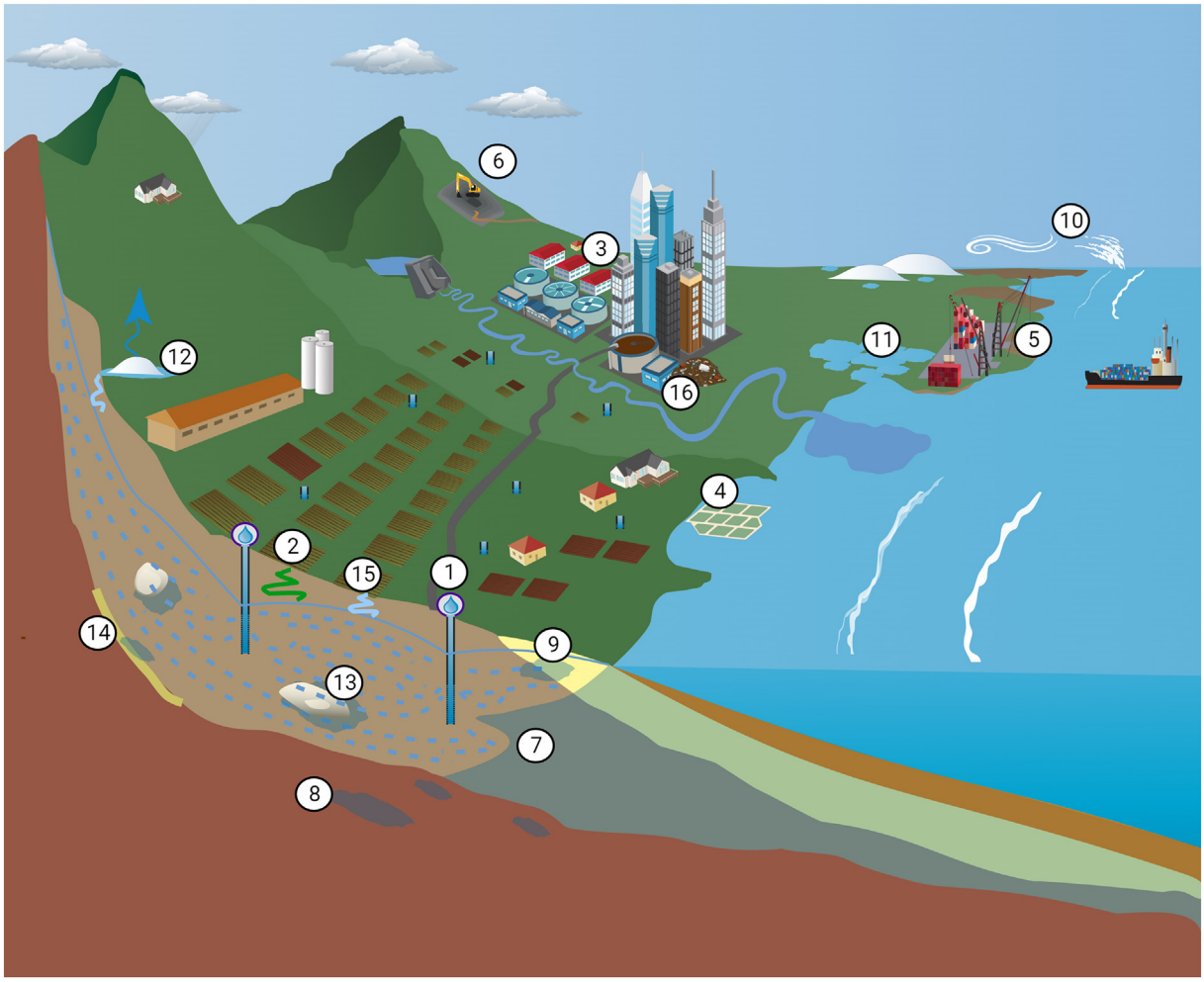


Fig. 1. Coastal groundwater squeeze. (1) Groundwater overexploitation; (2) agricultural contamination; (3) urban sprawl and development; (4) aquaculture contamination; (5) land reclamation dredging and navigation; (6) mining contamination; (7) seawater intrusion; (8) connate saline water; (9) marine transgression; (10) sea spray; (11) episodic flooding; (12) evaporative concentration; (13) rocks dissolution; (14) salt filtering by clay; (15) irrigation return flow; and (16) effluents and spills.

applied unsupervised learning tools to analyze hydrochemical data on coastal aquifers presumably affected by seawater intrusion. Subsequently, the articles were filtered using predetermined rules to eliminate incorrect selections. Finally, the selected articles were analyzed to assess how unsupervised learning tools have been applied to support seawater intrusion pattern recognition. The objective was to provide a reference for researchers and professionals who wish to apply these methodologies to identify hydrogeochemical processes in coastal aquifers. A database was established to generate a range of available similar studies to which a specific approach can be compared. Additionally, the most appropriate identified methods applied to the La Paz (Mexico) coastal aquifer case and the reviewed tools were delivered using an open-source data science software.

2. Background

2.1. Seawater intrusion identification

Seawater intrusion monitoring and assessment has been performed using four approaches: hydraulic head measurements, groundwater models, geophysical methods, and environmental tracer analysis (Cao et al., 2021; Werner et al., 2013). Hydraulic head measurements consider hydraulic gradient evaluations because landward gradients can reveal seawater intrusion susceptibility (Jasechko et al., 2020). However, hydraulic heads in the mixing zone are difficult to interpret owing to salinity; density variations occur at different elevations of an observed piezometer (Werner et al., 2013). In groundwater flow and transport models, multiple types of environmental data are integrated to numerically simulate variable density flow (Costall et al., 2020). Although such models are among the most complete approaches, the calibration stage can be tedious and time-consuming, and obtained results may be unsatisfactory (Carrera et al., 2010). Geophysical methods are used to map the subsurface groundwater salinity distribution in one, two, and three dimensions based on the differences in the electromagnetic properties of fresh and salty water (Werner et al., 2013). These methods also require calibration to differentiate the received signals from the lithological structure and salinity distribution (Cao et al., 2021).

Environmental tracers, such as major ions, have been broadly recognized as valuable tools for determining seawater intrusion (Li et al., 2020; Mirzavand et al., 2020). Compared with other approaches, tracers rely on multivariate analysis and are used to describe the groundwater system; moreover, hydrochemical data are easy to obtain owing to their low test costs and the high demand for exploration (Liu et al., 2021). Major ions are analyzed at different sample points based on multiple bivariates and composite plots, such as Piper, Stiff, and Gibbs diagrams (Mirzavand et al., 2020). These tracers have concentrations higher than 5 mg/L and account for over 95 % of the total solute content (Poeter et al., 2020). Generally, terrestrial groundwater tends to be of the alkaline earth bicarbonate type, and the concentration of Ca^{2+} often exceeds that of Mg^{2+}. Seawater has much higher concentrations of major ions (except for Ca^{2+} and HCO_3^-) and some minor ions than groundwater, and its water composition does not vary significantly at the global scale, because the long residence time in the ocean implies mixing and homogenization (Jiao and Post, 2019). The characteristics of groundwater from different coastal aquifers and standard

Table 1
Median values of the chemical composition of different coastal aquifer case studies and standard seawater. Values in brackets show the variation range.

Parameter	La Paz (Mexico) - Alluvium Arid (Tamez-Meléndez et al., 2016)	Göksu (Turkey) - Alluvium Mediterranean (Güner et al., 2021)	Jaffna (Siri Lanka) Karstic Monsoon (Chandrajith et al., 2016)	Shenzhen (China) Granite Monsoon (Shi et al., 2018)	Standard Seawater (Jiao and Post, 2019)
Chloride, Cl^- (mg/L)	385 (54.5–2960)	141.2 (72–1597.6)	250 (30–3500)	26.1 (4.1–3260)	19,804
Sodium, Na^+ (mg/L)	137 (36.7–1080)	122.3 (19.5–880.1)	130 (25–1500)	20.4 (4.2–1730)	11,033
Sulphate, SO_4^{2-} (mg/L)	64.3 (7.9–490)	193.6 (105.6–321.5)	49 (12–430)	29.5 (2.2–379.7)	2776
Magnesium, Mg^{2+} (mg/L)	39.5 (9.3–344)	31.1 (13.4–125.6)	24 (1.9–220)	4.4 (0.3–225)	1314
Calcium, Ca^{2+} (mg/L)	104 (27.8–658)	50.7 (15.8–136.4)	90 (7.6–660)	55.4 (1–214.8)	422
Potassium, K^+ (mg/L)	4.1 (1.5–14)	5.84 (2.4–34.5)	9.8 (0.9–44)	7.7 (0.7–61.8)	408
Bicarbonate, HCO_3^- (mg/L)	325 (166–1290)	246.1 (77.8–453.7)	250 (92–601)	165 (0.6–611.2)	107
Nitrate, NO_3^- (mg/L)	24.92 (0.7–216)	12.4 (12.1–13.3)	3.1 (1.8–26.1)	2.75 (0–26.5)	–
Total dissolved solids, TDS (mg/L)	1054 (348–5828)	554 (157–3941)	765 (221–6604)	279 (23–6760)	36,000
pH	7.2 (6.8–8.3)	7.8 (7.5–8.2)	7.59 (6.8–8.2)	7.4 (4.6–7.9)	8.1
T (°C)	29.7 (25–33.4)	20.9 (20.2–23.1)	30.5 (29.6–31.8)	23.4 (15.4–29.5)	–

seawater are listed in Table 1, and a Piper diagram of these data is shown in Fig. 2.

Tracer analysis encompasses many parameters in addition to the chemical composition, including biological, physical, and physicochemical parameters, such as temperature (T), pressure, density, electrical conductivity (EC), sampling position location, sampling date, sampling depth, and isotopic signatures, such as $\delta\ ^2H$ (deuterium) and $\delta^{18}O$ (oxygen-18) (Jiao and Post, 2019; Werner et al., 2013). Consequently, meaningful conclusions regarding water sources can only be made by combining the different parameter types. Exploring these multivariate data relationships is supported by the application of unsupervised learning techniques (Liu et al., 2021; Werner et al., 2013). For example, Table S1 shows multivariate data from the sampling campaign performed in August 2013 for the Mexican La Paz coastal aquifer (Section S1). The aquifer and database were analyzed by Tamez-Meléndez et al. (2016) and Torres-Martínez et al. (2021), who identified different salinization and nitrification sources supported by the multivariate analysis of hydrogeochemical and isotopic tracers.

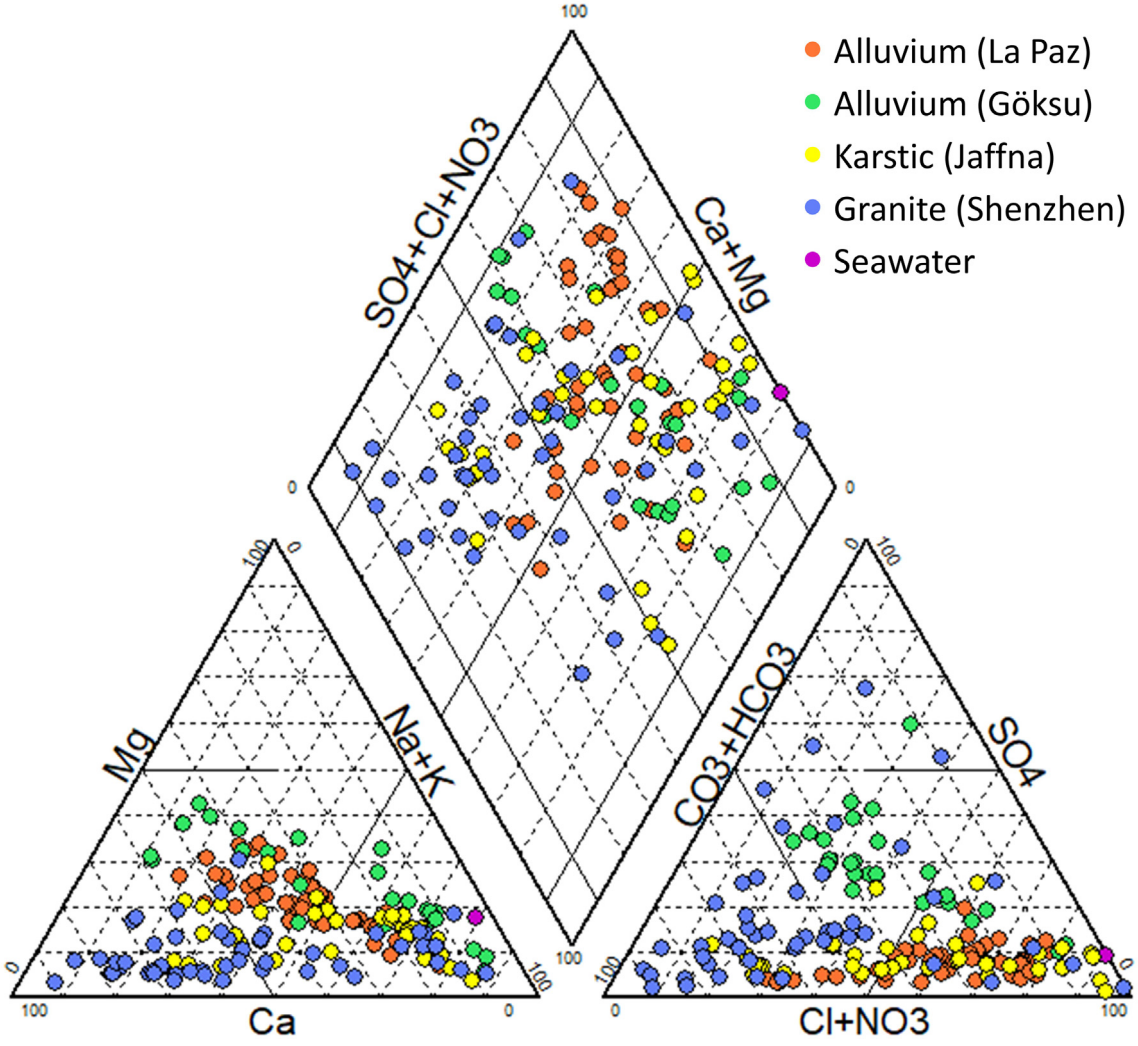


Fig. 2. Piper diagram for the different coastal aquifer case studies and seawater listed in Table 1.

2.2. Unsupervised pattern recognition in hydrogeology

The spatial variability of water samples can provide insights into aquifer heterogeneity and connectivity; thus, a robust classification and association scheme is important for the characterization of hydrogeological systems (Güler et al., 2002). Multivariate analysis of the physical, chemical, and biological characteristics of water resources supported by unsupervised learning tools has been used for decades to characterize data, select meaningful variables, and recognize pattern data structures and trends (Gredilla et al., 2013; Sergeant et al., 2016). Unlike supervised methods, which aim to predict variables and optimize parameters based on known outputs, unsupervised techniques aim to delineate the underlying internal relationships of the features (Díaz-Alcaide and Martínez-Santos, 2019; Berry et al., 2020). In the absence of certainty about "real outputs" from the provenance of water samples, unsupervised methods are preferred for the broad identification of the latent features, because the output is not restricted to a specific response (Tahmasebi et al., 2020).

Unsupervised pattern recognition techniques do not necessarily establish cause-and-effect relationships; they rather present information in a compact format as the first step in the complete analysis for generating hypotheses and performing hydrogeochemical data interpretation (Berry et al., 2020). Unsupervised techniques for pattern recognition have been commonly employed to analyze complex datasets and integrate environmental and pollution data (Fdez-Ortiz de Vallejuelo et al., 2011). These techniques apply mathematical methods to generate object data, graphical representations of the newly generated data, and interpretations of the resulting objects (Gredilla et al., 2013). These techniques can be divided into three main groups: (1) dimensional reduction methods (DRM), (2) cluster analysis (CA), and (3) artificial neural networks (ANN) (Fig. 3).

DRM aims to limit n-dimensional information about objects to a set of reduced and more representative dimensions (Ayesha et al., 2020). In this manner, each observation can be graphically depicted in 2D or 3D plots that show the most relevant relations in the database. Principal component analysis (PCA) is the most commonly used method, and a full review of the most relevant methods can be found in Ayesha et al. (2020). CA techniques are the most widely used pattern recognition methods, and their objective is to assign observations to the same cluster based on the degree of similarity among the variables (properties) that characterize the observation (Gredilla et al., 2013). A full review of traditional and recent developments in CA techniques can be found in Saxena et al. (2017). Finally, the use of ANN has been increasing, and this method simulates the nervous system in human beings to create models for pattern recognition (Gredilla et al., 2013). The most popular method for multivariate analysis is the self-organizing map (SOM), which performs segmentation similar to CA and allows for a topological representation of the database. A review of general ANN for pattern recognition can be found in Abiodun et al. (2019).

3. Material and methods

3.1. Focused question and search strategy

To delineate the type and methods of unsupervised learning techniques that have been used to identify seawater intrusion in coastal aquifers, this bibliometric review study follows the above-mentioned PRISMA statement (Moher et al., 2009). A systematic computerized literature search was conducted in May 2022 using Scopus and the Web of Science. The combination of three keywords ["seawater AND intrusion AND cluster", or "seawater AND intrusion AND unsupervised", or "seawater AND intrusion AND multivariate", or "saltwater AND intrusion AND cluster", or "saltwater AND intrusion AND unsupervised", or "saltwater AND intrusion AND multivariate"] was used to search for original articles released from 2000 to 2022, based on the search fields of "keyword", "abstract" and "title". After obtaining raw data, articles were excluded based on the following criteria: duplicate studies, no DOI, papers published before 2021 that did not have any citations, analyses that did not consider major ions, articles in which the search keywords did not match unsupervised machine learning or multivariate methodologies applied for water quality, analyses that were not applied to groundwater, and full-text articles that were not written in English, Spanish, Portuguese, or French.

3.2. Data extraction and analysis

The full text of the final papers selected for the analysis were assessed. General information on the article identification process, study case

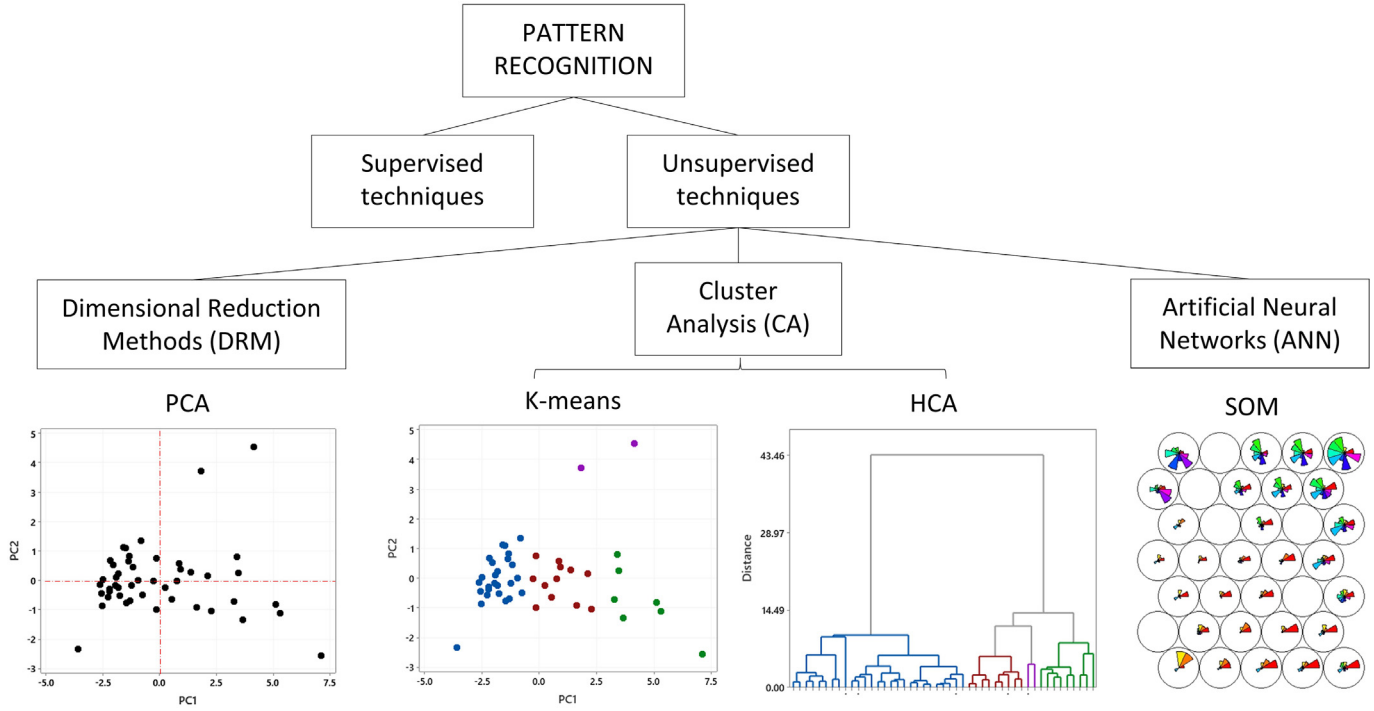


Fig. 3. Classification of pattern recognition techniques.

description, unsupervised learning application, and hydrogeochemical techniques was extracted. The data variables extracted for each section are shown in Table S2. If the articles did not specify the variable "climate", then this variable was included based on the location of the study area and the Köppen climate classification (Cui et al., 2021). Furthermore, if the article did not specify the variable "surface of the study area", then this variable was computed using QGIS software (v.3.26, QGIS Development Team, 2022) based on the study boundaries shown in the study area figure without considering the sea surface.

Once the database was filled, a frequency analysis was performed for categorical variables of interest. Additionally, the ratios of "number of samples" to "surface area" (sample density per area) and "number samples" to "number of variables" (sample density per variable) were computed. If a study used different databases separately, then the databases were considered independent in the analysis. In parallel, the application of the most relevant unsupervised learning techniques was described and discussed. The methods were applied to the La Paz aquifer database (Table S1) for a practical demonstration using the open-source R studio software.

4. Results and discussion

Fig. 4 illustrates the scheme of the methodology used. Of the 199 identified documents, 94 were excluded because they did not comply with the filtration characteristics detailed in the methodology section. Data extracted from the remaining 102 articles were analyzed (Table S2).

4.1. Bibliometric analysis

4.1.1. Research characteristics

Most of the research articles belonged to the subject areas of environmental science (43 %), earth and planetary sciences (29 %), and agricultural and biological sciences (11 %) (Fig. 5a). This associations related to the principal objective of the research studies, which was to explore the natural and anthropogenic processes of the study areas based on a physical context using an interdisciplinary approach. The three subject areas medicine (4 %), engineering (3 %), and social science (3 %) are associated with the studies focused on understanding the effect of groundwater processes on health and the social relationships of the study cases. Interestingly, although search keywords for computer science were included, no items were associated with this field.

The areas investigated in the reviewed studies were located in 31 countries (Fig. 5b). The country with the greatest number of studies was India at 19, followed by Tunisia and China at ten and seven, respectively. Similar to the seawater intrusion phenomenon, problems in coastal aquifers are usually associated with overexploitation in arid and semi-arid climates where little groundwater recharge occurs (Parizi et al., 2019). Of the 102 studies, 32 were associated with this type of climate (Köppen classification) in Egypt, Saudi Arabia, Oman, Mexico, Tunisia, Iran, and Djibouti. However, this phenomenon has also been identified in countries with Mediterranean climate, such as Turkey, Greece, Morocco, Lebanon, and Italy. Similarly, countries with humid and subhumid climates and with greater water availability, such as Ghana, Bangladesh, India, Mozambique, and Thailand, also presented seawater intrusion concerns.

4.1.2. Research data

Hydrochemical analyses in the reviewed studies were highly variable owing to differences in the biophysical and socioeconomic settings and sampling strategies. The most extensively analyzed area was in Morocco (710,850 km^2), with 542 samples analyzed for nine variables (Ez-zaouy et al., 2022), while the smallest area studied was a portion of Manukan Island in Malaysia (0.02 km^2), with 162 samples analyzed for 13 variables (Aris et al., 2012). Thus, while the Morocco study had a sample density per

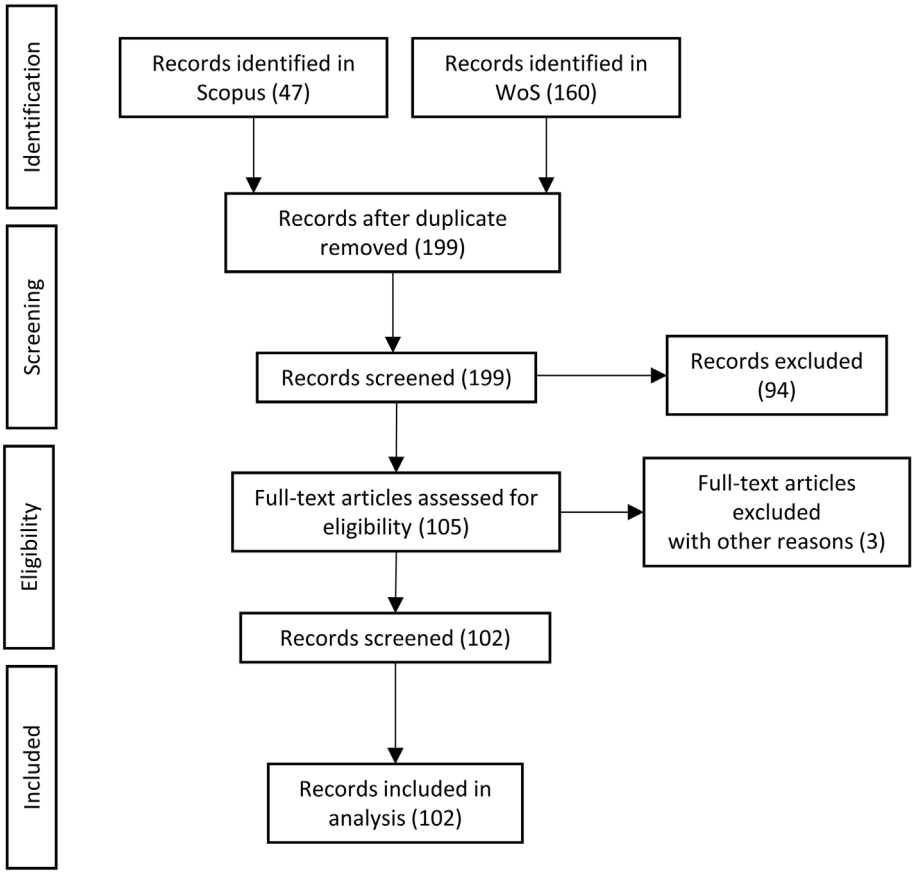


Fig. 4. Flow chart of the literature search and identification process.

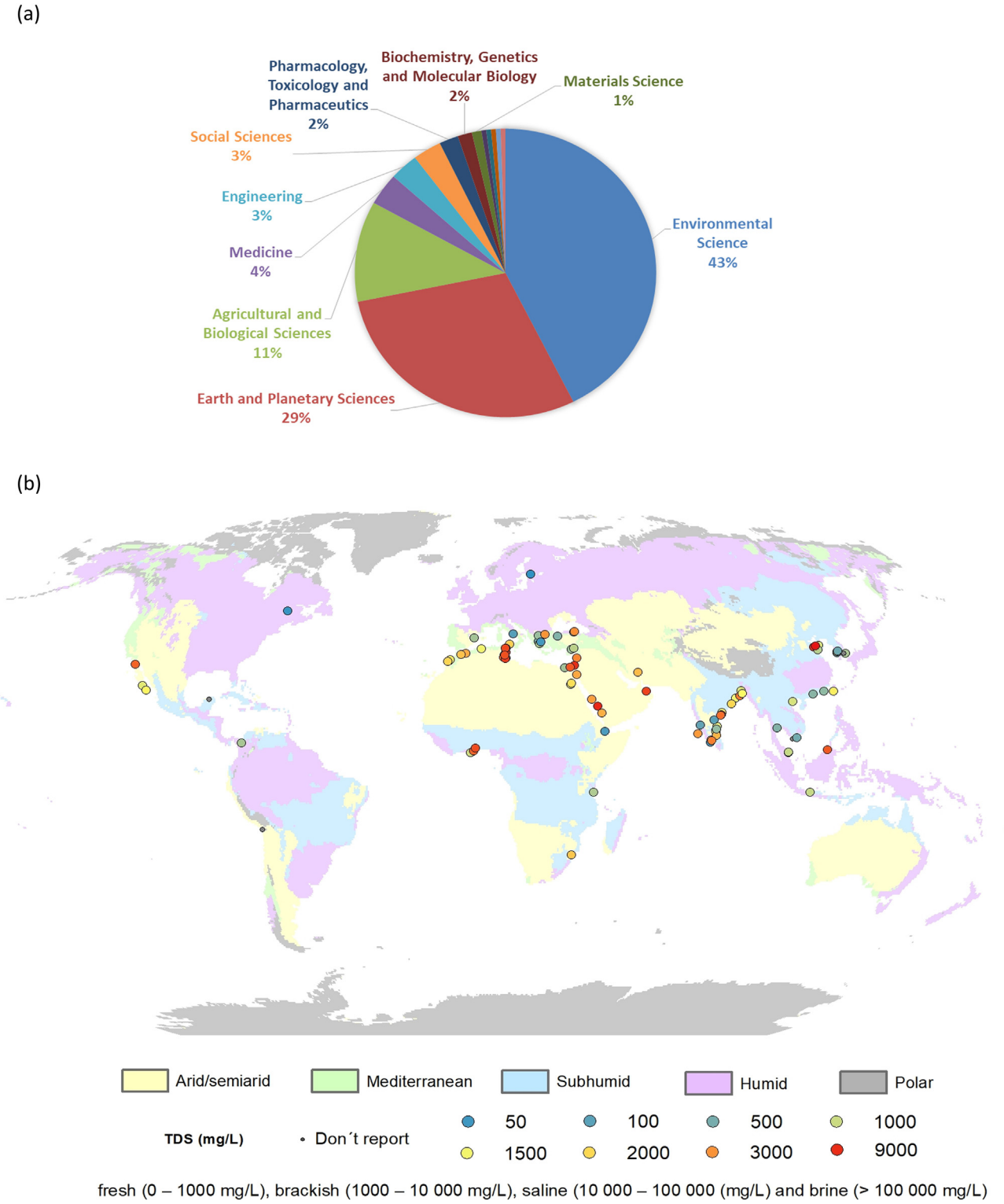


Fig. 5. Research characteristics. (a) Documents by Scopus subject area; (b) Study cases location and salinity (total dissolved solids; TDS < 600 mg/L can be considered as good water quality) (WHO, 2011).

area of 0.0004 samples/km^2 and a sample density per variable of 30.11, the Manukan Island study had a sample density per area of 8100 samples/km^2 and a sample density per variable of 12.46. From a performance perspective, the higher the number of samples and the higher the sample density per area, the better the analysis applied to the data sets (USGS, 2018). For multivariate analysis, the sample density per variable should be as large as possible, although there is no clear rule on the most relevant proportion (Knapp and Campbell-Heider, 1989). Table S2 shows the number of samples, variables, sample density per area, and sample density per variable for each document, and Table 2 summarizes the 102 studies.

Table 2
Data characteristics of the reviewed studies.

Variable	Mean	Minimum	Median	Maximum
Samples	403	9	59	30,809
Variables	14	5	12.5	31
Ratio sample/area (samples/km^2)	90.56	0.0004	0.09	8100
Ratio sample/variable	13.38	0.75	4	669.76

4.2. Description of unsupervised learning techniques

4.2.1. Principal components analysis

PCA is the most commonly used DRM for reducing larger sets of correlated variables into smaller and interpreting datasets regarding the variability of the information (Jolliffe and Cadima, 2016; Björklund, 2019; Ayesha et al., 2020). Principal components (PCs) are uncorrelated linear combinations of the original variables such that the sum of their explained variance is equal to that of the original variables. The variances of the PCs are eigenvalues, whereas the coefficients of the linear combinations are eigenvectors extracted from the covariance or correlation matrix of the data set (Olsen et al., 2012). Typically, a correlation matrix is used if the variables have different measurement scales to standardize the effects and no distributional assumptions are required (Jolliffe, 2002; Jolliffe and Cadima, 2016). In addition to the coefficients, the original variables are associated with the PCs using values between −1 and 1 "loadings", representing each variable's influence on each component. Values close to −1 or 1 indicate significant positive or negative influence, and values close to 0 indicate little influence (Jolliffe and Cadima, 2016).

There are many ways to adapt the PCA method to achieve modified objectives or analyze data of different types (Jolliffe and Cadima, 2016). For instance, including categorical variables is possible with a generalization of the PCA method, called multiple correspondence analysis (MCA) (Audigier et al., 2017; Greenacre and Pardo, 2011). Another important adaptation is orthogonal rotation, which usually applies varimax criterion to simplify the interpretation of the previously computed PCs (Jolliffe and Cadima, 2016). The varimax rotation goal is to maximize the variance of the loadings within the components, thus making larger loadings even larger and smaller loadings even smaller, while preserving the cumulative variance of the components (Denis, 2020). The rotation idea is borrowed from factor analysis (FA), which consists of an array of multivariate statistical methods and is sometimes confused with the specific PCA concept (Jolliffe and Cadima, 2016; Marefat et al., 2019). Currently, statistical packages compute PCs using FA, which allows for rotation to be applied in the same software module (Ayesha et al., 2020; IBM, 2021; Minitab, 2022).

In water research, the PCA method seeks to represent a set of multivariate observations in a lower data matrix arranged along interpretable axes corresponding to known environmental gradients (e.g., warming water temperature and decreasing dissolved oxygen). With the variability aligned to the gradients, it is possible to determine which individual variables are responsible for the greatest observed variation in each axis. This helpful characteristic of PCA may assist monitoring programs by prioritizing limited resources by measuring variables that explain the majority of water quality regime variations (Sergeant et al., 2016). In general, the use of PCA in the reviewed studies was exploratory and, mainly determined the processes and variables that explained most of the variance in the aquifer samples. Additionally, four of these studies used PCA as a pre-processing step for some cluster methods (Table S2). In this manner, the database was reduced to the principal components and clusters were created using the PCs as new variables.

PCA was applied in 77 of the 102 studies. In addition, one study applied MCA, and ten studies applied FA without specifying the extraction method (e.g., PCA, maximum-likelihood method, unweighted least-squares method); however, PCA is set by default in many software programs, such as SPSS and Minitab (IBM, 2021; Minitab, 2022). Of the 88 studies, 63 extracted PCs for analysis based on the Kaiser rule (eigenvalues >1) (Braeken, 2017), and 66 used varimax rotation. Seventy-nine studies associated the first component (PC1) or its equivalent in MCA with words and expressions related to the salinization process, including seawater intrusion, and five of these mentioned mixed salinization and anthropogenic influences (Table S2). Other expressions associated with PC1 were "natural processes", "hardness", "contamination sources", and "dilution of groundwater". On average, the variance of the first component of the 66 studies, with PC1 related to salinization, was 48 %. Variables with loadings >0.8 in PC1 were recorded for each study (Table S1). The most frequent variables in descending order were Cl^-, Na^+, electrical conductivity (EC), total dissolved solids (TDS), Mg^{2+}, Ca^{2+}, SO_4^{2-}, K^+, TH (total hardness), Br^-, HCO_3^-, Sr^{2+}, and salinity, which are typically involved in the salinization process (Jiao and Post, 2019).

Two characteristics that may affect PCAs of coastal aquifers were identified: sample density per variable and the redundancy of variables. Table 2 shows that the sample density per variable of the studies had a median of 4. In the review, 17 out of 102 studies (14 out of 88 in PCA) used less than two samples per variable, and the most extreme study had a ratio of 0.75 (9 samples and 12 variables) (Kumar et al., 2020). Regarding the redundancy of variables, some have been identified as strongly correlated or highly linearly dependent on others in the form $z = m{\cdot}x + n{\cdot}y + \ldots + b$ (King and Jackson, 1999; Senawi et al., 2017). The variables salinity, TDS, and EC generally have very high correlations with each other and can be considered redundant. Moreover, TDS can also be considered redundant when all major ions are considered (e.g., Celestino et al., 2018; Tiwari et al., 2019; Salem et al., 2021), because TDS can be understood as linear combination of these ions (sum). Similarly, TH can be interpreted as redundant (e.g., Gilabert-Alarcón et al., 2018; Sangadi et al., 2022), because it represents a linear combination of bivalent cations (Boyd et al., 2016). According to the above, 91 of the 102 studies used redundant variables (81 of 88 in PCA).

To compare different methods of use, PCA was applied to the La Paz coastal aquifer database (Table S1). The correlation matrix gave the same weight to the variables and, considered all variables, excluding EC and TDS. Table 3 shows the components and their loadings extracted based on the Kaiser rule (eigenvalues >1) and their varimax rotation when excluding EC and TDS. Fig. 6 shows a plot of the scores in the first two directions. PC1 had high loadings in EC, TDS, and major ions, the second had a good loading in NO_3^-, and the third component had moderate loadings in T, pH, and DO, which were associated with the salinization process, nitrification, and oxygen solubility change, respectively. The high affinity of PC1 with the major ions indicated the salinization process (the closer to −1 or 1 the loading, the greater the affinity), nitrification based on the affinity of PC2 with NO_3^- variations; and oxygen solubility based on the affinity of PC3 with T and DO, except with varimax rotation, in which pH loading was increased and T was reduced. Although the MCA method was applied in Hajji et al. (2020), this study was excluded from the comparison, since there were no defined categorical variables in the study case; however Section S2 provides documentation for its application.

Table 3 shows the effects of variable redundancy on the PCA of La Paz. PC1 explained 62.8 % of the variance when all variables were considered, and EC and TDS could be considered redundant ($R^2 = 0.99$). The variance of PC1 decreased to 60 % when EC was removed from the analysis. In the third case, which did not consider EC or TDS, PC1 had a variance of 56.6 %. It should be noted that when considering TDS and EC, their loadings were 0.99 in PC1, indicating that the variance of the component was in the same direction as that of the variables. With the above, redundant variables artificially increased the apparent relevance of some processes and a significant loss of information did not occur when they were removed (Fig. 6). On the contrary, identifying and eliminating this class of variables increased the sample density per variable from 3.6 to 4.2. Regarding the application of the varimax rotation, the redistribution of the accumulated variance stands out, and it reduced the importance of the rotated PC1 and increased the variance in the other two. Fig. 6 shows the scores of the rotated PC1 and PC2, where the rotation of the data pattern to the new orthogonal basis was inferred.

Table 3
Unrotated and varimax rotated factor loadings for the first three PCs with different variables in the La Paz database.

Variable	All Variables			Without EC			Without EC & TDS			Without EC & TDS (varimax)		
	PC1	PC2	PC3	PC1	PC2	PC3	PC1	PC2	PC3	RPC1	RPC3	RPC2
T	−0.18	−0.41	**−0.62**	−0.18	−0.41	**−0.62**	−0.19	−0.42	**0.62**	−0.03	0.16	−0.75
pH	−0.58	−0.28	0.52	−0.59	−0.26	0.52	−0.60	−0.24	−0.51	−0.40	**−0.72**	0.00
DO	−0.34	−0.56	**0.61**	−0.35	−0.56	**0.61**	−0.36	−0.54	**−0.61**	−0.05	**−0.89**	−0.08
EC	**0.99**	−0.12	0.00	–	–	–	–	–	–	–	–	–
Ca	**0.92**	−0.24	−0.13	**0.91**	−0.25	−0.13	**0.91**	−0.27	0.12	**0.92**	0.23	−0.01
Mg	**0.92**	−0.20	−0.06	**0.91**	−0.22	−0.06	**0.91**	−0.24	0.04	**0.92**	0.20	0.07
Na	**0.89**	0.10	0.18	**0.89**	0.08	0.18	**0.89**	0.07	−0.19	**0.78**	0.19	0.42
K	**0.82**	−0.38	0.05	**0.82**	−0.4	0.05	**0.82**	−0.41	−0.07	**0.92**	−0.02	−0.01
HCO_3	**0.80**	0.27	0.16	**0.80**	0.25	0.16	**0.80**	0.24	−0.16	**0.63**	0.28	0.50
Cl	**0.97**	−0.19	−0.02	**0.96**	−0.21	−0.02	**0.96**	−0.22	0.01	**0.96**	0.19	0.11
NO_3 as N	0.47	**0.72**	0.20	0.48	**0.70**	0.2	0.49	**0.70**	−0.19	0.16	0.42	**0.75**
SO_4	**0.88**	0.02	0.14	**0.88**	0.00	0.14	**0.88**	−0.01	−0.15	**0.81**	0.18	0.34
TDS	**0.99**	−0.06	0.04	**0.99**	−0.08	0.04	–	–	–	–	–	–
Variance	8.17	1.45	1.16	7.20	1.43	1.16	6.22	1.42	1.16	5.31	1.80	1.70
%	**63 %**	11 %	9 %	**60 %**	12 %	10 %	**57 %**	13 %	11 %	**48 %**	16 %	15 %

Note: Factor loadings >0.6 are in bold.

4.2.2. Hierarchical cluster analysis

Hierarchical clustering (HCA) forms clusters by dividing data patterns using a divisive or agglomerative approach (Saxena et al., 2017). In the agglomerative approach, the first connections correspond to the closest pairs of objects based on Euclidean distance or other indicators. Subsequently, the initial groups are connected to the closest group based on their similarities through a linkage algorithm (e.g., complete-linkage or Ward's-linkage), and the process is repeated until only one group remains (Strauss and von Maltitz, 2017; Székely and Rizzo, 2014). The results are presented in a dendrogram that shows the connection between the different group levels and the linkage distance regarding the samples (R-mode) or the variables (Q-mode). In general, for this and the other clustering methods, attributes are "normalized" (standardized) with functions such as the z-score to give all attributes appropriate and comparable importance (Bouguettaya et al., 2015). While this method can consider observations and variables at multiple levels of grouping, the disadvantage of HCA is that it is sensitive to noise and outliers (Saxena et al., 2017), although the most robust linkages (e.g., average and Ward's) may help limit this effect (Bu et al., 2020).

One application of HCA in hydrology is to depict correlation patterns among water samples, thus enabling a more rapid identification of the main hydrogeochemical processes than with the use of only descriptive statistics (Nogueira et al., 2019). Among the reviewed studies, 77 of 102 studies applied HCA in their analysis, 64 applied Q-mode, 25 applied R-mode, and 8 applied both Q- and R-mode. Of the 77 studies that applied HCA, 64 also applied PCA to interpret hydrogeochemical data as a complementary analysis, thus highlighting that HCA and PCA have a strong relationship. Studies that performed HCA (R-mode) showed similar variable associations of the hydrogeochemical processes obtained using the PCA method. As mentioned in Section 4.2.1, PCA was used in three studies to reduce the data dimensions before clustering through HCA (Table S2). Of the 60 studies that applied the Q-mode, 45 related at least one cluster to the influence of seawater intrusion and one related at least one cluster to the influence of brackish/saltwater. Of the 25 studies that applied the R-mode, 20 related one variable cluster to seawater intrusion or other salinization sources.

Ward's-linkage was the preferred linkage algorithm in these studies. Of the 77 studies that applied HCA, 52 applied this linkage rule. Ward's method is distinct from all other methods because it uses an analysis of variance approach to evaluate cluster distances (Sharma and Batra, 2019; Ward, 1963). Ward's-linkage seems to perform significantly better than other clustering procedures based on empirical studies that compare the methods (Willett et al., 1998). Furthermore, three studies used complete linkage (Saxena et al., 2017), while the remaining did not specify a linkage method. According to the same trend, 56 studies established the distance

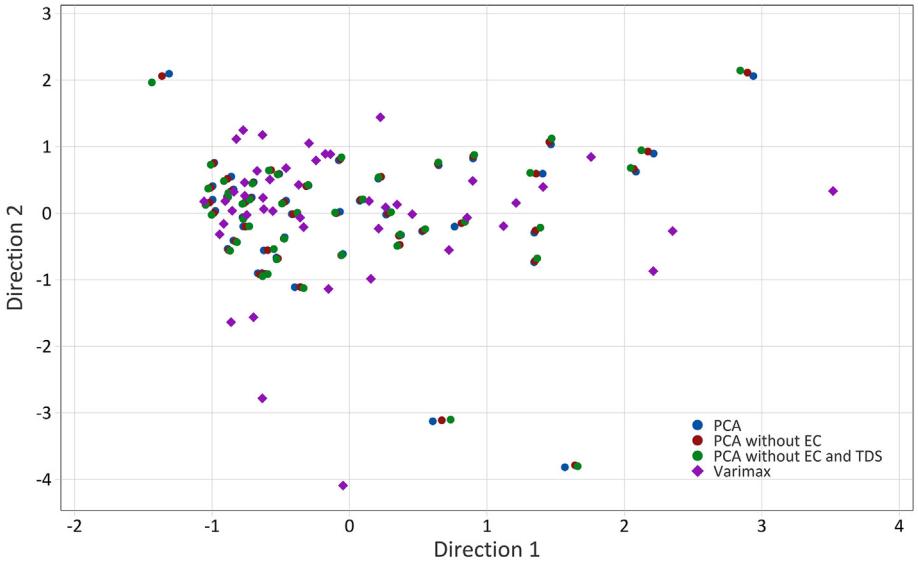


Fig. 6. First two unrotated and varimax rotated direction scores for different numbers of variables in the La Paz database.

criterion, in which the most used similarity measure was Euclidean distance in 54 studies. The other two distance criteria were Manhattan (Strauss and von Maltitz, 2017) and Pearson (Székely and Rizzo, 2014) (Table S2). The lack of information regarding the used linkage method and distance criterion can make it difficult to replicate and validate the results of the studies.

As seen with PCA, TDS and EC can be removed from HCA clustering. When redundant, these variables give greater importance to a specific variance direction without adding new relevant information and removing them reduces the complexity of clustering (Fraiman et al., 2008; Mitra et al., 2002). To demonstrate this, Ward's method of HCA with Euclidean distance was applied to the La Paz coastal aquifer database for the different sets of variables used in PCA, using the R script described in Section S3. *Z*-score standardization was used to assign the same weight to all variables, and four clusters were extracted. Fig. 7 shows that the dendrogram structure changed by a minor degree when EC or TDS were removed and the linkage distances were reduced. Clusters that did not consider EC or TDS were raised in the first two PCs in Fig. 8a to understand their relationship with the database. From the PCA, Cluster 1 (C1) had a minor salinization effect; C2 and C4 had an intermediate effect, and C3 had the highest values in the PC1 direction. Regarding nitrification (PC2), the only cluster with higher values relative to others was C3. Fig. S2 shows the location of the sampling points indicating the cluster membership.

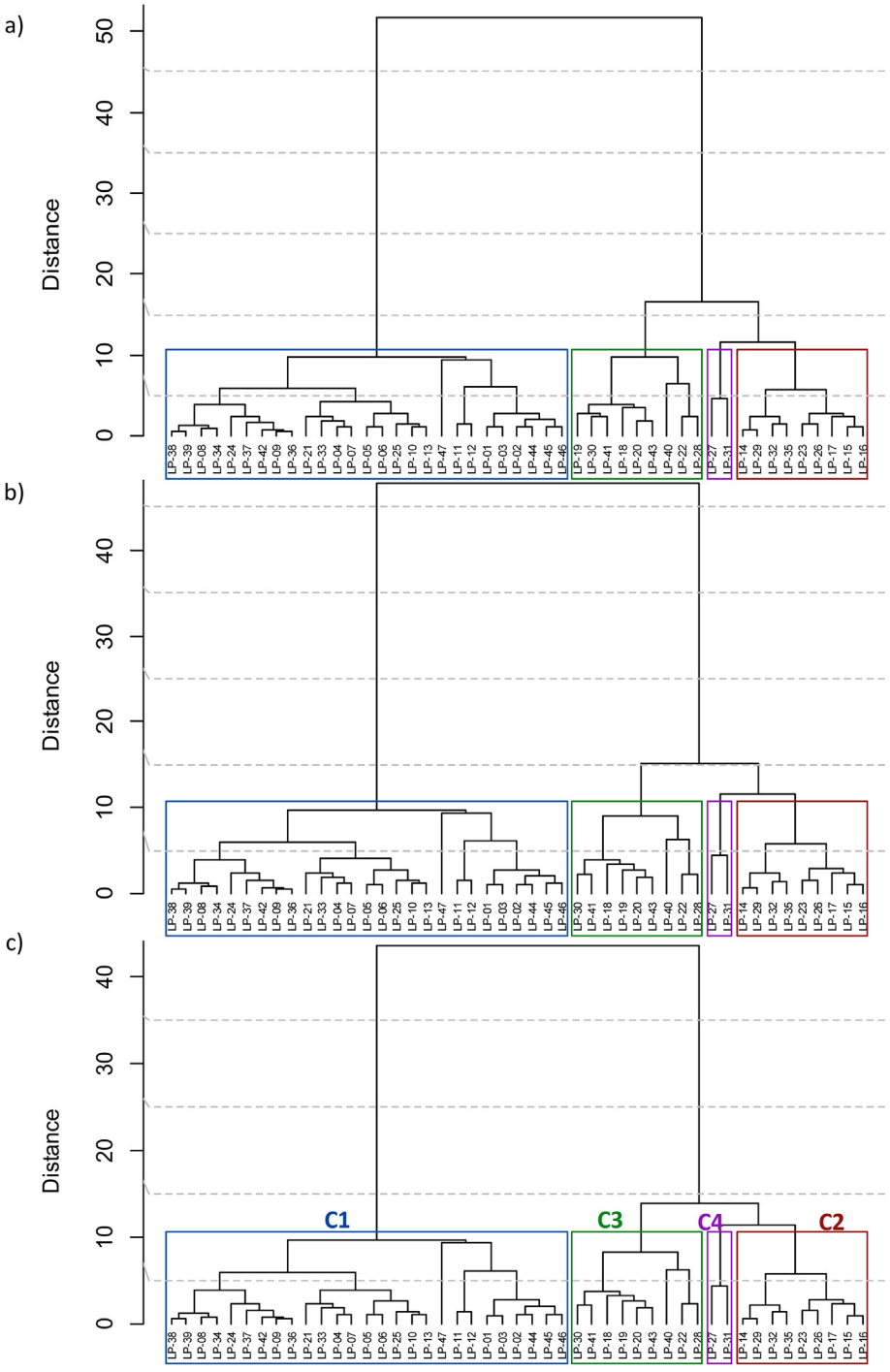


Fig. 7. Hierarchical clustering (HCA) dendrogram for different variables of the La Paz database. (a) All variables; (b) without electrical conductivity (EC); (c) without EC or total dissolved solids (TDS).

The PCA and HCA information, sample point location, and extended hydrogeochemical facies analysis of the groups in Section S3 indicates that C1 is associated with recharging groundwater, C2 is associated with seawater intrusion mixture, C3 is associated with terrestrial salinization, and C4 is associated with nitrate-polluted groundwater mixed with seawater intrusion.

4.2.3. K-means clustering

K-means clustering uses a predefined number (K) of centroids to generate the best adjustment between these centroids and the input data through an iterative process (Saxena et al., 2017). In the first step, a few centroid points are randomly selected. Each data point is then assigned to the closest centroid. The next step is to update the centroids by calculating the central points of these newly formed clusters using the Euclidean distance. Subsequently, the last two steps are repeated until no object changes the cluster assignment (Bouguettaya et al., 2015). The results may vary because of the initial random location of the centroids. The algorithm can then be run multiple times to choose the best result based on the minimum value of the total sum of squares (TSS). The predefined number (K) of cluster centroids depends on the modeler criteria, although there are heuristic rules to infer the optimal number. The elbow criteria usually repeats the process by increasing the number of centroids, and optimal clusters are inferred by selecting the elbow of the curve between the distortion measurement and the number of centroids (Yuan and Yang, 2019). Similar to HCA, the disadvantage of this method is that it is sensitive to noise and outliers (Saxena et al., 2017).

The fuzzy C-means (FCM) method is a modified form of the K-means method that uses fuzzy logic theory, in which objects can be partially assigned to multiple clusters (Mohammadrezapour et al., 2020). The proportion of membership to each cluster depends on the closeness of the data object to the centroids, and a fuzzy matrix describes this characteristic with "n" rows (objects) and "c" columns (clusters) (Izakian and Abraham, 2011). This partial assignment increases the expressiveness of the clustering analysis, thereby presenting a more comprehensive view of the relationships in the data (Stetco et al., 2015). To perform the segmentation, the fuzziness must be indicated through the fuzzy partition matrix exponent "m", with m >1 (Saxena et al., 2017). Similar to K-means, centroids are located randomly and the best combination can be chosen by selecting the final TSS minimum value between multiple runs. Despite improvements in the conventional partitioning method, FCM is still sensitive to noise and outliers (Saxena et al., 2017).

In hydrology, K-means and FCM have been used to depict associations between water samples and infer sources, which is similar to HCA (Mohammadrezapour et al., 2020). This review showed that K-means was used in two studies to identify groups for hydrogeochemical analysis (Table S2). In these studies, the analysis of the groups indicated the association of at least one group with the influence of seawater intrusion. The interpretation of the K-means results was used with complementary techniques, such as clustering PCA results by Celestino et al. (2018) and segmentation SOM results by Yin et al. (2021). FCM was applied by Güler et al. (2012), which enabled the identification of a cluster associated with seawater intrusion. Additionally, this study allowed the identification of transitional zones between the clusters through the georeferenced graphical representation of the fuzzy transitions.

K-means was applied to the La Paz coastal aquifer database for different variables to test the method and to determine the influence of redundant variables. *Z*-score standardization was used to assign the same weight to all the variables, and four clusters were extracted. The algorithm was executed 100 times using the R script described in Section S3, and the result with the lowest TSS was selected. The sampling cluster assignation had the same results for all three cases: all variables were considered, EC was omitted, and EC and TDS were omitted. A fewer number redundant variables corresponded to a lower TSS (i.e., 598, 552, and 506 for the three cases, respectively). This highlights the redundancy of TDS and EC with other chemical parameters, which implies that they can be omitted. Clusters were raised in the first two main components (Fig. 8b), which generally preserves the same pattern as that observed in the HCA clustering results (Fig. 8a) but generates a different pattern in the three sample classifications (Lp-21, Lp-20, and Lp-43). Fig. S7 shows the location of the sampling points indicating the cluster membership.

As an example of the fuzzy concept, FCM was applied to the La Paz coastal aquifer database without considering TDS or EC. *Z*-score standardization, four centroid clusters with 100 runs, and a fuzzy partition matrix exponent of m = 1.3 were set using the R script described in Section S4. The result with the lowest TSS was 502.17, and the fuzzy matrix was extracted

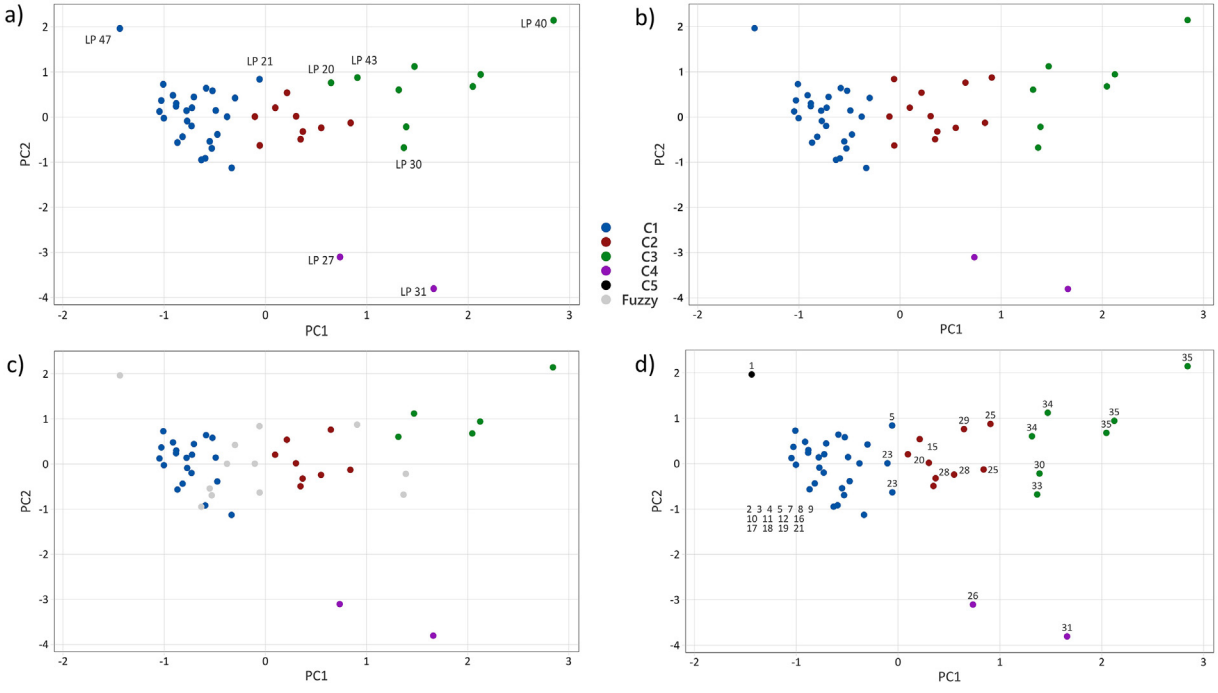


Fig. 8. Clustering without electrical conductivity (EC) or total dissolved solids (TDS) on the first component (PC1) vs. second component (PC2). (a) Hierarchical clustering (HCA); (b) k-means clustering; (c) fuzzy C-means; and (d) Self-organizing map (SOM) neural segmentation without EC or TDS.

as shown in Section S4. Fig. 8C shows the FCM cluster pattern on the two first PCs, in which samples that have been assigned with a proportion <90 % to a single cluster membership are marked as fuzzy. For instance, 77 % of the Lp-21 sample belongs to C2, 22 % belongs to C1, and <1 % is assigned between C3 and C4; 81 % of the Lp-30 sample belongs to C3, 17 % belongs to C2, and <2 % is assigned between C1 and C4; and 76 % of Lp-47 belongs to C1, 20 % belongs to C2, and <4 % is assigned between C3 and C4. Fig. S8 shows the location of sampling points that indicate the fuzzy cluster membership.

4.2.4. Self-organizing map

The SOM or Kohonen map is a specific type of ANN for visualizing and clustering high-dimensional data. It converts the non-linear statistical relationships between high-dimensional data into simple geometric relationships projected on a low-dimensional display, usually a regular two-dimensional grid of neurons (Kohonen, 2001; Wehrens and Kruisselbrink, 2018). The SOM neural network consists of an input and output layer (Kohonen layers). The input layer contains as many nodes as variables in the data set, whereas the output layer neurons are connected to every neuron from the input layer through adjustable weights or network parameters that form the weight vectors. The weight vectors constrain the reproduction of the input objects through the output layer in an ordered but not regular mesh that preserves the data topology (Kalteh et al., 2008; Wehrens and Kruisselbrink, 2018). After training, the input objects are assigned to the output layer neurons, and some neurons may not contain objects, thus increasing the difficulty of interpreting the SOM map. Reports have suggested that the map size should be varied to avoid as many empty neurons as possible since these do not represent the data pattern (Céréghino and Park, 2009; Li et al., 2018).

The first and most important step in applying the SOM method is data gathering and standardization to prevent variables from having a higher impact than others, such as in clustering techniques. The second step consists of defining several neurons associated with the weight vectors; typically, the heuristic rule of $w = 5\sqrt{m}$ is used, where m is the number of samples and w is the number of output layer neurons (Céréghino and Park, 2009; Yin et al., 2021). The initial values of the weight vectors of the neurons are established randomly. The next step is training, in which the weight vectors simultaneoulsly update their relative values for one input pattern and a neighborhood function. The neuron with the closest match to the presented input pattern is called the winner neuron or best-matching unit, and the next input pattern is used as the new target. It is recommended that the number of iterations be at least 500 times the number of neurons in the output layer. The last step is information extraction and visualization. Typically, a 2D projection of the final output neural mesh is used (Kalteh et al., 2008).

From the perspective of water research, a trained SOM map is a valuable tool for visualizing the data and obtaining insights into the system under investigation, such as satellite imagery data classification, rainfall-runoff analysis, and water quality associations (Kalteh et al., 2008; Olkowska et al., 2014). This review found two studies that combined SOM with K-means and HCA clustering to visualize and cluster hydrogeochemical data (Nguyen et al., 2015). This visualization highlighted the patterns of each variable's influence in 2D neuron maps and showed the sample clusters obtained in the SOM map projection. The combined methodologies permitted a better interpretation of the hydrogeochemical processes and identified at least one cluster associated with seawater intrusion.

SOM was applied to the La Paz database. *Z*-score standardization was used to assign the same weight to all variables, and 35 neurons ($w = 5\sqrt{47}$) with an array of five columns and seven rows were used, while 18,000 (35 × 500) iterations were set. The R script and additional parameters are described in Section S5. Fig. 9 shows the results of the final representative weight vectors and their neuron locations. In general, the patterns of the three analyses were very similar. In the upper right part, there were samples with the highest values of major elements, TDS and EC; in the bottom part, there were the samples with the highest DO and pH values; and in the middle left part, there were the samples with the lowest values of all variables. Neurons 26 and 31 showed the highest nitrate values, and EC and TDS were redundant in the analysis because the weight vectors of the major ions had the same pattern as EC and TDS, indicating that there was no new information when these variables were considered.

To provide a more robust interpretation of the results, HCA clustering was used with the neuron weight vector results when excluding EC and TDS (Ward's method, Euclidian distance, and 5 clusters). The SOM grid map segmentation with clusters is presented in Fig. 10 with the mean results of TDS, NO_3-N, and DO on the SOM grid. The neurons associated with the samples and neuron clusters are shown in Fig. 8c, and Fig. S9

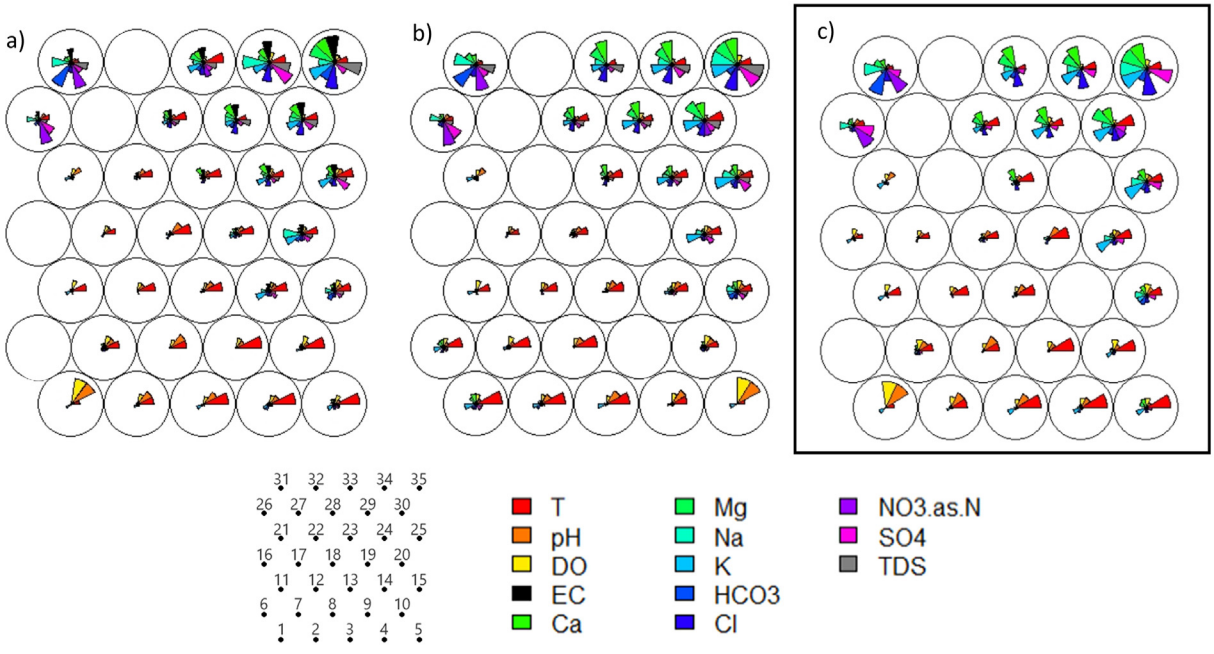


Fig. 9. Self-organizing map (SOM) grid for different numbers of variables of the La Paz database. (a) All variables; (b) without electrical conductivity (EC); (c) without EC or total dissolved solids (TDS).

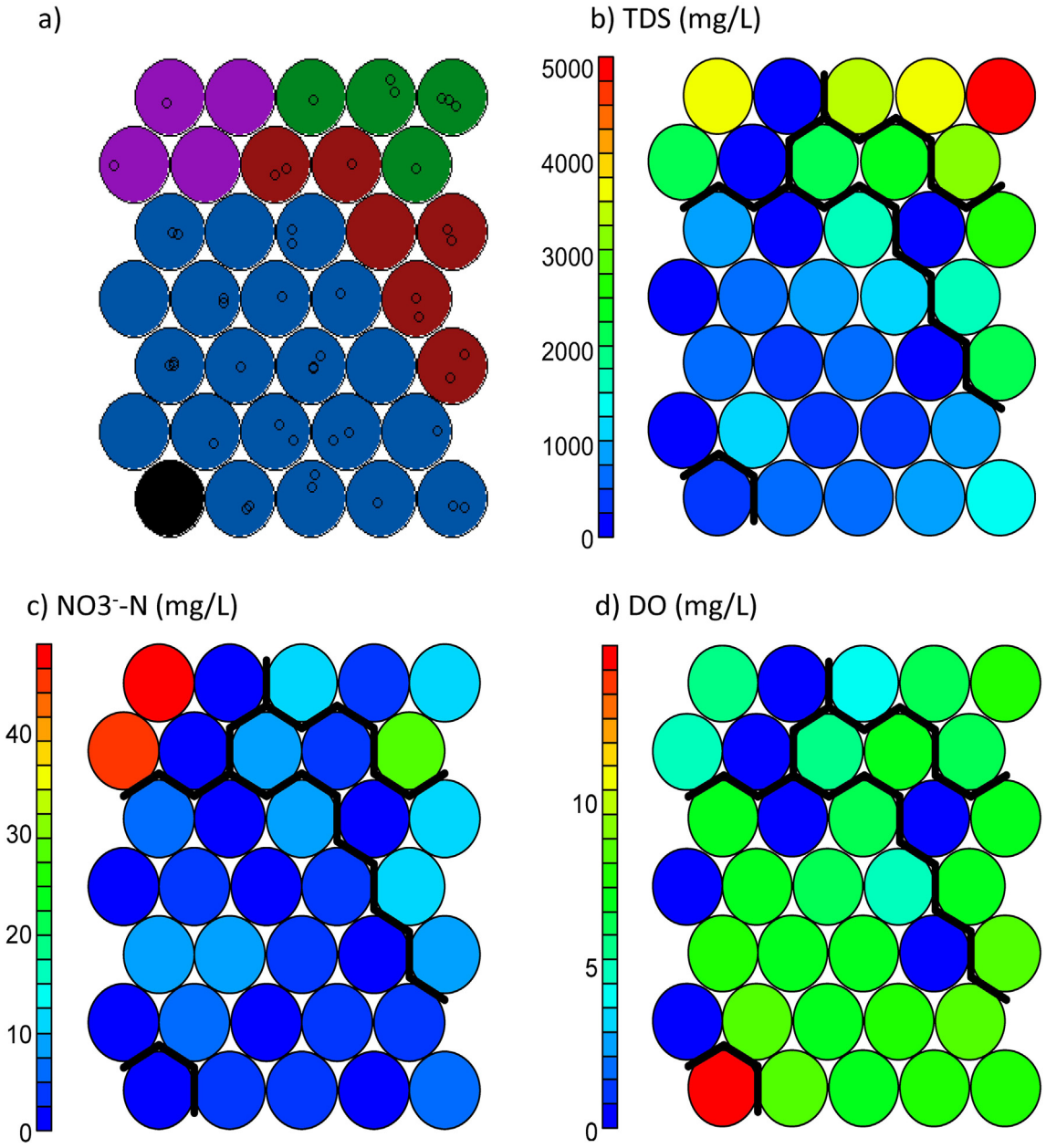


Fig. 10. Self-organizing map (SOM) grid of La Paz database: (a) Weight vector hierarchical clustering (HCA); (b) Total dissolved solids (TDS) mean value projections in the SOM grid; (c) HCO_3^- N mean value projections in the SOM grid; (d) Dissolved oxygen (DO) mean value projections in SOM grid.

shows the sampling point location, indicating the SOM code cluster membership. The clustering result and its interpretation are similar to those of HCA, K-means, and FCM. In contrast, SOM shows the internal structure of the groups and their relationship to topology. It should also be noted that five clusters are required to arrive at a segmentation similar to the other methods. When four clusters with the indicated SOM features were set, C2 and C3 belonged to the same group and Lp-47 sample formed a single cluster.

4.3. Seawater intrusion pattern recognition

The reviewed unsupervised learning techniques were used to conduct exploratory analyses of the hydrogeochemical processes governing coastal groundwater quality. The PCA, MCA, and HCA (R-mode) techniques were used to associate hydrogeochemical processes based on the similarity and variance of the data variables. Grouping and segmentation techniques, such as SOM, HCA (Q-mode), K-means, and FCM, made it possible to assemble water samples with similar characteristics, mainly hydrogeochemical facies, which were associated with different sources, such as salinization from evaporation, seawater intrusion, or anthropogenic impacts, based on the support of different discrimination criteria. However, owing to their unsupervised nature, these techniques do not allow the clusters to be directly associated with the sources. It is undeniable that simplification of the databases and pattern recognition facilitated the interpretation of hydrogeological processes in each case study.

Monitoring major ions as chemical tracers of environmental processes is of enormous importance because they account for 95 % of TDS (Poeter et al., 2020). In general, PCA also highlighted the relevance of the variance of these constituents in PC1. On average, the salinization process was the most relevant and explained 48 % of the data variance in the studies. Although it could be argued that there is a bias because major ions are always considered for sampling analysis, 22 studies with a PC1 average variance of 43 % considered at least 15 different variables, which is twice the number of major ions. Other variables that were highly relevant in PC1 when

associated with salinization processes were F^- (Askri et al., 2022), Fe (Awaleh et al., 2018; Galazoulas and Petalas, 2014; Kim et al., 2005), chemical oxygen demand (Wang et al., 2022), Sr (Hyung et al., 2021), Mn (Hyung et al., 2021), B (Güner et al., 2021), Se (Papazotos et al., 2020), Br (Sae-Ju et al., 2020), Cr (Galazoulas and Petalas, 2014), NH_4^+ (Salem et al., 2021), total viable count (Gokul et al., 2019), Li (Souid et al., 2018), As (Houssein et al., 2017), sodium adsorption ratio (Taşana et al., 2022), and seawater intrusion generalized index (Alameddine and Fadel, 2021). Table S2 shows other studies in which these variables were relevant to PC1.

Because the greatest variance is in the direction of salinization, it can be understood that this process partially biases the formed clusters. For the case of La Paz, it can be seen in Fig. 8 that data clusters are partitioned in the directions of the first two principal components, although separation is more predominant by PC1. Although clusters appear to be effective, it should be noted that clusters methods are sensitive to "outliers" or "anomalies" (Saxena et al., 2017), such as Lp-47 (water recharge), Lp-40 (high salinity), and Lp-31 (high nitrate concentration) (Fig. 8a). These outlier observations are of great interest and should be identified because they can affect the interpretation of the related clusters. HCA and SOM present features against this because it is possible to identify the outliers in the dendrogram and SOM segmentation pattern, respectively. Another characteristic for differentiating groundwater groups is the optimal number of clusters. An average of 3.7 clusters was used in the studies that relate at least one group to seawater intrusion influence, and 91 % of these studies did not use a criterion for selecting the optimal number of clusters. As Pacheco Castro et al. (2018) stated, a hydrogeological sense should be used to select the final number of clusters, and the number should be increased until significance is observed.

The most commonly used criteria for identifying seawater intrusion in the cluster results are associated with the chemical characteristics of seawater and its interaction with the aquifer. Associations were sought based on samples similar to seawater, such as those with high salinity values, high chlorine concentrations, and Na^+-Cl^- facies (Jiao and Post, 2019). The interaction between seawater and the solid aquifer matrix is also an indicator because when the sea wedge advances (seawater intrusion), reverse cation exchange may occur, in which Na^+ is exchanged for Ca^{2+} in the signature of groundwater; however, when the wedge recedes (freshening), direct cation exchange occurs, with Ca^+ exchanged for Na^+ (Giménez-Forcada, 2010). These evolution trends are understood from different diagrams that have been developed over the years, such as Piper (Moreno Merino et al., 2021), Durov (Chadha, 1999), Stiff (Lee, 1998), and HFE-D (Giménez-Forcada, 2010). Cation exchange in water samples can also be identified by comparing the excess and deficit of ions in the theoretical mixture of recharge water and seawater (fraction of seawater) through end-members (Nogueira et al., 2019; Papazotos et al., 2020). These discrimination techniques based on major ions appear to be effective in places where seawater intrusion is the main source of salinization. However, under arid and hyper-arid conditions, it is difficult to differentiate the sources of high salinization values with similar signatures (Sabarathinam et al., 2021).

Apart from facies and tracers, the salt origin can be inferred by comparing the chemical and isotopic concentration ratios of groundwater samples and seawater, such as Na^+/Cl^-, Ca^{2+}/Cl^-, Mg^{2+}/Ca^{2+}, Ca^{2+}/Mg^{2+}, Cl^-/HCO_3^-, (Jiao and Post, 2019; Lee and Song, 2007), Cl^-/Br^- (Bertrand et al., 2022), Cl^-/Si (Sabarathinam et al., 2021), $\delta^{34}S(SO_4^{2-})$ (Hyung et al., 2021; Kim et al., 2019), $\delta^{13}C$(Dissolved Inorganic Carbon), $\delta^{11}B$, B/Cl^-, and $^{87}Sr/^{86}Sr$ (Awaleh et al., 2018). In addition, a more composited relationship based on major ions has been used to infer the influence of salinization intrusion, such as the Simpson ratio for evaluating the salinization degree and the chloro-alkaline index (CAI) for evaluating the degree of ion exchange (Ha et al., 2022; Wang et al., 2022). Depending on the major ions, the results obtained from the CAI and Simpson index reflect the results of different salinization sources, not only seawater intrusion. Overall, most of the reviewed studies used the Piper diagram and/or major ion relations as discriminant techniques, and only four studies included major ion ratios as inputs in the multivariate analysis (Table S2). In contrast, 20 articles used isotopes (mostly δ^2H and $\delta^{18}O$) for analysis, of which only five included isotopes as inputs for multivariate analysis (only δ^2H and $\delta^{18}O$) (Table S2). However, these two variables are related to water sources, which can be seawater, among others, and not directly related to salinization.

The lithology of the study area is also a discriminating factor for salinization identification. The saturation indices of minerals based on thermodynamic and geochemical mass balances help to characterize groundwater influenced by different sources of mineralization, showing whether water is under- or oversaturated with respect to given minerals, such as halite, gypsum, calcite, and dolomite (Parkhurst and Appelo, 2013). Generally, the zone affected by seawater intrusion is undersaturated with respect to halite (Güler et al., 2012; Sabarathinam et al., 2021). Less complex techniques to compare the groundwater interaction with the solid matrix are the Gibbs diagram (Marandi and Shand, 2018) and Na-normalized diagrams (Gaillardet et al., 1999). The first compares water samples to the pattern of world water resources, thus indicating the influence of evaporites (along with seawater intrusion) and rock interactions with recharge water, while the second associates the water samples with different registered carbonates, silicates, and evaporite rock end members. In these three discriminant techniques, the influence of seawater intrusion can be confused with that of evaporation and evaporite mineral dissolution. Thus, the results should be compared with the lithologies of the study area to identify associations and discrepancies. Table S2 lists the studies that have used these techniques.

Hydraulic characteristics are also important for interpreting hydrogeochemical data. Knowledge of flows, hydraulic gradients, and sample grid spatial configuration helps to identify the mineralization process to which water is subjected along flow lines. For instance, proximity to the coast is associated with the influence of seawater intrusion, such as in Hajji et al. (2020), El Yaouti et al. (2009), and Yik et al. (2012). In addition, when inland salinity is unclear, the hydraulic gradient, sea-level rise, and groundwater overexploitation may suggest associations with seawater intrusion extension (Ferguson and Gleeson, 2012). For instance, while the Red River delta aquifer (hydraulic gradients $<10^{-4}$) in Vietnam has a seawater extension of hundreds of kilometers and salinization deposits from the Holocene seawater intrusion (Larsen et al., 2017), the Caplina/Concordia aquifer system (hydraulic gradients of the order of 10^{-2}) in Peru/Chile has a seawater intrusion extension of approximately 10 km, which is mainly due to overexploitation (Narvaez-Montoya et al., 2022). This review identified four studies that used hydraulic conditions (distance to the coast, field slope, and hydraulic head) as inputs for unsupervised techniques (Table S2). It is necessary to continue including this type of variable in this class of studies since they present a meaningful value for the associations and interpretations of coastal hydrogeology.

Even with an understanding of the study area, data, and unsupervised applied techniques, it is difficult to distinguish seawater intrusion from other phenomena with total confidence. Mechanisms that involve salty water of different origins from the sea or fossil seawater stored inside the aquifer can result in similar geochemical signatures, such as high concentrations of ions and alteration of facies through cation exchange. Therefore, data can be misinterpreted. In most studies, it is not assured that the water samples originate from seawater intrusion; rather, the samples are usually associated with this source. Apart from the environmental tracer and general hydraulic feature analysis, it is necessary to implement complementary studies, such as geophysical methods and numerical groundwater models, to understand this phenomenon. In this review, only eight case studies of this type were identified (Table S2).

4.4. Recommendations for unsupervised hydrogeochemistry data analysis

Several aspects were identified that might be useful for researchers and professionals to improve the analysis using reviewed unsupervised techniques. Except for coastal hydrogeology, most of the recommendations to be adopted for water research are based on data analysis.

For the preprocessing of raw data, TDS, EC, and TH can be considered redundant variables when used together and when major ions are considered, such as in the Laz Paz analysis example. The exclusion of these variables helps limit the complexity of the analysis by eliminating the multiplicity of the same effect without losing important information. However, the variables must be redundant; for instance, TDS and EC were not considered redundant in the study of Zhu et al. (2020) because their correlation coefficient was 0.73. To detect undesirable redundancies, tools for computing the correlation matrix and detecting linear dependencies are included in Section S7. Similarly, increasing the sample density per variable can improve the significance of the results of multivariate analysis. Multiple rules of thumb have been generated to designate the minimum value of this relationship, however, the values differ considerably (2:1 to 30:1) because these studies applied different multivariate techniques and different methodologies (Knapp and Campbell-Heider, 1989).

Second, data should be standardized to assign the same weight to all meaningful variables. Most studies that performed standardization used the z-score, which consists of centering the variables with a mean of zero, scaling to unit variance, and retaining the magnitude proportions. Other methods for standardizing the variables can be found in Miuigan and Cooper (1988). Note that PCA does not require standardization to assign equal importance to the variables if the correlation matrix (by default) is used, because the raw data correlation matrix is equal to the standardized data correlation matrix (Jolliffe and Cadima, 2016). Moreover, using logarithm transformations for preprocessing is not recommended since the reviewed techniques are exploratory and do not require distributional assumptions (Jolliffe, 2002; O'Hara and Kotze, 2010). Most environmental data, including geochemistry data, follow a skewed positive distribution (Andersson, 2021; Govett et al., 1975). Log-transforming search strategies are usually used for normal distributions, although such transformations are not necessary, such as in Pacheco Castro et al. (2018). The reason for this is paradigmatic and unclear, although the transformation implies that the mechanisms are multiplicative on the scale of the raw data (Govett et al., 1975; O'Hara and Kotze, 2010), and which distorts the data's internal relationships. Of the 102 studies, 21 transformed their data (20 %).

Reproducibility of the results and further exploration of related hypotheses require access to raw data (Alsheikh-Ali et al., 2011). Although journals encourage the sharing of data and other useful materials related to research, such sharing does not occur in many cases. Only 39 of the 102 reviewed studies (38 %) provided the raw data for their analysis. These relevant data can be placed in the supporting material of studies, which generally permits multiple formats; moreover, data can be both shared and protected (through DOI's) in ad-hoc repositories, such as Zenodo (Sicilia et al., 2017). Regarding the accessibility of unsupervised technique tools, there is no major problem in accessing different software, such as Minitab, SPSS, Stata/SE, and STATISTICA. However, the tools must be made available to the general public for reproducibility and validation.

Furthermore, given that this research is applied to water resources, which are associated with the human right to access water and UN-Sustainable Development Goal 6 (clean water and sanitation), it is necessary to socialize raw data, methods, and results as best as possible. This work highlights the importance of sharing data and using open source software to validate, reproduce, and replicate the research. The techniques applied to the La Paz database were executed using open-source R Studio environment. R scripts and documentation for discriminant techniques, such as Piper, are provided in the Supplementary material.

4.5. Research opportunities

Most studies have used several techniques independently to explore and interpret hydrogeochemical data. The integration of at least two techniques was identified in a few studies that integrated HCA and K-means with prior dimension reduction via PCA (e.g., Aris et al., 2012; Celestino et al., 2018; Hasan et al., 2021; Osiakwan et al., 2021) and that combined HCA or K-means with SOM to obtain clusters over the data topology (e.g., Nguyen et al., 2015; Yin et al., 2021). It is possible to continue integrating these techniques with other diagrams, thus creating new analysis strategies. Moreover, although unsupervised learning has a bias component because the modeler is the one who gives meaning to the results, the new techniques and procedures must be more objective and consider elements of the data, such as the variance and amount of data.

One problem with interpreting cluster data is that when considering all variables, all processes are integrated; therefore, the mechanisms featuring the highest variance can overshadow the others. To solve this problem, Mora et al. (2021) used an HCA double clustering approach, whereby the HCA (R-mode) was first executed on the data, and each determined set of variable groups was associated with a hydrogeochemical process, such as salinization. Sample clusters were then formed from each group of variables (process) by applying HCA (R-mode) to the data. In this way, the samples were clustered according to an identified process, thus identifying the level of influence and associating different sources. One way to improve the strategy of sample clustering using a different process is to use PCA instead of HCA (R-mode) to determine the variables and infer the general process in the first stage. In this way, the variables are not only associated with a process, but clarity is also obtained based on the importance of the processes that lead to data variance. For instance, three relevant processes were identified in the La Paz case, and their variance contributions were calculated (salinization accounted for 62.8 % of the total variance, nitrate contamination accounted for 6 %, and oxygen solubility change accounted for 5.6 %). Thus, three sets of HCAs (Q-mode) clusters can be generated with the most relevant variables of each component (highest loading).

Exploring coastal hydrogeochemical data with alternatives to the reviewed techniques is possible. For instance, independent component analysis (ICA), a DRM, can be used instead of or parallel to traditional PCA. ICA extracts independent sources (directions) by exploring statistically independent patterns from the observations of an unknown linear mixture. This technique is more powerful than others based on uncorrelated components, such as PCA (Calabrese, 2019; Kano et al., 2004). Another DRM that can be used to reveal the global structure of hydrogeological datasets and is not based on linear relations but rather on probabilistic distributions, is T-stochastic neighbor embedding. This technique maximizes the relationship between the closest observations and minimizes the influence of the most distant observations (Ayesha et al., 2020).

For novel clustering applications in coastal aquifers, model-based clustering can associate observations to multiple clusters since this method creates groups based on a mixture of component models (Fürnkranz et al., 2011); and density-based clustering can create clusters based on contiguous dense regions and identifying and eliminating the influence of “anomalies” in the data (Hahsler et al., 2019). Another cluster that can avoid the large influence of anomalies is K-medoids, in which an actual point (medoid) is used to represent the cluster center rather than the mean point as the center of a cluster (k-means) (Mannor et al., 2011). Another methodology that can be used to detect anomalies is insolation forest (Lesouple et al., 2021).

Although the previously detailed methods may constitute the recognition pattern methodologies with multiple applications, cases with more complicated features use the ANN. SOM has been extensively used to identify internal relationships in hydrogeochemistry. However, other ANN present good alternatives to the Kohonen map. Adaptive resonance theory 2 is an unsupervised network that classifies samples based on their memory, which makes it possible to include new samples after training, classify existing clusters, or create new ones (Fan et al., 2008). Moreover, gas neural networks can be used to preserve data topology, such as SOM, but avoids empty neurons (Du, 2010). Another neural network that can be applied to reduce dimensions and extract the most relevant information from a database is an autoencoders (Fdez-Ortiz de Vallejuelo et al., 2011). In addition, graph neural networks can represent database interdependencies and non-linear relationships (Wu et al., 2021).

Delimiting the seawater intrusion phenomenon as much as possible is of great importance because a misinterpretation can lead to incorrect decision-making regarding the management of the aquifer. As indicated in Section 4.3, hydrogeochemical and hydraulic variables could constitute relevant environmental gradients in the salinization process, although their use is

still limited. It is necessary to continue advancing in the use of these variables and understanding their significance for seawater identification. Other variables that could be used and related to seawater intrusion are those associated with microplastics and stygofauna (groundwater fauna) (Li et al., 2021; Shapouri et al., 2016). On the other hand, the variables used in water research studies are a mixture of compositional data, which are part of a whole (e.g., chemical compounds) and non-compositional data (e.g., physical variables) (Herms et al., 2021). Generally, the compositional nature of some variables is analyzed using composite plots such as the Piper diagram (Section 2.1) after applying unsupervised techniques. The study of Boente et al. (2018) stated that applying the compositional data analysis (CoDa) approach for a complementary multivariate analysis could enable the exploration of relative enrichment spots and evaluation enrichment trends, thus complementing the results when the compositional nature is not considered.

5. Conclusions

This work reviews how unsupervised learning has supported seawater intrusion pattern recognition in coastal aquifers worldwide over the last 22 years. PCA, the most frequently used DRM, enabled the identification of environmental gradients, among which the most relevant was associated with salinization and presented an average explained variance of 48 %. Meanwhile, HCA, K-means, FCM, and SOM facilitated the segmentation of samples into clusters, which were subsequently assigned to hydrochemical impacts and sources, thus delineating seawater intrusion. The application of the reviewed techniques to the La Paz case study enabled the visualization and comparison of their performances. It was shown that redundant variables, such as TH, EC, and TDS, do not provide new information and further complicate the analysis. On the other hand, the clustering methods and SOM applications did not show relevant changes in cluster patterns. However, HCA and SOM present advantages for outlier identification, while FCM represents transitional zones because it can assign samples to multiple clusters. Although the application of these methods has supported the identification of seawater intrusion, new techniques with greater precision for differentiating sources must be adopted.

In addition to advancing pattern recognition techniques, the need to complement studies with other approaches, such as flow models and geophysical methods, was shown. By collecting information of a different nature, the phenomena of seawater intrusion can be better delimited spatially and temporally, which enables appropriate management. Furthermore, this review supports the idea that both journals and authors are responsible for uploading the necessary information to reproduce and validate studies. In addition to ensuring that access to information makes the studies reproducible, it also favors the socialization of information of general interest for water resources management.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.160933.

CRediT authorship contribution statement

Herewith we state that all authors participated in the development of the manuscript. In the following an accurate and detailed description of their diverse contributions to the work:

Christian Narvaez-Montoya: Initial idea and conceptualization, Data curation, Data analysis, Investigation, Writing-Original draft preparation, Visualization.

Jürgen Mahlknecht: Supervision, Funding acquisition, Project administration, Reviewing and editing.

Juan Antonio Torres-Martínez: Validation, Reviewing and editing.

Abrahan Mora: Validation, Reviewing and editing.

Guillaume Bertrand: Validation, Reviewing and editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors state that the submission of the manuscript implies that the work described has not been published previously, that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright holder.

Acknowledgments

We gratefully thank Consejo Nacional de Ciencia y Tecnología (CONACyT) (CVU: 1014283) and Tecnologico de Monterrey for providing scholarship and tuition to the lead author of the Ph.D. degree program. Symbols for the graphical abstract were obtained from the Integration and Application Network of the University of Maryland Center for Environmental Science (ian.umces.edu/symbols/).

References

Abiodun, O.I., Kiru, M.U., Jantan, A., Omolara, A.E., Dada, K.V., Umar, A.M., et al., 2019. Comprehensive review of artificial neural network applications to pattern recognition. IEEE Access 7, 158820–158846. https://doi.org/10.1109/ACCESS.2019.2945545.

Abu-alnaeem, M.F., Yusoff, I., Ng, T.F., 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. https://doi.org/10.1016/j.scitotenv.2017.09.320.

Alameddine, G.R.I., Fadel, M.El., 2021. Management of saltwater intrusion in data - scarce coastal aquifers: impacts of seasonality, water deficit, and land use. Water Resour. Manag., 5139–5153 https://doi.org/10.1007/s11269-021-02991-4.

Alfarrah, N., Walraevens, K., 2018. Groundwater overexploitation and seawater intrusion in coastal areas of arid and semi-arid regions. Water 10 (2), 143. https://doi.org/10.3390/w10020143.

Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., Ioannidis, J.P.A., 2011. Public availability of published research data in high-impact journals. PLoS ONE 6 (9), e24357. https://doi.org/10.1371/journal.pone.0024357.

Amanambu, A.C., Obarein, O.A., Mossa, J., Li, L., Ayeni, S.S., Balogun, O., et al., 2020. Groundwater system and climate change: present status and future considerations. J. Hydrol. 589 (December 2019), 125163. https://doi.org/10.1016/j.jhydrol.2020.125163.

Andersson, A., 2021. Mechanisms for log normal concentration distributions in the environment. Sci. Rep. 11 (1), 16418. https://doi.org/10.1038/s41598-021-96010-6.

Aris, A.Z., Praveena, S.M., Abdullah, M.H., Radojevic, M., 2012. Statistical approaches and hydrochemical modelling of groundwater system in a small tropical island. J. Hydroinf. 14 (1), 206–220. https://doi.org/10.2166/hydro.2011.072.

Askri, B., Ahmed, A.T., Bouhlila, R., 2022. Origins and processes of groundwater salinisation in Barka coastal aquifer, Sultanate of Oman. Phys.Chem.Earth Parts A/B/C 126, 103116. https://doi.org/10.1016/j.pce.2022.103116.

Audigier, V., Husson, F., Josse, J., 2017. MIMCA: multiple imputation for categorical variables with multiple correspondence analysis. Stat. Comput. 27 (2), 501–518. https://doi.org/10.1007/s11222-016-9635-4.

Awaleh, M.O., Boschetti, T., Soubaneh, Y.D., Kim, Y., Baudron, P., Kawalieh, A.D., et al., 2018. Geochemical, multi-isotopic studies and geothermal potential evaluation of the complex Djibouti volcanic aquifer (republic of Djibouti). Appl. Geochem. 97 (August), 301–321. https://doi.org/10.1016/j.apgeochem.2018.07.019.

Ayesha, S., Hanif, M.K., Talib, R., 2020. Overview and comparative study of dimensionality reduction techniques for high dimensional data. Inform.Fusion 59 (January), 44–58. https://doi.org/10.1016/j.inffus.2020.01.005.

Berry, M.W., Azlinah, M., Wah Yap, B., 2020. Supervised and Unsupervised Learning for Data Science. Springer Nature Switzerland, Switzerland https://doi.org/10.1007/978-3-030-22475-2.

Bertrand, G., Petelet-Giraud, E., Cary, L., Hirata, R., Montenegro, S., Paiva, A., et al., 2022. Delineating groundwater contamination risks in southern coastal metropoles through implementation of geochemical and socio-environmental data in decision-tree and geographical information system. Water Res. 209, 117877. https://doi.org/10.1016/j.watres.2021.117877.

Björklund, M., 2019. Be careful with your principal components. Evolution 73 (10), 2151–2158. https://doi.org/10.1111/evo.13835.

Boente, C., Albuquerque, M.T.D., Fernández-Braña, A., Gerassis, S., Sierra, C., Gallego, J.R., 2018. Combining raw and compositional data to determine the spatial patterns of potentially toxic elements in soils. Sci. Total Environ. 631–632, 1117–1126. https://doi.org/10.1016/j.scitotenv.2018.03.048.

Bouguettaya, A., Yu, Q., Liu, X., Zhou, X., Song, A., 2015. Efficient agglomerative hierarchical clustering. Expert Syst. Appl. 42 (5), 2785–2797. https://doi.org/10.1016/j.eswa.2014.09.054.

Boyd, C.E., Tucker, C.S., Somridhivej, B., 2016. Alkalinity and hardness: critical but elusive concepts in aquaculture. J. World Aquacult. Soc. 47 (1), 6–41. https://doi.org/10.1111/jwas.12241.

Braeken, J., 2017. An empirical Kaiser criterion. Psychol. Methods 22 (3), 450–466. https://doi.org/10.1037/met0000074.

Bu, J., Liu, W., Pan, Z., Ling, K., 2020. Comparative study of hydrochemical classification based on different hierarchical cluster analysis methods. Int. J. Environ. Res. Public Health 17 (24). https://doi.org/10.3390/ijerph17249515.

Burrell, A.L., Evans, J.P., De Kauwe, M.G., 2020. Anthropogenic climate change has driven over 5 million km2 of drylands towards desertification. Nat. Commun. 11 (1), 3853. https://doi.org/10.1038/s41467-020-17710-7.

Busico, G., Cuoco, E., Kazakis, N., Colombani, N., Mastrocicco, M., Tedesco, D., Voudouris, K., 2018. Multivariate statistical analysis to characterize/discriminate between anthropogenic and geogenic trace elements occurrence in the Campania Plain, Southern Italy. Environ. Pollut. 234, 260–269. https://doi.org/10.1016/j.envpol.2017.11.053.

Calabrese, B., 2019. Data reduction. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 480–485 https://doi.org/10.1016/B978-0-12-809633-8.20460-3.

Cao, T., Han, D., Song, X., 2021. Past, present, and future of global seawater intrusion research: a bibliometric analysis. J. Hydrol. 603 (PA), 126844. https://doi.org/10.1016/j.jhydrol.2021.126844.

Carrera, J., Hidalgo, J.J., Slooten, L.J., Vázquez-Suñé, E., 2010. Computational and conceptual issues in the calibration of seawater intrusion models. Hydrogeol. J. 18 (1), 131–145. https://doi.org/10.1007/s10040-009-0524-1.

Celestino, A.E.M., Cruz, D.A.M., Sánchez, E.M.O., Reyes, F.G., Soto, D.V., 2018. Groundwater quality assessment: an improved approach to K-means clustering, principal component analysis and spatial analysis: a case study. Water (Switzerland) 10 (4), 1–21. https://doi.org/10.3390/w10040437.

Céréghino, R., Park, Y.-S., 2009. Review of the self-organizing map (SOM) approach in water resources: commentary. Environ. Model Softw. 24 (8), 945–947. https://doi.org/10.1016/j.envsoft.2009.01.008.

Chadha, D.K., 1999. A proposed new diagram for geochemical classification of natural waters and interpretation of chemical data. Hydrogeol. J. 7 (5), 431–439. https://doi.org/10.1007/s100400050216.

Chandrajith, R., Diyabalanage, S., Premathilake, K.M., Hanke, C., van Geldern, R., Barth, J.A.C., 2016. Controls of evaporative irrigation return flows in comparison to seawater intrusion in coastal karstic aquifers in northern Sri Lanka: evidence from solutes and stable isotopes. Sci. Total Environ. 548–549, 421–428. https://doi.org/10.1016/j.scitotenv.2016.01.050.

Costall, A.R., Harris, B.D., Teo, B., Schaa, R., Wagner, F.M., Pigois, J.P., 2020. Groundwater throughflow and seawater intrusion in high quality coastal aquifers. Sci. Rep. 10 (1), 9866. https://doi.org/10.1038/s41598-020-66516-6.

Cui, D., Liang, S., Wang, D., 2021. Observed and projected changes in global climate zones based on Köppen climate classification. Wiley Interdiscip. Rev. Clim. Chang. 12 (3), 1–28. https://doi.org/10.1002/wcc.701.

Damonte, G., Boelens, R., 2019. Hydrosocial territories, agro-export and water scarcity: capitalist territorial transformations and water governance in Peru's coastal valleys. Water Int. 44 (2), 206–223. https://doi.org/10.1080/02508060.2018.1556869.

Denis, D., 2020. Univariate, Bivariate, and Multivariate Statistics Using R: Quantitative Tools for Data Analysis and Data Science. John Wiley & Sons, Boston, MA.

Díaz-Alcaide, S., Martínez-Santos, P., 2019. Review: advances in groundwater potential mapping. Hydrogeol. J. 27 (7), 2307–2324. https://doi.org/10.1007/s10040-019-02001-3.

Du, K., 2010. Clustering: A Neural Network Approach. 23, pp. 89–107. https://doi.org/10.1016/j.neunet.2009.08.007.

El Yaouti, F., El Mandour, A., Khattach, D., Benavente, J., Kaufmann, O., 2009. Salinization processes in the unconfined aquifer of Bou-Areg (NE Morocco): a geostatistical, geochemical, and tomographic study. Appl. Geochem. 24 (1), 16–31. https://doi.org/10.1016/j.apgeochem.2008.10.005.

Enemark, T., Peeters, L.J.M., Mallants, D., Batelaan, O., 2019. Hydrogeological conceptual model building and testing: a review. J. Hydrol. 569, 310–329. https://doi.org/10.1016/j.jhydrol.2018.12.007.

Ez-zaouy, Y., Bouchaou, L., Saad, A., Hssaisoune, M., Brouziyne, Y., Dhiba, D., Chehbouni, A., 2022. Morocco's coastal aquifers: recent observations, evolution and perspectives towards sustainability. Environ. Pollut. 293 (November 2021), 118498. https://doi.org/10.1016/j.envpol.2021.118498.

Fan, J., Song, Y., Fei, M., 2008. Neurocomputing ART2 Neural Network Interacting With Environment. 72, pp. 170–176. https://doi.org/10.1016/j.neucom.2008.02.026.

Fdez-Ortiz de Vallejuelo, S., Arana, G., de Diego, A., Madariaga, J.M., 2011. Pattern recognition and classification of sediments according to their metal content using chemometric tools. A case study: the estuary of Nerbioi-Ibaizabal River (Bilbao, Basque Country). Chemosphere 85 (8), 1347–1352. https://doi.org/10.1016/j.chemosphere.2011.07.054.

Ferguson, G., Gleeson, T., 2012. Vulnerability of coastal aquifers to groundwater use and climate change. Nat. Clim. Chang. 2 (5), 342–345. https://doi.org/10.1038/nclimate1413.

Fraiman, R., Justel, A., Svarc, M., 2008. Selection of variables for cluster analysis and classification rules. J. Am. Stat. Assoc. 103 (483), 1294–1303. https://doi.org/10.1198/016214508000000544.

Fürnkranz, J., Chan, P.K., Craw, S., Sammut, C., Uther, W., Ratnaparkhi, A., et al., 2011. Model-based clustering. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning. Springer, US, Boston, MA, pp. 686–689 https://doi.org/10.1007/978-0-387-30164-8_554.

Gaillardet, J., Dupré, B., Louvat, P., Allègre, C.J., 1999. Global silicate weathering and CO2 consumption rates deduced from the chemistry of large rivers. Chem. Geol. 159 (1–4), 3–30. https://doi.org/10.1016/S0009-2541(99)00031-5.

Galazoulas, E.C., Petalas, C.P., 2014. Application of multivariate statistical procedures on major ions and trace elements in a multilayered coastal aquifer: the case of the south Rhodope coastal aquifer. Environ. Earth Sci. 72 (10), 4191–4205. https://doi.org/10.1007/s12665-014-3315-5.

Gilabert-Alarcón, C., Daesslé, L.W., Salgado-Méndez, S.O., Pérez-Flores, M.A., Knöller, K., Kretzschmar, T.G., Stumpp, C., 2018. Effects of reclaimed water discharge in the maneadero coastal aquifer, Baja California,Mexico. Appl. Geochem. 92 (March), 121–139. https://doi.org/10.1016/j.apgeochem.2018.03.006.

Giménez-Forcada, E., 2010. Dynamic of sea water Interface using hydrochemical facies evolution diagram. Ground Water 48 (2), 212–216. https://doi.org/10.1111/j.1745-6584.2009.00649.x.

Gokul, M.S., Dahms, H.U., Muthukumar, K., Henciya, S., Kaviarasan, T., James, R.A., 2019. Multivariate drug resistance and microbial risk assessment in tropical coastal communities. Hum. Ecol. Risk. Assess. 25 (5), 1073–1095. https://doi.org/10.1080/10807039.2018.1447361.

Govett, G.J.S., Goodfellow, W.D., Chapman, R.P., Chork, C.Y., 1975. Exploration geochemistry—distribution of elements and recognition of anomalies. J. Int. Assoc. Math. Geol. 7 (5–6), 415–446. https://doi.org/10.1007/BF02080498.

Gredilla, A., De Vall, S.F., Amigo, J.M., Diego, A.De, Madariaga, J.M., 2013. Unsupervised pattern-recognition techniques to investigate metal pollution in estuaries. Trends Anal. Chem. 46. https://doi.org/10.1016/j.trac.2013.01.014.

Greenacre, M., Pardo, R., 2011. Multiple correspondence analysis of a subset of response categories. SSRN Electron. J. https://doi.org/10.2139/ssrn.847647.

Güler, C., Thyne, G.D., McCray, J.E., Turner, K.A., 2002. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrogeol. J. 10 (4), 455–474. https://doi.org/10.1007/s10040-002-0196-6.

Güler, C., Kurt, M.A., 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. https://doi.org/10.1016/j.jhydrol.2011.11.021.

Güner, E.D., Cekim, H.O., Seçkin, G., 2021. Determination of water quality assessment in wells of the Göksu Plains using multivariate statistical techniques. Environ. Forensic 22 (1–2), 172–188. https://doi.org/10.1080/15275922.2020.1834025.

Ha, Q.K., Tran Ngoc, T.D., Le Vo, P., Nguyen, H.Q., Dang, D.H., 2022. Groundwater in Southern Vietnam: understanding geochemical processes to better preserve the critical water resource. Sci. Total Environ. 807, 151345. https://doi.org/10.1016/j.scitotenv.2021.151345.

Hahsler, M., Piekenbrock, M., Doran, D., 2019. dbscan: fast density-based clustering with R. J. Stat. Softw. 91 (1). https://doi.org/10.18637/jss.v091.i01.

Hajji, S., Nasri, G., Boughariou, E., Bahloul, M., Allouche, N., Bouri, S., 2020. Towards understanding groundwater quality using hydrochemical and statistical approaches: case of shallow aquifer of Mahdia-Ksour Essaf (Sahel of Tunisia). Environ. Sci. Pollut. Res. 27 (5), 5251–5265. https://doi.org/10.1007/s11356-019-06982-2.

Hasan, M.N., Siddique, M.A.B., Reza, A.H.M.S., Khan, R., Akbor, M.A., Elius, I.Bin, et al., 2021. Vulnerability assessment of seawater intrusion in coastal aquifers of southern Bangladesh: water quality appraisals. Environ.Nanotechnol.Monit.Manag. 16 (February), 100498. https://doi.org/10.1016/j.enmm.2021.100498.

Herms, I., Jódar, J., Soler, A., Lambán, L.J., Custodio, E., Núñez, J.A., et al., 2021. Evaluation of natural background levels of high mountain karst aquifers in complex hydrogeological settings. A Gaussian mixture model approach in the Port del Comte (SE, Pyrenees) case study. Sci. Total Environ. 756, 143864. https://doi.org/10.1016/j.scitotenv.2020.143864.

Houssein, A., Elmi, W., Zghibi, A., 2017. Assessment of chemical quality of groundwater in coastal volcano- sedimentary aquifer of Djibouti, Horn of Africa. J. Afr. Earth Sci. 131, 284–300. https://doi.org/10.1016/j.jafrearsci.2017.04.010.

Hyung, T., Sang, K., Chung, Y., Senapathi, V., Sekar, S., Eldin, H., 2021. Groundwater decrease and contamination around subway tunnels in a coastal area of Busan City, Korea. Environ. Earth Sci. https://doi.org/10.1007/s12665-021-09829-7.

IBM, 2021. Factor Analysis Extraction (SPSS Statistics 25). Retrieved June 1, 2022, from. https://www.ibm.com/docs/vi/spss-statistics/25.0.0?topic=analysis-factor-extraction.

Izakian, H., Abraham, A., 2011. Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Syst. Appl. 38 (3), 1835–1838. https://doi.org/10.1016/j.eswa.2010.07.112.

Jasechko, S., Perrone, D., Seybold, H., Fan, Y., Kirchner, J.W., 2020. Groundwater level observations in 250,000 coastal US wells reveal scope of potential seawater intrusion. Nat. Commun. 11 (1), 3229. https://doi.org/10.1038/s41467-020-17038-2.

Jiao, J., Post, V., 2019. Coastal hydrogeology. Paper Knowledge. Toward a Media History of Documentsvol. 5. Cambridge University Press. https://doi.org/10.1017/9781139344142.

Jolliffe, I.T., 2002. Principal component analysis for special types of data. Principal Component Analysis. Springer-Verlag, New York, pp. 338–372 https://doi.org/10.1007/0-387-22440-8_13.

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: analysis, modelling and application. Environ. Model. Softw. 23 (7), 835–845. https://doi.org/10.1016/j.envsoft.2007.10.001.

Kano, M., Hasebe, S., Hashimoto, I., Ohno, H., 2004. Evolution of multivariate statistical process control: application of independent component analysis and external analysis. Comput. Chem. Eng. 28, 1157–1166. https://doi.org/10.1016/j.compchemeng.2003.09.011.

Kim, J.-H., Kim, R.-H., Lee, J., Cheong, T.-J., Yum, B.-W., Chang, H.-W., 2005. Multivariate statistical analysis to identify the major factors governing groundwater quality in the coastal area of Kimje,South Korea. Hydrol. Process. 19 (6), 1261–1276. https://doi.org/10.1002/hyp.5565.

Kim, R., Kim, J., Ryu, J., Koh, D., 2019. Hydrogeochemical Characteristics of Groundwater Influenced by Reclamation, Seawater Intrusion, and Land Use in the Coastal Area of Yeonggwang, Korea. 23(4), pp. 603–619. https://doi.org/10.1007/s12303-018-0065-5.

King, J.R., Jackson, D.A., 1999. Variable selection in large environmental data sets using principal components analysis. Environmetrics 10 (1), 67–77. https://doi.org/10.1002/(SICI)1099-095X(199901/02)10:1<67::AID-ENV336>3.0.CO;2-0.

Knapp, T.R., Campbell-Heider, N., 1989. Numbers of observations and variables in multivariate analyses. West. J. Nurs. Res. 11 (5), 634–641. https://doi.org/10.1177/019394598901100517.

Kohonen, T., 2001. Self-Organizing Maps. Retrieved from. http://www.springer.de/phys/.
Kumar, P.J.S., Jegathambal, P., Babu, B., Kokkat, A., James, E.J., 2020. A hydrogeochemical appraisal and multivariate statistical analysis of seawater intrusion in point calimere wetland, lower Cauvery region, India. Groundw. Sustain. Dev. 11 (October 2019), 100392. https://doi.org/10.1016/j.gsd.2020.100392.
Lall, U., Josset, L., Russo, T., 2020. A snapshot of the world's groundwater challenges. Annu. Rev. Environ. Resour. 45, 171–194. https://doi.org/10.1146/annurev-environ-102017-025800.
Larsen, F., Tran, L.V., Van Hoang, H., Tran, L.T., Christiansen, A.V., Pham, N.Q., 2017. Groundwater salinity influenced by Holocene seawater trapped in incised valleys in the Red River delta plain. Nat. Geosci. 10 (5), 376–381. https://doi.org/10.1038/ngeo2938.
Lee, T.-C., 1998. LEEGRAM: a program for normalized stiff diagrams and quantification of grouping hydrochemical data. Comput. Geosci. 24 (6), 523–529. https://doi.org/10.1016/S0098-3004(98)00073-9.
Lee, J.Y., Song, S.H., 2007. Groundwater chemistry and ionic ratios in a western coastal aquifer of Buan, Korea: implication for seawater intrusion. Geosci. J. 11 (3), 259–270. https://doi.org/10.1007/BF02913939.
Lesouple, J., Baudoin, C., Spigai, M., Tourneret, J.-Y., 2021. Generalized isolation forest for anomaly detection. Pattern Recogn. Lett. 149, 109–119. https://doi.org/10.1016/j.patrec.2021.05.022.
Li, T., Sun, G., Yang, C., Liang, K., Ma, S., Huang, L., 2018. Using self-organizing map for coastal water quality classification: towards a better understanding of patterns and processes. Sci. Total Environ. 628–629, 1446–1459. https://doi.org/10.1016/j.scitotenv.2018.02.163.
Li, C., Gao, X., Li, S., Bundschuh, J., 2020. A review of the distribution, sources, genesis, and environmental concerns of salinity in groundwater. Environ. Sci. Pollut. Res. 27 (33), 41157–41174. https://doi.org/10.1007/s11356-020-10354-6.
Li, M., Zhang, M., Rong, H., Zhang, X., He, L., Han, P., Tong, M., 2021. Transport and deposition of plastic particles in porous media during seawater intrusion and groundwater-seawater displacement processes. Sci. Total Environ. 781, 146752. https://doi.org/10.1016/j.scitotenv.2021.146752.
Liu, Q., Zhang, Z., Zhang, B., Mu, W., Zhang, H., Li, Y., Xu, N., 2021. Hydrochemical analysis and identification of open-pit mine water sources: a case study from the Dagushan iron mine in Northeast China. Sci. Rep. 11 (1), 23152. https://doi.org/10.1038/s41598-021-02609-0.
Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., et al., 2011. K-medoids clustering. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning. Springer, US, Boston, MA, pp. 564–565 https://doi.org/10.1007/978-0-387-30164-8_426.
Marandi, A., Shand, P., 2018. Groundwater chemistry and the Gibbs diagram. Appl. Geochem. 97, 209–212. https://doi.org/10.1016/j.apgeochem.2018.07.009.
Marefat, F., Saeedian, A., Mozaffari, S.H., Khanlarzadeh, N., Kardoust, A., Mirzaei, M., Fatalaki, J.A., 2019. Advancing quantitative methods in second language research. Innov. Lang. Learn. Teach. 13 (3), 299–302. https://doi.org/10.1080/17501229.2019.1566910.
Mianabadi, A., Derakhshan, H., Davary, K., Hasheminia, S.M., Hrachowitz, M., 2020. A novel idea for groundwater resource management during megadrought events. Water Resour. Manag. 34 (5), 1743–1755. https://doi.org/10.1007/s11269-020-02525-4.
Michael, H.A., Post, V.E.A., Wilson, A.M., Werner, A.D., 2017. Science, society, and the coastal groundwater squeeze. Water Resour. Res. 53 (4), 2610–2617. https://doi.org/10.1002/2017WR020851.
Minitab, 2022. Enter your data for Factor Analysis (MINITAB 18). Retrieved June 1, 2022, from https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/factor-analysis/perform-the-analysis/enter-your-data/.
Mirzavand, M., Ghasemieh, H., Sadatinejad, S.J., Bagheri, R., 2020. An overview on source, mechanism and investigation approaches in groundwater salinization studies. Int. J. Environ. Sci. Technol. 17 (4), 2463–2476. https://doi.org/10.1007/s13762-020-02647-7.
Mitra, P., Murthy, C.A., Pal, S.K., 2002. Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. 24 (3), 301–312. https://doi.org/10.1109/34.990133.
Miuigan, G.W., Cooper, M.C., 1988. A study of standardization of variables in cluster analysis. J. Classif. 204, 181–204. https://doi.org/10.1007/BF01897163.
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. & Applic. 32 (8), 3763–3775. https://doi.org/10.1007/s00521-018-3768-7.
Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 6 (7), e1000097. https://doi.org/10.1371/journal.pmed.1000097.
Mora, A., Torres-Martínez, J.A., Moreau, C., Bertrand, G., Mahlknecht, J., 2021. Mapping salinization and trace element abundance (including as and other metalloids) in the groundwater of north-central Mexico using a double-clustering approach. Water Res. 205, 117709. https://doi.org/10.1016/j.watres.2021.117709.
Moreno Merino, L., Aguilera, H., González-Jiménez, M., Díaz-Losada, E., 2021. D-Piper, a modified piper diagram to represent big sets of hydrochemical analyses. Environ. Model Softw. 138, 104979. https://doi.org/10.1016/j.envsoft.2021.104979.
Narvaez-Montoya, C., Torres-Martínez, J.A., Pino-Vargas, E., Cabrera-Olivera, F., Loge, F.J., Mahlknecht, J., 2022. Predicting adverse scenarios for a transboundary coastal aquifer system in the Atacama Desert (Peru/Chile). Sci. Total Environ. 806, 150386. https://doi.org/10.1016/j.scitotenv.2021.150386.
Naser, A.M., Unicomb, L., Doza, S., Ahmed, K.M., Rahman, M., Uddin, M.N., et al., 2017. Stepped-wedge cluster-randomised controlled trial to assess the cardiovascular health effects of a managed aquifer recharge initiative to reduce drinking water salinity in south-west coastal Bangladesh: study design and rationale. BMJ Open 7 (9), 1–11. https://doi.org/10.1136/bmjopen-2016-015205.
Nguyen, T.T., Kawamura, A., Tong, T.N., Amaguchi, H., Nakagawa, N., Gilbuena, R., Bui, D. Du., 2015. Identification of spatio-seasonal hydrogeochemical characteristics of the unconfined groundwater in the red River Delta,Vietnam. Appl. Geochem. 63, 10–21. https://doi.org/10.1016/j.apgeochem.2015.07.009.
Nogueira, G., Stigter, T.Y., Zhou, Y., Mussa, F., Juizo, D., 2019. Understanding groundwater salinization mechanisms to secure freshwater resources in the water-scarce city of Maputo, Mozambique. Sci. Total Environ. 661, 723–736. https://doi.org/10.1016/j.scitotenv.2018.12.343.
O'Hara, R., Kotze, J., 2010. Do not log-transform count data. Nature Precedings https://doi.org/10.1038/npre.2010.4136.1.
Olkowska, E., Kudłak, B., Tsakovski, S., Ruman, M., Simeonov, V., Polkowska, Z., 2014. Assessment of the water quality of Kłodnica River catchment using self-organizing maps. Sci. Total Environ. 476–477, 477–484. https://doi.org/10.1016/j.scitotenv.2014.01.044.
Olsen, R.L., Chappell, R.W., Loftis, J.C., 2012. Water quality sample collection, data treatment and results presentation for principal components analysis – literature review and Illinois River watershed case study. Water Res. 46 (9), 3110–3122. https://doi.org/10.1016/j.watres.2012.03.028.
Osiakwan, G.M., Appiah-Adjei, E.K., Kabo-Bah, A.T., Gibrilla, A., Anornu, G., 2021. Assessment of groundwater quality and the controlling factors in coastal aquifers of Ghana: an integrated statistical, geostatistical and hydrogeochemical approach. J. Afr. Earth Sci. 184 (August), 104371. https://doi.org/10.1016/j.jafrearsci.2021.104371.
Pacheco Castro, R., Pacheco Ávila, J., Ye, M., Cabrera Sansores, A., 2018. Groundwater quality: analysis of its temporal and spatial variability in a karst aquifer. Groundwater 56 (1), 62–72. https://doi.org/10.1111/gwat.12546.
Papazotos, P., Vasileiou, E., Perraki, M., 2020. Elevated groundwater concentrations of arsenic and chromium in ultramafic environments controlled by seawater intrusion, the nitrogen cycle, and anthropogenic activities: the case of the Gerania Mountains, NE Peloponnese,Greece. Appl. Geochem. (July), 104697 https://doi.org/10.1016/j.apgeochem.2020.104697.
Parizi, E., Hosseini, S.M., Ataie-Ashtiani, B., Simmons, C.T., 2019. Vulnerability mapping of coastal aquifers to seawater intrusion: review, development and application. J. Hydrol. 570 (August 2018), 555–573. https://doi.org/10.1016/j.jhydrol.2018.12.021.
Parkhurst, D.L., Appelo, C.A.J., 2013. Description of input and examples for PHREEQC version 3: a computer program for speciation, batch-reaction, one-dimensional transport, and inverse geochemical calculations. Techniques and Methods https://doi.org/10.3133/tm6A43 Reston, VA.
Poeter, E., Fan, Y., Cherry, J., Wood, W., Mackay, D., 2020. Groundwater in Our Water Cycle: Getting to Know Earth's Most Important Fresh Water Source. The Groundwater Project, Ontario, Canada https://doi.org/10.21083/978-1-7770541-1-3.
Polemio, M., Zuffianò, L.E., 2020. Review of utilization management of groundwater at risk of salinization. J. Water Resour. Plan. Manag. 146 (9), 03120002. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001278.
Rajabi, M.M., Ataie-Ashtiani, B., Simmons, C.T., 2018. Model-data interaction in groundwater studies: review of methods, applications and future directions. J. Hydrol. 567 (September), 457–477. https://doi.org/10.1016/j.jhydrol.2018.09.053.
Rajoub, B., 2020. Supervised and unsupervised learning. Biomedical Signal Processing and Artificial Intelligence in Healthcare, pp. 51–89 https://doi.org/10.1016/b978-0-12-818946-7.00003-2 (January).
Rakib, M.A., Sasaki, J., Matsuda, H., Quraishi, S.B., Mahmud, M.J., Bodrud-Doza, M., et al., 2020. Groundwater salinization and associated co-contamination risk increase severe drinking water vulnerabilities in the southwestern coast of Bangladesh. Chemosphere 246. https://doi.org/10.1016/j.chemosphere.2019.125646.
Sabarathinam, C., Bhandary, H., Ali, A., 2021. Strategies to characterize the geochemical interrelationship between coastal saline groundwater and seawater. Environ. Earth Sci. 80 (18), 642. https://doi.org/10.1007/s12665-021-09924-9.
Sae-Ju, J., Chotpantarat, S., Thitimakorn, T., 2020. Hydrochemical, geophysical and multivariate statistical investigation of the seawater intrusion in the coastal aquifer at Phetchaburi Province, Thailand. J. Asian Earth Sci. 191 (December 2018), 104165. https://doi.org/10.1016/j.jseaes.2019.104165.
Salem, Z.E.-S., Abdelrahman, K., Kováčiková, S., Badran, O.M., 2021. Use of various statistical techniques to assess the vertical and lateral change in the groundwater chemistry of quaternary aquifer in an irrigated highly populated area. J.King Saud Univ.Sci. 33 (7), 101556. https://doi.org/10.1016/j.jksus.2021.101556.
Sangadi, P., Kuppan, C., Ravinathan, P., 2022. Effect of Hydro-geochemical Processes and Saltwater Intrusion on Groundwater Quality and Irrigational Suitability Assessed by Geo-statistical Techniques in Coastal Region of eastern Andhra Pradesh, India. 175 (December 2021).
Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O.P., Tiwari, A., et al., 2017. A review of clustering techniques and developments. Neurocomputing 267, 664–681. https://doi.org/10.1016/j.neucom.2017.06.053.
Senawi, A., Wei, H.-L., Billings, S.A., 2017. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. Pattern Recogn. 67, 47–61. https://doi.org/10.1016/j.patcog.2017.01.026.
Sergeant, C.J., Starkey, E.N., Bartz, K.K., Wilson, M.H., Mueter, F.J., 2016. A practitioner's guide for exploring water quality patterns using principal components analysis and procrustes. Environ. Monit. Assess. 188 (4), 249. https://doi.org/10.1007/s10661-016-5253-z.
Shapouri, M., Cancela da Fonseca, L., Iepure, S., Stigter, T., Ribeiro, L., Silva, A., 2016. The variation of stygofauna along a gradient of salinization in a coastal aquifer. Hydrol. Res. 47 (1), 89–103. https://doi.org/10.2166/nh.2015.153.
Sharma, S., Batra, N., 2019. Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon). IEEE, pp. 568–573 https://doi.org/10.1109/COMITCon.2019.8862232.
Shi, X., Wang, Y., Jiao, J.J., Zhong, J., Wen, H., Dong, R., 2018. Assessing major factors affecting shallow groundwater geochemical evolution in a highly urbanized coastal area of

Shenzhen City, China. J. Geochem. Explor. 184, 17–27. https://doi.org/10.1016/j.gexplo.2017.10.003.
Sicilia, M.-A., García-Barriocanal, E., Sánchez-Alonso, S., 2017. Community curation in open dataset repositories: insights from Zenodo. Procedia Comput.Sci. 106, 54–60. https://doi.org/10.1016/j.procs.2017.03.009.
Souid, F., Agoubi, B., Telahigue, F., Chahlaoui, A., Kharroubi, A., 2018. Groundwater salinization and seawater intrusion tracing based on lithium concentration in the shallow aquifer of Jerba Island, southeastern Tunisia. J. Afr. Earth Sci. 138, 233–246. https://doi.org/10.1016/j.jafrearsci.2017.11.013.
Stetco, A., Zeng, X.-J., Keane, J., 2015. Fuzzy C-means + +: fuzzy C-means with effective seeding initialization. Expert Syst. Appl. 42 (21), 7541–7548. https://doi.org/10.1016/j.eswa.2015.05.014.
Strauss, T., von Maltitz, M.J., 2017. Generalising Ward's method for use with Manhattan distances. PLOS ONE 12 (1), e0168288. https://doi.org/10.1371/journal.pone.0168288.
Székely, G.J., Rizzo, M.L., 2014. Partial distance correlation with methods for dissimilarities. Ann. Stat. 42 (6). https://doi.org/10.1214/14-AOS1255.
Tahmasebi, P., Kamrava, S., Bai, T., Sahimi, M., 2020. Machine learning in geo- and environmental sciences: from small to large scale. Adv.Water Resour.J. 142. https://doi.org/10.1016/j.advwatres.2020.103619.
Tamez-Meléndez, C., Hernández-Antonio, A., Gaona-Zanella, P., Ornelas-Soto, N., Mahlknecht, J., 2016. Isotope signatures and hydrochemistry as tools in assessing groundwater occurrence and dynamics in a coastal arid aquifer. Environ. Earth Sci. 75 (830). https://doi.org/10.1007/s12665-016-5617-2.
Taşana, M., Demir, Y., Taşan, S., 2022. Groundwater quality assessment using principal component analysis and hierarchical. Water Supply 22 (3), 3431–3447 0.2166/ws.2021.390.
Tiwari, A.K., Pisciotta, A., De Maio, M., 2019. Evaluation of groundwater salinization and pollution level on Favignana Island, Italy. Environ. Pollut. 249, 969–981. https://doi.org/10.1016/j.envpol.2019.03.016.
Torres-Martínez, J.A., Mora, A., Mahlknecht, J., Kaown, D., Barceló, D., 2021. Determining nitrate and sulfate pollution sources and transformations in a coastal aquifer impacted by seawater intrusion—a multi-isotopic approach combined with self-organizing maps and a Bayesian mixing model. J. Hazard. Mater. 417, 126103. https://doi.org/10.1016/j.jhazmat.2021.126103.
Tully, K., Gedan, K., Epanchin-Niell, R., Strong, A., Bernhardt, E.S., BenDor, T., et al., 2019. The invisible flood: the chemistry, ecology, and social implications of coastal saltwater intrusion. Bioscience 69 (5), 368–378. https://doi.org/10.1093/biosci/biz027.
USGS, 2018. Preparations for Water Sampling. Geological Survey Techniques and Methods. https://doi.org/10.3133/tm9A1 book 9, chap. A1, 42 p. Reston.
Vaux, H., 2011. Groundwater under stress: the importance of management. Environ. Earth Sci. 62 (1), 19–23. https://doi.org/10.1007/s12665-010-0490-x.
Wang, H., Yang, Q., Liang, J., 2022. Interpreting the salinization and hydrogeochemical characteristics of groundwater in Dongshan Island,China. Mar. Pollut. Bull. 178, 113634. https://doi.org/10.1016/j.marpolbul.2022.113634.
Ward, J.H., 1963. Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. 58 (301), 236–244. https://doi.org/10.1080/01621459.1963.10500845.
Wehrens, R., Kruisselbrink, J., 2018. Flexible self-organizing maps in kohonen 3.0. J. Stat. Softw. 87 (7). https://doi.org/10.18637/jss.v087.i07.
Werner, A.D., Bakker, M., Post, V.E.A., Vandenbohede, A., Lu, C., Ataie-Ashtiani, B., et al., 2013. Seawater intrusion processes, investigation and management: recent advances and future challenges. Adv. Water Resour. 51, 3–26. https://doi.org/10.1016/j.advwatres.2012.03.004.
WHO, 2011. Guidelines for Drinking-water Quality. Fourth. Gutenberg, Malta.
Willett, P., Barnard, J.M., Downs, G.M., 1998. Chemical similarity searching. J. Chem. Inf. Comput. Sci. 38 (6), 983–996. https://doi.org/10.1021/ci9800211.
Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P.S., 2021. A comprehensive survey on graph neural networks. IEEE Trans.Neural Netw.Learn.Syst. 32 (1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386.
Wunderlin, D.A., María Del Pilar, D., María Valeria, A., Fabiana, P.S., Cecilia, H.A., De Los, María, Ángeles, B., 2001. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquía River basin (Córdoba-Argentina). Water Res. 35 (12), 2881–2894. https://doi.org/10.1016/S0043-1354(00)00592-3.
Yik, C., Harun, M., Mangala, S., Hawa, A., Yahaya, B., 2012. Delineation of temporal variability and governing factors influencing the spatial variability of shallow groundwater chemistry in a tropical sedimentary island. J. Hydrol. 432–433, 26–42. https://doi.org/10.1016/j.jhydrol.2012.02.015.
Yin, Z., Luo, Q., Wu, J., Xu, S., Wu, J., 2021. Identification of the long-term variations of groundwater and their governing factors based on hydrochemical and isotopic data in a river basin. J. Hydrol. 592 (October 2020), 125604. https://doi.org/10.1016/j.jhydrol.2020.125604.
Yuan, C., Yang, H., 2019. Research on K-value selection method of K-means clustering algorithm. J 2 (2), 226–235. https://doi.org/10.3390/j2020016.
Zhu, H., Zhou, J., Feng, H., Liu, H., Zhu, H., Liu, Z., et al., 2020. Influences of natural and anthropogenic processes on the groundwater quality in the Dagujia River basin in Yantai, China. J.Water Supply: Res. Technol. - AQUA 69 (2), 184–196. https://doi.org/10.2166/aqua.2019.113.