JEE Journal of Ecological Engineering

Journal of Ecological Engineering 2022, 23(8), 327–340
https://doi.org/10.12911/22998993/151073
ISSN 2299–8993, License CC-BY 4.0

Received: 2022.05.30
Accepted: 2022.06.17
Published: 2022.07.01

Use of Multivariate Statistical Analysis of Hydrochemical Data for the Identification of the Geochemical Processes in the Tirana-Fushe Kuqe Alluvial Aquifer, North-Western Albania

Endri Raço[1], Arjan Beqiraj[2], Sabina Cenameri[3], Aurela Jahja[2]

[1] Mathematical Engineering Department, Faculty of Mathematics and Physics Engineering, Polytechnic University of Tirana, Sulejman Delvina St., Tirana, Albania

[2] Earth Sciences Departament, Faculty of Geology and Mining, Polytechnic University of Tirana, Street Elbasani, Tirana, Albania

[3] EU Support to Integrated Water Management in Albania Project, Street Pjeter Budi, Tirana, Albania

* Corresponding author's e-mail: e.raco@fimif.edu.al

ABSTRACT

During the research, 71 groundwater samples were collected over a 300 km^2 area of Tirana-Fushe Kuqe alluvial aquifer extension (central-western Albania) and subsequently analyzed for 11 parameters (pH, K^+, Na^+, Ca^{2+}, Mg^{2+}, HCO_3^-, Cl^-, SO_4^{2-}, NO_3^-, TH and TDS). Both geochemical conventional (Piper and Chadha diagrams) methods of groundwater classification and multivariate statistical (principal components analysis – PCA and hierarchical cluster analysis – HCA) methods were applied to the dataset to evidence the geochemical processes controlling groundwater geochemistry evaluation through the aquifer. The conventional geochemical methods revealed four (G1–G4) hydrochemical groups where the dominant group is G2 the samples of which are from unconfined to semiconfined recharge zone and the majority of them have Ca-Mg-HCO_3 groundwater. Group G3 includes the samples from the confined coastal aquifer having Na-Cl groundwater. Group G1 includes three groundwater samples of Ca-Mg-SO_4 from the central part of the aquifer, while group G4, the samples of which are spatially located between G3 and G2 zones, has Na-HCO_3 groundwater. The first four components of the PCA account for 85.35% of the total variance. Component PC1 is characterized by very high positive loadings of TH, Ca^{2+}, and Mg^{2+}, suggesting the importance of dissolution processes in the aquifer recharge zone. Component PC2 is characterized by very high positive loadings in Na^+, K^+, and Cl^- and moderate to high loadings of TDS, revealing the involvement of seawater intrusion and diffusion from clay layers. On the basis of their variable loadings, the first two components are defined as the "hardness" and "salinity", respectively. The HCA produced four geochemically distinct clusters, C1–C4. The samples of cluster C1 are from the coastal confined aquifer and their groundwater belongs to the Na-Cl type. The samples from cluster C2 are located in the south and east recharge areas and most of them have Ca–Mg–HCO_3 groundwater, while the samples from cluster C3, which are located in the northeastern recharge zone, have Mg-Ca–HCO_3 groundwater. Finally, cluster C4 includes two groundwater subgroups having Na-Cl-HCO_3 and Na-Mg-Cl-HCO_3 groundwater in the vicinity of cluster C1 as well as Na-HCO_3-Cl and Na-Mg-HCO_3-Cl groundwater next to cluster C2 and C3.

Keywords: Tirana-Fushe Kuqe aquifer, groundwater, multivariate analysis, principal component analysis, hierarchical cluster analysis.

INTRODUCTION

Albania is rich in both surface and groundwater; the latter is mainly related to karst and alluvial aquifers. The Tirana-Fushe Kuqe alluvial aquifer, which extends from Tirana at SE to the Mat river at NW (Figure 1), covering an area of over 300 km^2 in central-western Albania, represents not only an interesting hydrologic unit but also an important basin for drinking water supply of about 1/3 of the country population. A groundwater quantity of about 2000 l/s was pumped from the alluvial

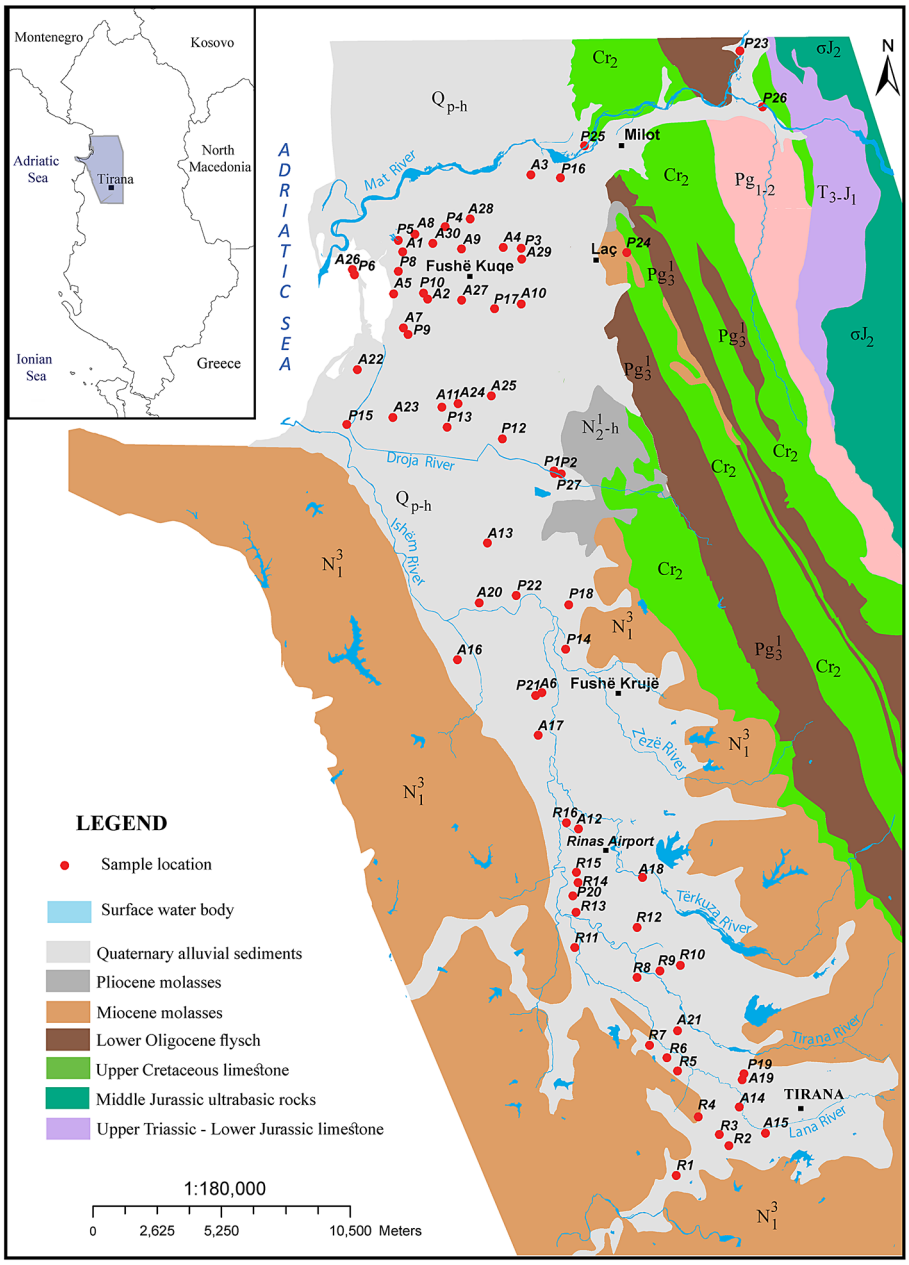

Figure 1. Geological map of the Tirana-Fushe Kuqe region (alluvial aquifer extension along with sampling sites are shown)

aquifer of Tirana-Fushe Kuqe by public pumping stations [Tartari and Dakoli, 2001] which after 5 years was increased to 2500 l/s [Eftimi et al., 2006]. Nowadays, the pumped groundwater quantity from this aquifer is over 3000 l/s due to an additional pumped water quantity of 600 l/s for the supply of Durresi coastal city.

This aquifer is among the most studied; abundant quantitative and qualitative data are generated since the year 1961 when the first drillings started. Tartaei and Dakoli [2001] constructed a hydrochemical map of the TDS and total hardness values for the whole Tirana-Fushe Kuqe aquifer extension, based on the chemical data of previous studies accomplished by ex-Albanian Hydrogeological Enterprise. Systematic hydrochemical monitoring (thirty sites), during the hydrological year 2000–2001, was realized as part of Beqiraj's post-doctoral studies at the Department of Earth Sciences (University of Rome "La Sapienza") whose preliminary results were presented by Beqiraj et al. [2002]. Within the framework of the bilateral project between the Institute of Geosciences and Georesources (Pisa-Italy) and the Faculty of Geology and Mining (Tirana-Albania) during the period 2017 to 2018, about 30 samples were collected and

analyzed for the main chemical parameters and H-2 and O-18 stable isotopes, the results of which were integrated by Cenameri [2021].

This paper aimed to bridge a gap that consists of the missing multivariate analysis of the abundant hydrochemical data of the aquifer groundwater. In addition to routine statistical estimations like basic statistics, correlations between variables, and boxplots, two well-known multivariate methods of geochemical data were used, i.e. principal components analysis (PCA) and hierarchical cluster analysis (HCA) to better evidence the factors and processes that control the formation of geochemical composition of groundwater. This was done by confronting and integrating the results of PCA and HCA with those of traditional geochemical groundwater classifications like PIPER, CHADHA, etc.

The multivariate analysis is widely used to supplement classical hydrogeochemical methods helping to solve different compositional issues of groundwater: to identify groundwater sources [Steinhorst and Williams, 1985], to evidence major water groups and factors affecting groundwater quality [Melloul et al. 1992], to identify rock–water interaction processes [Farnham, et al., 2003], to understand groundwater flow in the aquifer [Stetzenbach, et al., 2001], and to indicate the hydrogeochemical evolution of groundwater [Cloutier et al., 2008].

GEOLOGY AND HYDROGEOLOGY OF THE STUDY AREA

The Tirana-Fushë Kuqe basin extends from Tirana city in the southeast to River Mat in the northwest, covering an area of over 300 km^2. It belongs to the Tirana depression, a flat physiographic area, gently sloping toward the northwest. In the east, south, and southwest the study area is bordered by hilly terrains. To the west and northwest, it is bordered by the Adriatic Sea coastal line and the River Mat watercourse, respectively. The plain area is crossed by several rivers that drain from the eastern mountains among which River Mat represents the most important hydrographic feature of the area, followed by River Droja and Ishmi southward. The latter collects the waters of some small rivers that flow from the mountains of Dajti and Kruja such as Zeza, Terkuza, Tirana and Lana (Figure 1).

Tirana-Fushe Kuqe depression was created during the Quaternary as a consequence of earth sinking confronted with a raising up regime of the older (Neogene to Cretaceous) eastern geological formations (Figure 1). The Neogene formations that represent the basement of the alluvial sediments compose an intermountain southeast-northwest extending syncline and consist of Miocene mudstone – siltstone - sandstone intercalations, which crop out at the foot of the bordering hills. The depression was successively filled up mainly with alluvial sediments the thickness of which ranges from about 20–30 m in the South (Tirana) through 60–80 m in the Center (Fushe Kruja) up to about 200 m in the North (Fushe Kuqe) [Tartari and Dakoli, 2001]. Quaternary sediments, gravel and sand sandwiched between clay layers, are of Pleistocene age, formed after the last glacial maximum when the sea level rose from 120 below the present level and the seashore retreated from a position about 50 km offshore outside the present coastline [Kumanova et al., 2014]. The clay cover is likely to have been formed during the rapid seawater regression later in the Holocene [Kumanova, et al., 2014].

The Quaternary sediments consist of discontinuous gravel-sand intercalations with silt-clay layers. Gravel crops out at the rivers outlet in the plain, but among them the most widespread riverbed gravel outcrop is that of River Mat, which extends over an area of several kilometers. Only isolated gravel patches occur along the courses of the Droja, Zeza, and Terkuza rivers.

However, gravel sediments are widely distributed at depth in the whole valley, indicating two evident increasing tendencies of their thickness [Tartari and Dakoli, 2001] from southeast to northwest and from east to west. The aquifer changes from one layer to multilayer due to the impermeable clay intercalations within the gravel section, following the above-mentioned directions. The alluvial aquifer system is unconfined in the southeastern (Tirana) region and changes gradually versus semiconfined conditions in the central (Rinas-Fushe Kruja) region up to a complete confined aquifer in the northwestern (Fushe Kuqe) area. Following the above direction, the aquifer transmissivity increases from about 100 m^2/d to more than 8000 m^2/d [Eftimi et al., 2006]. The main groundwater direction in the southern and central part of the basin is from southeast to northwest, whereas a second important groundwater movement, in the northern part of the basin, is from the east river outlets to the Adriatic Sea in the west.

HYDROGEOCHEMISTRY

The groundwater indicates variable geochemical composition due to variable hydrogeological conditions along the aquifer extension, and influence of seawater intrusion in the coastal parts of the aquifer [Cenameri and Beqiraj, 2018, Eftimi et al., 2006; Beqiraj et al., (2002); Tartari and Dakoli, 2001]. Beqiraj et al. (2002) identified six main hydrochemical groundwater types, the distribution of which in the aquifer is closely related to its hydrogeological conditions. The Ca-HCO_3 and Ca–Mg–HCO_3 types are typical for the southern and eastern recharge areas where phreatic conditions dominate, followed westward by a mixed zone under semiconfined and confined conditions where Mg-Ca-HCO_3, Na-Mg-HCO_3-Cl, and Na-Cl-HCO_3 types are present. Finally, the Na–Cl groundwater characterizes the completely under confined conditions coastal areas of the aquifer. The chloride content in the groundwater of the southern – central areas of the basin is generally less than 20 mg/l, which is compatible with its content in the rivers crossing this area such as Lana (10 mg/l), Tirana (11.1 mg/l), Terkuza (16 mg/l) and Zeza (12 mg/l) [Eftimi et al., 2006]. There is an apparent increase, from east to west, of the chloride content in the groundwater of the northern aquifer extension, which becomes faster in its coastal area where values over 500 mg/l Cl^- are met [Cenameri and Beqiraj 2016].

METHODOLOGY

Water sampling and testing

In the period 2016–2018, 67 groundwater samples were collected, mostly from private wells and only a few from municipal wells. In addition, four samples were taken from the surface (3 from rivers Mat, Fan and Droja and one from a Laç karst spring) water. The sampling location was recorded by a portable GPS device tip Garmin - GPSMAP 62 series. The sampled wells are distributed over the whole aquifer extension area (Figure 1). The SEBA Liquid Sampler KLL-S was used for groundwater sampling from the wells. The temperature, pH and conductivity were measured on site. The pH and temperature were measured using a Hanna Testo 205 pH meter, having accuracy of ±0.02 pH/±0.4 °C and resolution of 0.01 pH/0.1°C. Conductivity was measured with a Hanan HI 99300, which, within the range 0–3999μS/cm, has accuracy of ±2% and resolution of 1μS/cm.

Polyethylene bottles, sealed with double cap, with a volume of 1.5 L, preliminarily acidified with nitric acid diluted to 1:1, were used. Bottles were rinsed 2–3 times with the sampling water before sampling. Chemical analysis was performed in Chemical Laboratory of the Albanian Geological Survey. Ca^{2+} and Mg^{2+}, hardness, and HCO_3^- were analyzed via titration; Cl^- and SO_4^{2-} by photometry; pH and electric conductivity by electrochemistry; Na^+, and K^+ were analyzed in AAS "PERKIN ELMER" in both flame technic 400 AANALYST and 900 AA model with graphite furnace.

Organizing data for multivariate statistical analysis

Multivariate statistical techniques serve as commonly approach when it comes to groundwater classification. These techniques enable grouping of groundwater samples and calculation of correlations between chemical parameters and groundwater samples [Cloutier, et al., 2008]. In this study, two multivariate methods were utilized taking advantage of the possibilities that R software provides [Ihaka et al., 1996]: Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA). Principal component analysis (PCA) is a popular strategy for reducing dimensions of dataset, thus increasing the interpretability with minimal information loss. The way PCA operates is by generating new uncorrelated variables that enable the maximization of variance [Jolliffe and Cadima, 2016; Zainol et al., 2021; Nakagawa, 2021; Zhang et al., 2021]. Basically, the problem of defining the new dataset dimensions is reduced in a matrix decomposition, with eigenvalues/vectors interpretation described in many works [Beattie et al., 2021; Torokhti and Friedland, 2009; Gewers, 2021]. HCA is a common approach to solve the problem of reducing the dimensions of the data set [Mushtaq, N. et al., 2020; Ebrahimi et al., 2021; Xu et al. 2021]. The approach is based on iteratively splitting data in smaller groups (clusters) based upon their similarity. In essence, the HCA method yields a hierarchical representation, where clusters of lower level are created by the division of clusters in the next upper level. The HCA method stops when each cluster from the lowest level accommodate an isolated feature [Xu et al., 2020]

The collected data were organized in a dataset consisting of 71 groundwater samples and 11 parameters, which include pH, K^+, Na^+, Ca^{2+}, Mg^{2+}, HCO_3^-, Cl^-, SO_4^{2-}, NO_3^-, Total Hardness (TH) and Total Dissolved Solids (TDS). In the dataset, there were 15 samples that have missing values of NO_3^- and the sample P22 which has missing values for Na^+ and Cl^-. Various statistical techniques and geochemical relationships can be utilized to estimate the missing values [Jack et al., 2002]. Because the missing data samples wells are spread out over the study area, the missing NO_3^- values were estimated by averaging the values of nearest sampling wells. Because the water of the well P22 is typically brackish and with known salinity (TDS = 10815.93 mg/l), chloride content was calculated using Vernier equation: salinity (ppt) = 0.0018066 × Cl^- (mg/L), while the Na^+ content was estimated from the electro-neutrality of the water. At the end, a dataset of 71 samples and 11 parameters was used for the multivariate statistical analysis. Before proceeding with PCA and HCA methods, an in-depth treatment of dataset was applied, to ensure that it was suitable for multivariate statistical analyses. Firstly, the hardness values were converted from German degrees to mg/l, satisfying a request of multivariate (PCA and HCA) analysis that all variables to be reported in the same concentration unit [Rock, 1988].

DESCRIPTIVE STATISTICS

Descriptive statistics of geochemical parameters are displayed in Table 1, where it is noticeable that the standard deviation values are higher than mean and mean larger than median for most elements. Combined with skewness values boxplots for each variable of the dataset plotted using R library tidyverse (Wickham et al., 2019), it is an indicator of positive skewness and large variable outliers, which is a generally characteristic for distribution of geochemical data [Rock, 1988].

Normal distribution

Before performing the PCA procedure, it is a good practice to check first whether or not all variables follow normal distribution. In fact, most of the variables seem to have right-skewed distributions, a commonly observed characteristic in geochemical populations [Grunsky, 2010]. Reimann and Filzmoser [2000] also showed that in regional geochemical datasets, it is practically impossible to meet normal distribution.

Outliers

Most of variables seem to have right-skwed distribution, commonly observed characteristic in geochemical populations [Grunsky, 2010], which is often caused by outliers [Reimann and Filzmoser]. The outliers that were detected using box plot inner fence rule [Reimann et al, 2005], should be removed prior to entering a multivariate analysis, but since their removal would lead to a substantial reduction of data rows, the authors preferred to keep all data for our analysis, as proposed by [Grünfeld, 2005; Pison, 2003].

Data transformation

To reduce skewness of data, inverse hyperbolic sine function was applied. The authors

Table 1. Descriptive statistics of the chemical parameters for 71 water samples (TH = Total Hardness)

Parameters	Units	Minimum	Maximum	Median	Average	Standard deviation	Skewness
pH		6.26	8.95	7.52	7.63	0.59	0.09
K^+	mg/l	0.53	75.1	2.30	4.61	9.56	6.28
Na^+	mg/l	4.51	2641.61	34.00	135.88	342.12	5.97
Ca^{2+}	mg/l	7.58	460.92	62.80	73.71	65.68	2.98
Mg^{2+}	mg/l	6.75	644.48	26.10	39.00	76.57	7.33
HCO_3^-	mg/l	24.4	1265.75	330.00	326.63	175.58	2.11
Cl^-	mg/l	4.9	6008.85	24.85	191.22	738.16	7.27
SO_4^{2-}	mg/l	3.5	1502.8	51.20	89.04	203.73	6.06
NO_3^-	mg/l	0.1	182.4	3.72	9.15	22.58	6.72
TH	mg/l	52.24	3786.59	347.99	443.17	492.25	4.87
TDS	mg/l	117.99	10815.93	550.24	827.51	1325.87	6.40

proceeded with Arsinh function, instead of commonly used logarithmic transformation, in order to achieve better treatment of small and large values. Another reason for not choosing logarithmic approach is because this transformation does not take in consideration compositional nature of the geochemical dataset [Filzmoser, 2009]. Data transformation was made possible using best Normalize library [Peterson, 2020]. After the data transformation, the skewness was reduced and the variables follow a normal or nearly normal distribution (frequency diagrams not shown), thus being suitable for further multivariate analyses.

Correlation

Estimation of correlation coefficients can raise problems when it comes to geochemical datasets. [Gardner and Neufeld, 2013]. One of the most discussed problems for long time is the problem of spurious correlation [Buccianti, 2014; Aitchison, 1982]. Nevertheless, correlation analysis remains relevant when it comes to exploration of geochemical data due to its simplicity and consistency. The authors preferred to go with non-parametric Spearman correlation matrix because of the data limitations to be used for Pearson's r calculation [Cooksey, 2015]. Table 2, plotted using GGally library [Schloerke et al., 2020], illustrates Spearman coefficients calculated on dataset values. Na^+ and Cl^- have strong correlation with each other and both have moderate positive correlations with TDS, indicating saline water. Ca^{2+} and Mg^{2+} show a moderate relation, but high positive correlations with TH, as it might be expected. The high correlation of Ca^{2+} and low correlation of Mg^{2+} with HCO_3^- is an indication of dissolution and cation exchange processes, respectively. The low negative correlation between Ca^{2+} and Na^+ confirms that the cation exchange mostly influenced the relations of these individual cations.

Before stepping to multivariate analysis of the dataset the authors needed to check if the dataset met statistical requirements [Hair, 2019]. First, Kaiser-Meyer-Olkin (KMO) was check as a measure of overall adequacy. A minimal value of 0.5 (Bad) is required as a rule of thumb to consider items tolerable for performing factor analysis [Hadi et al., 2016]. For the parameter's dataset, KMO calculated using psych library [Revelle, 2021] has a value of 0.63, which means the correlation can be considered acceptable for proceeding multivariate analysis.

The next step was to check communality meaning the amount of variance shared among dataset parameters. Communalities take the values between 0 and 1. The values nearer to 1 imply variance is being represented by the factor. It is recommended that the items with cutoff value less than 0.5 should be removed prior starting with multivariate analysis methods [Hair, 2019; Hogarty et al., 2005; Yanai and Ichikawa, 1990]. Table 3 demonstrates that communalities for each parameter are reasonable for the multivariate analysis approach.

RESULTS AND DISCUSSION

Principal component analysis

After the preparatory stage, principal component analysis (PCA) was applied in the transformed geochemical dataset, aiming to reduce

Table 2. Spearman correlation coefficient between the major variables of groundwater composition

Variable	pH	K^+	Na^+	Ca^{2+}	Mg^{2+}	HCO_3^-	Cl^-	SO_4^{2-}	NO_3^-	TH	TDS
pH	1.000	-0.016	0.164	-.642**	-.238*	-.639**	0.095	-.253*	-.778**	-.406**	-.373**
K^+	-0.016	1.000	.615**	-0.105	-0.037	-0.030	.559**	0.192	0.120	0.006	.300*
Na^+	0.164	.615**	1.000	-0.177	0.073	-0.020	.838**	.398**	0.018	0.006	.563**
Ca^{2+}	-.642**	-0.105	-0.177	1.000	.563**	.701**	0.050	.342**	.514**	.845**	.529**
Mg^{2+}	-.238*	-0.037	0.073	.563**	1.000	.331**	0.124	.433**	0.194	.755**	.433**
HCO_3^-	-.639**	-0.030	-0.020	.701**	.331**	1.000	0.029	.307**	.497**	.585**	.509**
Cl^-	0.095	.559**	.838**	0.050	0.124	0.029	1.000	.435**	0.082	0.201	.611**
SO_4^{2-}	-.253*	0.192	.398**	.342**	.433**	.307**	.435**	1.000	.280*	.435**	.590**
NO_3^-	-.778**	0.120	0.018	.514**	0.194	.497**	0.082	.280*	1.000	.329**	.408**
TH	-.406**	0.006	0.006	.845**	.755**	.585**	0.201	.435**	.329**	1.000	.479**
TDS	-.373**	.300*	.563**	.529**	.433**	.509**	.611**	.590**	.408**	.479**	1.000

* p-value < 0.05; **p-value < 0.01.

Table 3. Variable communalities

Variable	Initial	Extraction
pH	1.000	0.874
K^+	1.000	0.622
Na^+	1.000	0.929
Ca^{2+}	1.000	0.940
Mg^{2+}	1.000	0.822
HCO_3^-	1.000	0.587
Cl^-	1.000	0.832
SO_4^{2-}	1.000	0.462
NO_3^-	1.000	0.822
TH	1.000	0.902
TDS	1.000	0.902

the dimensions. The reasons for using PCA over many factor analysis methods when it comes to geochemical data are explained in [Reimann, et al., 2022]. The PCA method consists in transforming original dataset to a new set of non-correlated parameters, which preserve the most variance of original intercorrelated parameters. As a rule, new created components are ordered in such way that the precedent component responds for more variance comparing to the component next in order [Reimann, et al., 2022]. In many studies, the components taken into consideration using eigenvalues criterion, would explain at least one variable's variability, if corresponding eigenvalue is greater than or equal to 1 [Reimann, et al., 2022; Basilevsky, 1994]. In the considered case, based on scree plot results (Figure 2) and individual characteristics of some of the sampling sites, it was concluded to take into consideration the first four principal components, thus placing a cut-off threshold for explained variance over 80%.

Varimax orthogonal rotation [Allen, 2017] was selected as a preferred method to obtain maximal variance out of first four principal components, when working with geochemical datasets [Reimann, et al., 2022]. For each component, eigenvalue and variance are shown in Table 4:

Table 5 shows loadings for the first four components, which account for 85.35% of the total variance (Table 4). The first two components (PC1 and PC2) explain 66.09% of the variance, bringing up the majority of the variance of the dataset. The two other extracted components (PC3 and PC4) are less important. Component PC1 explains the greatest (40.35%) amount of the overall variance, and is labeled by very high positive loadings of TH, Ca^{2+}, Mg^{2+} and moderate to high loadings of HCO_3^- and TDS (Table 5), suggesting the dissolution processes of carbonate rocks prevailed in the aquifer recharge area. This factor highlights the contribution of Ca^{2+}, Mg^{2+} and HCO_3^- to TH and as such it has been identified as the temporary hardness of the groundwater. Component PC2 explains 25.74% amount of the overall variance and is labeled by very high positive loadings in Na^+, K^+ and Cl^- and moderate to high loadings of TDS (Table 5). This factor indicates that the highest values of TDS in groundwater are closely related to elevated concentrations of Na^+ and Cl^-; thus, it accounts for the spatial salinity of the groundwater affected by seawater intrusion [Cenameri, et al., 2016] and/or by diffusion from intercalated clay layers [Jack et al., 2013]. Components PC3 and PC4 are characterized by high positive loadings of

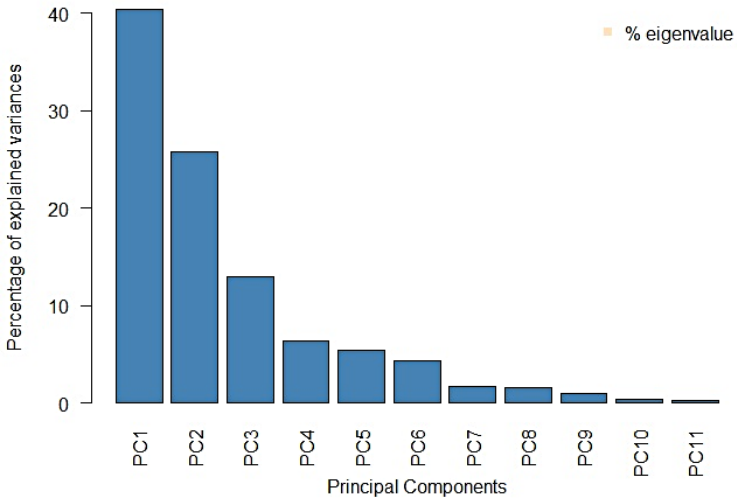

Figure 2. Scree plot

Table 4. Eigenvalues and variances

Components	Eigenvalue	Variance percent	Cumulative variance percent
PC1	4.43815310	40.3468463	40.34685
PC2	2.83187057	25.7442779	66.09112
PC3	1.42342415	12.9402196	79.03134
PC4	0.69516914	6.3197194	85.35106
PC5	0.59917261	5.4470238	90.79809
PC6	0.47086495	4.2805905	95.07868
PC7	0.18938914	1.7217195	96.80040
PC8	0.17399745	1.5817950	98.38219
PC9	0.11067478	1.0061344	99.38833
PC10	0.04300325	0.3909387	99.77927
PC11	0.02428084	0.2207349	100.00000

Table 5. PCA factor loadings of all variables

Variable	PC1	PC2	PC3	PC4
pH	-0.2288	0.1374	-0.8857	-0.1428
K^+	-0.1131	**0.8005**	0.2122	0.03293
Na^+	0.05854	**0.9077**	-0.1635	0.2736
Ca^{2+}	**0.872**	-0.1076	0.4116	0.05086
Mg^{2+}	**0.8565**	0.295	-0.1156	0.006588
HCO_3^-	**0.6511**	-0.01998	0.3069	0.3285
Cl^-	0.2382	**0.8752**	-0.2153	0.05676
SO_4^{2-}	0.1373	0.3121	0.1077	**0.8883**
NO_3^-	0.1243	0.06295	**0.9276**	0.004445
TH	**0.9342**	0.1064	0.1445	0.04926
TDS	**0.5579**	**0.6927**	0.05385	0.33

NO_3^- and SO_4^{2-}, respectively, and together account for 19.26% of the overall variance.

Hierarchical cluster analysis

The next step of the research was estimating the similarity between samples, using the Hierarchical Cluster Analysis (HCA) method [Everitt, 1974; Rousseeuw, 1990; Hastie, 2009]. The dendrogram in Figure 3 demonstrates the clustering based on the factor scores, produced previously with Principal Component Analysis. To calculate the dissimilarity between the variables, average linkage was used based on Euclidean distance [Xu, N. et al., 2021, Romesburg, 2004], while the distance between clusters was estimated by using the 'Ward.D2' method, which uses squares of dissimilarities before clustering [Legendre, 2014; Cavanaugh, 2018]. Ward's method performed better in building more or less homogenous clusters that are geochemically distinct from each other, against to other methods. In the dendrogram of the considered hydrochemical data, four distinct clusters are evidenced. The cluster C1 (excluding sample P22, which is not in coastal zone) has high salinity (average TDS = 1871.66 ± 672.70 mg/l) and is hydrochemically dominated by the Na-Cl water type, which is typical for the groundwater of aquifer coastal area. The cluster C2 is the largest, including 42 samples (or 59%) located in the south, southeast and east area of the aquifer extension. The groundwaters of this cluster have moderate salinity (average TDS = 549.86 ± 187.41 mg/l) and are mostly hard to very hard (average TH = 24.93 ± 11.92 ºdH). This group is mainly dominated by Ca-Mg-HCO_3 water type. The cluster C3 includes the samples located in the north-eastern extreme area of the aquifer, along the south side of the Mat River flow. The groundwater in this cluster has low content of both Cl^- (12.01 ± 5.39 mg/l) and TDS (253.70 ± 32.27 mg/l), mostly belonging to Mg-Ca-HCO_3 and Mg-Ca-HCO_3-SO_4. The cluster C4 includes the samples located between cluster C1 in the west and clusters C2 and

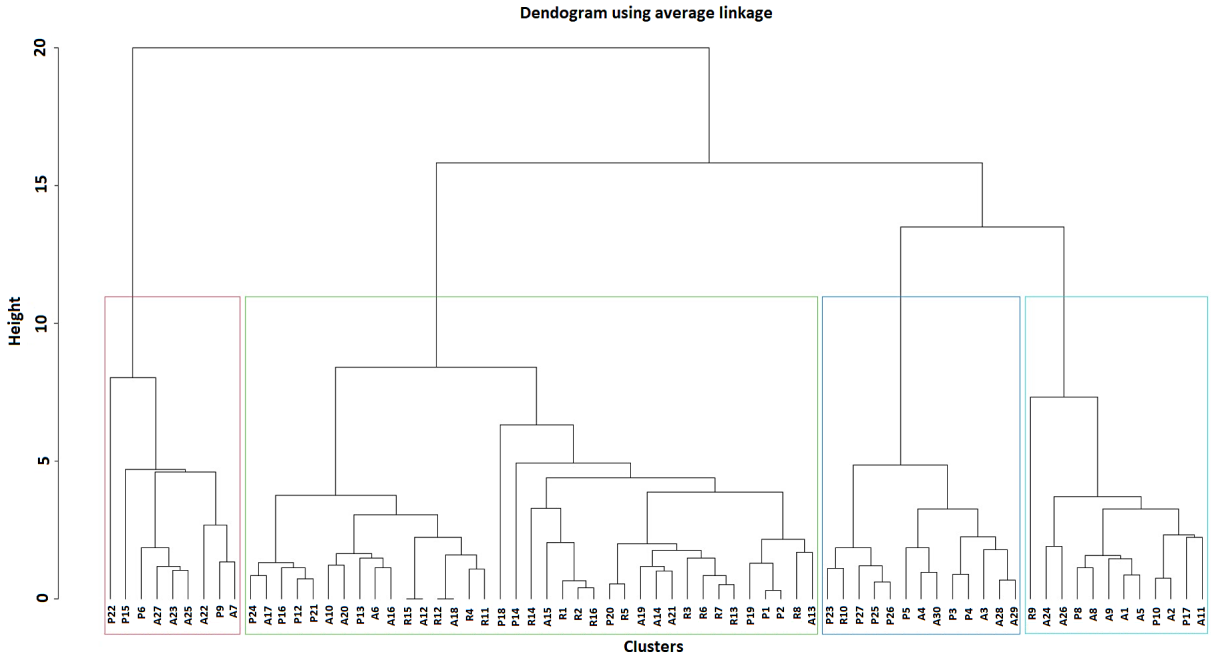

Figure 3. Sites clustered by HCA

C3 in the east. The groundwater in this cluster is characterized by moderate values of Cl^- and TDS (87.72 ± 31.60 mg/l and 445.69 ± 103.56 mg/l, respectively), but is typically soft (average TH = 7.17 ± 2.97 ºdH) water. It belongs mainly to the Na-Mg-HCO_3-Cl, Na-HCO_3-Cl and Na-Cl-HCO_3 water types.

Geochemical interpretation and discussion

Several diagrams are in use in the wide research literature of water chemistry which help to evidence the main hydrochemical water types and to highlight the dominated geochemical processes that control the composition of groundwater. In this paper, two most known (Piper and Chadha) diagrams were applied in combination with results of the multivariate statistical analysis. In the Piper Diagram (Figure 4), four different groups of samples were identified and summarized in Table 6.

Group G1 consists of 3 groundwater (R9, P15, P18) samples (or 4%) that fall within zone I of Piper diagram (Figure 4) belonging to the Calcium Magnesium-Sulfate (Ca-Mg-SO_4) water type. These groundwater samples fall within the Reverse Ion Exchange field of Chadha diagram (Figure 5). In addition, 49 groundwater samples (or 69%), herein named group G2, fall within the zone II of the diagram belonging to the Calcium Magnesium Bicarbonate (Ca-Mg-HCO_3) water type. They were sampled from the wells located in the recharge (southern, south-eastern and eastern zone (Figure 1) and are characterized as chemically young (immature) groundwaters [Allen and Kirste, 2012]. These samples have similar pH, total alkalinity, and major ion concentrations to the river waters that cross the aquifer plain from southeast-east to northwest-west. In fact, the river water points fall within the zone II of the Piper diagram, confirming that their riverbed water constitutes the main recharge source of the groundwater of the alluvial Tirana – Fushe Kuqe aquifer. Carbonate dissolution should be the primary source of Ca^{2+} and HCO_3^- in the groundwater of the recharge zone, while Mg^{2+} might be derived from both carbonate dissolution and weathering of silicate minerals. The saturation of groundwater with carbonates (calcite and dolomite) is well demonstrated by dominantly positive values of Langelier Saturation Index. This group of samples also falls in the Recharge Water field of Chadha diagram (Figure 5).-

Group G3 consists of 13 groundwater samples (or 18%) that fall within the zone III of the Piper diagram (Figure 4), belonging to the Sodium Chloride (Na-Cl) water type. These waters are mainly sampled from the wells near the coastal area and are characterized as chemically evolved (mature) groundwater [Allen and Kirste, 2012]. A few wells (P22, P14, A25), located about 10 km to the east of coast line, south of River Droja, also fall within the zone III. This brackish groundwater should rather be relict sea water [Kumanova et al., 2014], trapped

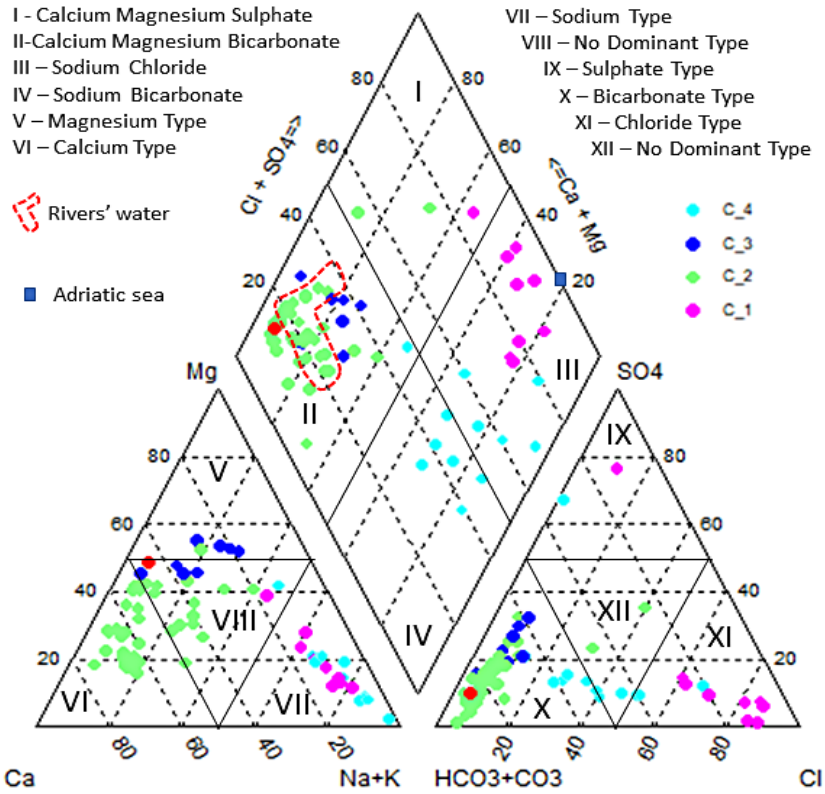

Figure 4. Piper diagram showing the hydrochemical groundwater facies; river waters and clusters are also indicated

Table 6. Classification of groundwater samples based on the piper diagram

Piper diagram zone	Group of samples	Groundwater hydrochemical types	No. of samples	%
I	G1	Calcium Magnesium Sulphate	3	4
II	G2	Calcium Magnesium Bicarbonate	49	69
III	G3	Sodium Chloride	13	18
IV	G4	Sodium Bicarbonate	6	8

within alluvial sediments after the sea water regression during Holocene or are likely derived by diffusion from intercalated clay layers [Jack et al., 2013] saturated with sea water salts. The points of the above-mentioned 13 samples fall within the Sea Water field of the Chadha's classification diagram (Figure 5). The groundwaters from coastal aquifers having the chloride content over 100mg/l are considered as an indication of sea water intrusion [Eftimi, 2003; Lyles, 2000]. Group G4 consists of 6 groundwater samples (or 8%) that fall within the zone IV of Piper diagram (Figure 4) belonging to Sodium Bicarbonate (Na- CO_3^- water type. These waters are sampled from the wells in the center of the Fushe Kuqe area and are characterized as chemically more evolved groundwater [Allen and Kirste, 2012]. The groundwater shows an increase in Na^+ and a respective decrease in Ca^{2+}, which is considered as a consequence Ca^{2+} - Na^+ ion exchange which is expressed by the change from Ca-Mg-HCO_3 to Na-HCO_3 water type [Thorstenson et al., 1979; Henderson, 1985]. In fact, these groundwater samples fall within the Basic Ion Exchange field of Chadha diagram (Figure 5) and represent the lower-chloride samples of the cluster 4.

Table 7, modified after [Cloutier, et al., 2008], indicates the relationships between clusters and principal components with hydrochemical groups by means of sample distribution among them. There is a good overlap between clusters and groups. Thus, the samples from clusters C1 and C2 mostly belong to the groups G3 and G2, respectively; all samples of C3 belong to G2, while in the case of cluster C4, 1 sample belongs to G2, 5 samples to G3 and 6 samples to G4. Clusters C2 and C3 correspond to PC1, while clusters C1 and C4 to PC2. The division of

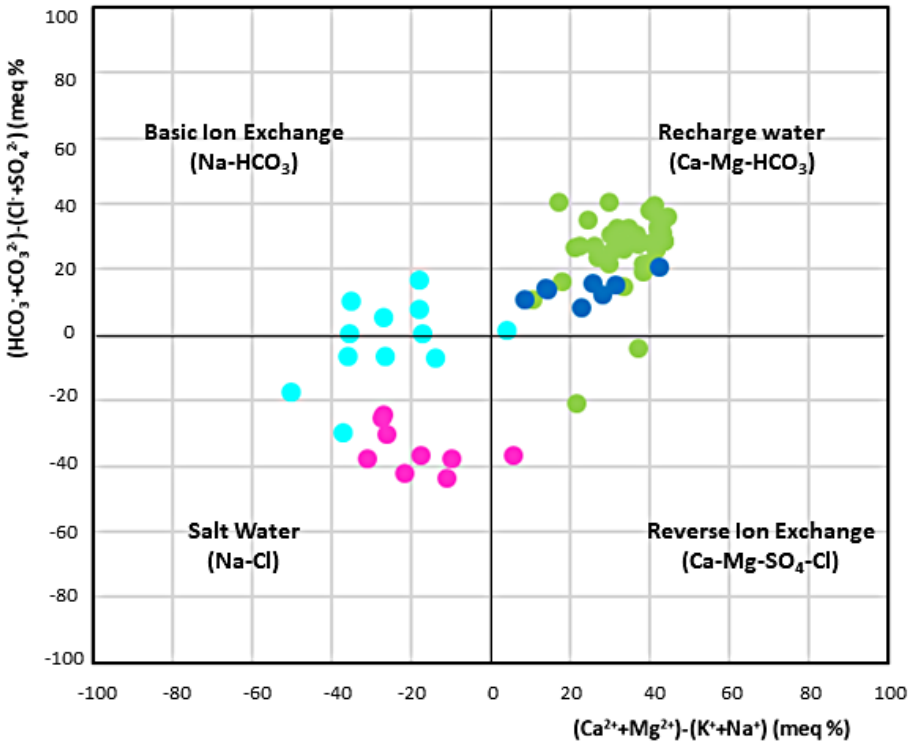

Figure 5. Chadha diagram demonstrating the hydrochemical classification of groundwater considering the dominant geochemical processes (the figurative points are the same as in the Piper diagram)

group G2 in two clusters (C2 and C3) is a better approach with the hydrogeological context of the aquifer. The samples of cluster C3 are located south of the River Mat and belong to the Mg-Ca-HCO_3 water type, while the majority of C2 samples belong to the Ca-HCO_3 and Ca-Mg-HCO_3 types and spread out in the south and central parts of the aquifer. Cluster C3 overlaps the highest hydraulic parameters zone (Hydraulic conductivity = 200–350m/day; transmissivity = 4000–8000m^2/day) of aquifer (Milot-Fushe Kuqe) [Tartari and Dakoli, 2001; Eftimi, 2012]. The groundwater of C2 results from the dissolution of carbonates that outcrop on the mountains to the east of the aquifer, whereas the groundwater of C3 is mostly affected by the silicate weathering of northeastern ophiolites. The samples of cluster C1 represent the coastal subgroup of G3 where groundwater is significantly affected by sea water intrusion and belongs to Na-Cl and Na-Mg-Cl types. Moreover, 5 samples of cluster C4, that are located to the east of samples of cluster C1, fall to G3 having lower content of Cl- than the C1 samples, but over 100 mg/l, and belong to the Na-Cl-HCO_3 and Na-Mg-Cl-HCO_3 types. The next 6 samples of C4 belong to G4, having the Cl content between 50–100 mg/l and belonging to the Na-HCO_3-Cl and Na-Mg-HCO_3-Cl types. Thus, cluster C4 defines a transitional zone between C1 (G3) to the west and C2 and C3 (G2) to the east, where decreasing seawater intrusion from west to east is confronted with saline flushing / cation exchange progressing from east to west. PC's give a more simplified compositional classification of groundwater, where PC1 includes the C2 and C3 samples, revealing the solution and/or weathering

Table 7. Relationship between groundwater clusters, groups and principal components (modified after [Cloutier, et al., 2008]

Clusters	Groups				Principal components		Total
	G1	G2	G3	G4	PC1	PC2	
C1	**1**	0	**8**	0	0	**9**	9
C2	**2**	35	0	0	**37**	0	37
C3	0	13	0	0	**13**	0	13
C4	0	1	**5**	**6**	0	**12**	12
Total	3	49	13	6	50	21	**71**

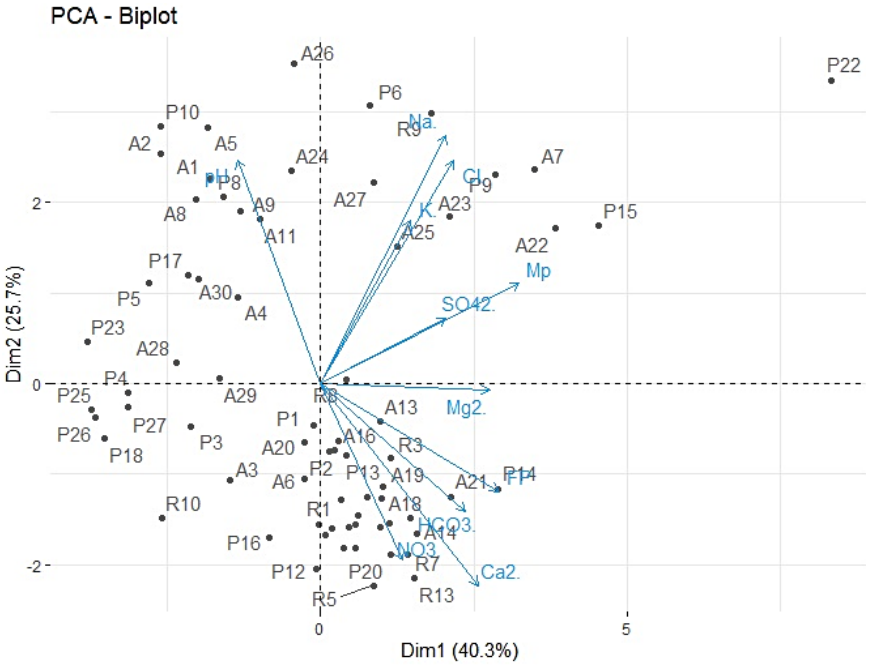

Figure 6. Plots of PCA loadings scores for dataset of groundwater samples

processes which dominate in the south-southeast and east-northeast recharge areas of the aquifer, while PC2 includes the C1 and C4 samples and is related with processes of seawater intrusion and cation exchange which are present along the coastal zone. This is better seen in the biplot score plot (Figure 6), where the first principal component (PC1) that includes the samples from the recharge zone, is characterized by high total hardness and high concentrations of Ca^{2+}, Mg^{2+} and HCO_3^-. The second principal component (PC2) includes the samples from the coastal area of the aquifer which are distinguished by high TDS content and high concentration of Cl^- and Na^+.

CONCLUSIONS

The multivariate statistical methods of PCA and HCA were applied in close combination with conventional techniques of geochemistry for a better understanding of the factors and processes that control groundwater geochemical composition. Piper and Chadha geochemical diagrams distinguished four compositional groundwater groups, among which group G1 consists of 3 groundwater samples belonging to the Ca-Mg-SO_4 water type and falling within the Reverse Ion Exchange field of Chadha diagram; group G2 belongs to the Ca-Mg-HCO_3 type and includes 49 samples from the recharge zone; group G3 consists of 13 groundwater samples from the coastal area which belong to Na-Cl) water type; group G4 comprises 6 groundwater samples from the center of the Fushe Kuqe area showing the Na-HCO_3 composition due to Basic Ion Exchange.

Component PC1 that explains 40.35% amount of the overall variance, is described by very high positive loadings of TH, Ca^{2+}, and Mg^{2+}, indicating the that dissolution of carbonate rocks was the main process in the aquifer recharge zone. Component PC2, that explains 25.74% amount of the total variance, is described by very high positive loadings in Na^+, K^+, and Cl^- and moderate to high loadings of TDS revealing the involvement of seawater intrusion and diffusion from clay layers.

HCA classified the 71 groundwater samples into four clusters (C1–C4). Cluster C1 is dominated by the Na-Cl water type and represents the coastal subgroup of G3 where groundwater is significantly affected by sea water intrusion. Cluster C2 includes 42 samples of G2 from south, southeast and east recharge area having the Ca-Mg-HCO_3 water type. The 7 samples of cluster C3 also belong to G2, but are located to the immediate south of River Mat bed which represents the highest hydraulic parameters zone from Milot to Fushe Kuqe, having the Mg-Ca-HCO_3 water type. Finally, 5 samples of cluster C4 fall to G3, having the Cl^- content higher than 100 mg/l, but lower than cluster C1, while 6 other samples of C4 belong to G4, having the Cl^- content between 50–100 mg/l.

REFERENCES

1. Aitchison J. 1982. The Statistical Analysis of Compositional Data. Journal of the Royal Statistical Society. Series B (Methodological), 44(2), 139–177.
2. Allen D.M., Kirste D. 2012. Results of the July 2011 groundwater chemistry sampling study on Mayne island, British Columbia. Mayne Island, BC: Mayne Island Integrated Water Systems Society.
3. Allen M. 2017. The sage encyclopedia of communication research methods. Thousand Oaks, CA: SAGE Publications, 1–4.
4. Basilevsky A.T. 1994. Statistical Factor Analysis and Related Methods: Theory and Applications. A Wiley-Interscience Publication, 733.
5. Beattie J.R., Esmonde-White F.W.L. 2021. Exploration of Principal Component Analysis: Deriving Principal Component Analysis Visually Using Spectra. Appl Spectrosc, 75(4), 361–375.
6. Beqiraj A., Masi U., Barbieri M. 2002. Geochemical and isotopic characteristics of groundwater in the Tirana-Fushë Kuqe hydrogeological basin and implications for quality. Elbasan, Albania: special edition. (in Italian)
7. Cenameri S., Beqiraj A. 2016. Seawater intrusion in Fushe Kuqe aquifer, geochemical considerations. Buleltin of Geological Sciences, 1/2018, 33–40.
8. Cenameri S., Beqiraj A. 2018. Groundwater Geochemistry of the Fushë Kuqe Aquifer, North-Western Albania.. Journal of Environmental Science and Engineering, A7, 354–360.
9. Cenameri S. 2021. Tracing of groundwater salinization processes in the Fushë Kuqe aquifer through geochemical and isotopic methods. PhD Thesis (in Albanian). National Biblioteca of Albania, Tirana, Albania.
10. Cloutier V., Lefebvre R., Therrien R. and Savard M. M. 2008. Multivariate statistical analysis of Geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. Journal of Hydrology, 353, 294–313.
11. Cooksey R. 2015. Illustrating statistical procedures: finding meaning in quantitative data. Choice Reviews Online, 52(7), 737.
12. Ebrahimi P., Albanese S., Esposito L., Zuzoloc D. and Cicchellac D. 2021. Coupling compositional data analysis (CoDA) with hierarchical cluster analysis (HCA) for preliminary understanding of the dynamics of a complex water distribution system: the Naples (South Italy) case study. Environmental Science: Water Research & Technology, 7, 1060–1077.
13. Eftimi R. 2003. Some Considerations on Seawater-Freshwater Relationship in Albanian Coastal Area. Alikante, Spain, IGME Madrid, 239–250.
14. Eftimi R. 2012. Hydrogeologic conditions of Fushe Kuqe basin. Water Project Implementation Unit. Tirana, Albania, 60.
15. Eftimi R., Gourcy R. L., Stichler W., Amataj S., Zoto J. 2006. Investigation of the Recharge Sources of Tirana Alluvial Basin by Means of Environmental Isotopes. Proc. Marrakech: In Integrated Water Resources Management and Challenges of the Sustainable Develop, 1–9.
16. Everitt B. 1974. Cluster Analysis. Heinemann Educational Books Ltd., London.
17. Farnham I.M., Johannesson K.H., Singh A.K., Hodge V. and Stetzenbach K.J. 2003. Factor analytical approaches for evaluating groundwater trace element chemistry data. Analytica Chimica Acta, 490, 123–138.
18. Filzmoser P. 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities, Sci Total Environ, 407(23), 6100–6108.
19. Gardner R.C., Neufeld R.W.J. 2013. What the correlation coefficient really tells us about the individual. Canadian Journal of Behavioural Science, Revue canadienne des sciences du comportement, 45(4), 313–319.
20. Gewers F.L., Ferreira G.R., De Arruda H.F., Silva F.N., Comin C.H., Amancio D.R., Costa L.D.F. 2021. Principal Component Analysis: A Natural Approach to Data Exploration. ACM Computing Survey, 54(4), 1–34.
21. Grünfeld K. 2005. Dealing with outliers and censored values in multi-element geochemical data - A visualization approach using XmdvTool. Applied Geochemistry, 20, 341–352.
22. Grunsky E.C. 2010. The interpretation of geochemical survey data. Geochemistry: Exploration, Environment, Analaysis, 10, 27–74.
23. Hadi N.U.I., Abdullah N., Sentosa I. 2016. An Easy Approach to Exploratory Factor Analysis: Marketing Perspective. Journal of Educational and Social Research, 6(1), 215–223.
24. Hair J.F. 2019. Multivariate data analysis, Eighth edition, Cengage, Andover, Hampshire, United Kingdom.
25. Hastie T., Tibshirani R., Friedman J. 2009. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer, 764.
26. Henderson T. 1985. Geochemistry of ground-water in two sandstone aquifer systems in the northern Great Plains in parts of Montana and Wyoming: U.S. Geological Survey, Professional Paper 1402-C, C1–C84.
27. Hogarty K.Y., Hines C.V., Kromrey J., Ferron J., Mumford K.R. 2005. The quality of factor solutions in exploratory factor analysis: the influence of sample size, communality, and overdetermination. Educational and Psychological Measurement, 65(2), 202–226.
28. Ihaka R., Gentleman R. 1996. R: A Language for Data Analysis and Graphics, Journal of Computational and Graphical Statistics, 5(3), 299–314.

29. Indu S.N., Rajaveni S.P., Schneider M., Elango L. 2015. Geochemical and Isotopic Signatures for the Identification of Seawater Intrusion in an Alluvial Aquifer. Earth Syst. Sci., 124(6), 1281–1291.
30. Jacks G., Kumanova Xh., Marku S. 2013. Isotopic assessment of the recharge of a coastal aquifer in North Albania. Geophysical Research Abstracts, EGU2013-4586, 15.
31. Jolliffe I.T., Cadima J. 2016. Principal component analysis: a review and recent developments. Phil. Trans. R. Soc. A, 374, 20150202.
32. Kumanova X., Marku S., Fröjdö S., Jacks G. 2014. Recharge and sustainability of a coastal aquifer in Northern Albania. Hydrogeology Journal, 22(4), 883–892.
33. Lyles J.R. 2000. Is seawater intrusion affecting ground water on Lopez Island, Washington?. USGS Numbered Series, U.S. Geological Survey, Fact Sheet FS-057-00.
34. Melloul A., Collin M. 1992. The 'principal components' statistical method as a complementary approach to geochemical methods in water quality factor identification; application to the Coastal Plain aquifer of Israel. Journal of Hydrology, 140(1–4), 49–73.
35. Murtagh F., Legendre P. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion?. Journal of Classification, 31(3), 274–295.
36. Mushtaq N., Masood N., Khattak J.A., Hussain I., Khan Q., Farooqi A. 2020. Health risk assessment and source identification of groundwater arsenic contamination using agglomerative hierarchical cluster analysis in selected sites from upper Eastern parts of Punjab province, Pakistan. Human and Ecological Risk Assessment: An International Journal, 27(4), 999–1018.
37. Nakagawa K., Amano H., Berndtsson R. 2021. Spatial Characteristics of Groundwater Chemistry in Unzen, Nagasaki, Japan. Water, 13, 426.
38. Pison G., Rousseeuw P.J., Filzmoser P., Crouxc C. 2003. Robust factor analysis. Journal of Multivariate Analysis, 84(1), 145—172.
39. Reimann C., Filzmoser P., Garrett R.G. 2005. Background and threshold: critical comparison of methods of determination. Science of The Total Environment, 346(1–3), 1–16.
40. Reimann C., Filzmoser P., Garrett R.G. 2022. Factor analysis applied to regional geochemical data:. Applied Geochemistry, 17, 185–206.
41. Reimann C., Filzmoser P. 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology, 39, 1001–1014.
42. Revelle W. 2021. Psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA.
43. Rock N.M.S. 1988. Summary statistics in geochemistry: A study of the performance of robust estimates. Mathematical Geology, 20(3), 243–275.
44. Romesburg H.C. 2004. Cluster Analysis for Researchers. Morrisville, NC: Lulu.com, 330. (Reprint of 1984 edition, with minor revisions)
45. Schloerke B., Cook D., Larmarange J., Briatte F., Marbach M., Thoen E., Elberg A., Toomet O., Crowley J., Hofmann H., Wickham H. 2020. GGally: Extension to "ggplot2." Mathematics.
46. Steinhorst R.K., Williams R.E. 1985. Discrimination of groundwater sources using cluster analysis, MANOVA, canonical analysis and discriminant analysis.. Water Resources Research, 21, 1149–1156.
47. Stetzenbach K.J., Hodge V.F., Guo C., Farnham I.M. 2001. Geochemical and statistical evidence of deep carbonate groundwater within averlying volcanic rock aquifers/aquitards of southern Neveda, USA. Journal of Hydrology, 243, 254–271.
48. Tartari M., Dakoli H. 2001. Quantitative and qualitative evaluation of Tirana-Lezha basin groundwater scoping to increase the pumping rates. Research report, Albanian Geological Survey, 50.
49. Thorstenson D.C., Fisher D.W., Croft M.G. 1979. The geochemistry of the Fox Hills-Basal Hell Creek Aquifer in Southwestern North Dakota and Northwestern South Dakota. Water Resources Research, 15(6), 1479–1498.
50. Torokhti A., Friedland Sh. 2009. Towards theory of generic Principal Component Analysis. Journal of Multivariate Analysis, 100, 661–669.
51. Wickham H., Averick M., Bryan J., et al., 2019. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686.
52. Xu N., Finkelman R.B., Dai S., Xu C., Peng M. 2021. Average Linkage Hierarchical Clustering Algorithm for Determining the Relationships between Elements in Coal. ACS Omega, 6(9), 6206–6217.
53. Xu N., Peng M., Li Q., Xu C. 2020. Towards Consistent Interpretations of Coal Geochemistry Data on Whole-Coal versus Ash Bases through Machine Learning. Minerals, 10(4), 328.
54. Yanai H., Ichikawa M. 1990. New lower and upper bounds for communality in factor analysis. Psychometrika, 55(2), 405–409.
55. Zainol N.F.M., Zainuddin A.H., Looi L.J., Aris A.Z., Isa N.M., Sefie A., Yusof K.M.K.K. 2021. Spatial Analysis of Groundwater Hydrochemistry through Integrated Multivariate Analysis: A Case Study in the Urbanized Langat Basin, Malaysia. Int. J. Environ. Res. Public Health, 18(11), 5733.
56. Zhang Y., He Z., Tian H., Huang X., Zhang Z., Liu Y., Xiao Y., Li R. 2021. Hydrochemistry appraisal, quality assessment and health risk evaluation of shallow groundwater in the Mianyang area of Sichuan Basin, southwestern China. Environmental Earth Sciences, 80(17).