2024 4th International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) | 979-8-3503-9118-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICECCME62383.2024.10796675

K-means method for clustering groundwater chemical parameters of Konya Closed basin

Carmen Maftei
Civil Engineering Faculty
Transilvania University of Brasov
Brasov, Romania
cemaftei@gmail.com

Cristina Serban
Mathematics and Computer Science Faculty
Ovidius University of Constanta
Constanta, Romania
cgherghina@gmail.com

Kubra Kucuk
Civil Engineering Faculty
Kto Karatay University
Konya, Türkiye
kubrakuchuk@gmail.com

Vahdettin Demir
Civil Engineering Faculty
Kto Karatay University
Konya, Türkiye
vahedettin.demir@karatay.edu.tr

Radu Muntean
Civil Engineering Faculty
Transilvania University of Brasov
Brasov, Romania
radu.m@unitbv.ro

***Abstract*—Cluster analysis, respectively K-Means method, is used to classify groundwater from Konya Closed Basin, located in Central Anatolia Region of Turkey. Datasets include the beginning (MB) and end (MS) of the season values for pH, Na and HCO_3 spanning 2015-2022 period. There are 163 water wells (stations) in Konya Closed Basin where chemical properties were measured. Additionally, an EDA and statistical analysis was performed. The international standards for pH, Na and HCO_3 in drinking water and irrigation water are generally respected. To evaluate the success of cluster separation using the K-Means technique, we estimated the silhouette coefficient. The results reveal that, generally, 2 to 4 clusters were determined for all models proposed. This approach can help to identify those wells that exceed the standard values of the investigated chemical parameters for drinking and irrigation water. In this way, the water management department could take improvement measures over time.**

Keywords—groundwater, chemical parameters, cluster analysis

I. INTRODUCTION

Water is an essential resource for arid and semi-arid areas.

To analyze the chemical properties of groundwater or surface water, different methods are used: statistical methods US salinity diagram, Piper diagram, Geographical Information System (GIS), etc. In [1], twenty-seven groundwater samples were picked from shallow-dug wells and the groundwater hydrochemistry was established by means of bivariate plots, a Piper plot, and a Durov Plot for drinking water quality assessment. A Piper three-line diagram to characterize the chemical types of groundwater, the Gibbs model and ion ratio method to determine the main controlling factors, the Kriging interpolation method to describe the spatial distribution and variation of chemical elements, and a statistical analysis were conducted by [2] to discuss chemical characteristics and influencing factors of groundwater. Piper three-line diagram, Gibbs diagram and ion combination ratio method were also used by [3] to observe the temporal–spatial evolution of groundwater chemical characteristics and controlling factors in the study area.

Several studies have focused over the years on Konya Closed Basin (Turkey), mainly investigating the trend of water-level change in lakes or wells, to devise effective water management strategies. The study in [4] aims to determine the changes of the groundwater level of 10 observation stations in Konya Closed Basin, from 1978 to 2020, using the parametric Linear Trend method and the non-parametric Mann-Kendall. The same topic is treated by [5], which employs GIS to estimate the water level changes in 18 wells in some districts in the borders of Konya Closed Basin, thus establishing the groundwater database of the region. Research in [6] is done by applying remote sensing techniques. Temporal gridded datasets of Gravity Recovery and Climate Experiment (GRACE) and the Global Land Data Assimilation System (GLDAS) are used to estimate and monitor the groundwater storage changes in the Konya Closed Basin.

In this study, we conducted an analysis on the chemical characteristics of several groundwater drills situated in Konya Closed Basin. The research focused on three chemical parameters, namely pH, Na (Sodium) and HCO3 (Bicarbonates), employing classical models (like statistics methods, Q-Q plot, Shapiro and D'Agostino-Pearson tests) and cluster analysis, to identify a correlation between the investigated parameters.

Cluster analysis is a useful data mining technique that allows researchers to gain meaningful insights into the structure of their data [7, 8, 9–11]. Generally, there are two data science approaches: supervised and unsupervised. Clustering is considered an unsupervised learning due to its lack of a class label, which is present in supervised learning such as classification. In supervised learning, the algorithm

"learns" from the training dataset by repeatedly making predictions on the data and adapting for the correct answer. While supervised learning methods tend to be more accurate than unsupervised learning ones, they require human intervention to label the data appropriately. Unsupervised learning models discover hidden patterns in data without human intervention, requiring it only for validating results.

Traditionally, cluster methods involve using exploratory algorithms to group data points into clusters (group or subset of data points), based on their similarities and differences, thus identifying relationships and patterns within complex data sets. The classification into clusters is done using various criteria such as smallest distances, density of data points, or various statistical distributions.

There are many methods used in cluster analysis. The Principal Component Analysis (PCA) technique can be employed to (i) assess the clustering and similarities present in the datasets, (ii) investigate the consistency and overlap of the clusters, and (iii) pinpoint the sources of variation among the parameters [7]. While PCA simplifies the data by reducing its dimensionality, it may lead to loss of some information from the original data. In addition, the new components generated by PCA may not have a straightforward interpretation. Hierarchical clustering is also a method of cluster analysis which attempts to group data points into a tree of clusters, without having fixed number of clusters [7]. This method is characterized by ease of handling of any forms of similarity or distance. However, it requires the computation and storage of a squared distance matrix, which, for very large datasets, can be expensive and slow. The K-means method enables the classification of monitoring stations by grouping them according to the similarities found in their samples [12]. A major advantage of the K-means algorithm, apart of being fast and scalable, is its capacity to evaluate clustering results in a quantitative and objective manner using cluster validity indices, which helps determine the optimal number of hidden clusters in the dataset [8].

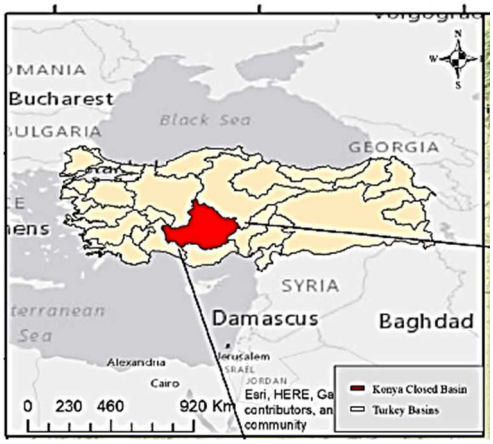
In this research, we chose to apply on our datasets the K-Means method of clustering, an iterative algorithm that finds in optimal time a predetermined number (k) of clusters by minimizing the mean distance between geometric points.

II. MATERIALS AND METHODS

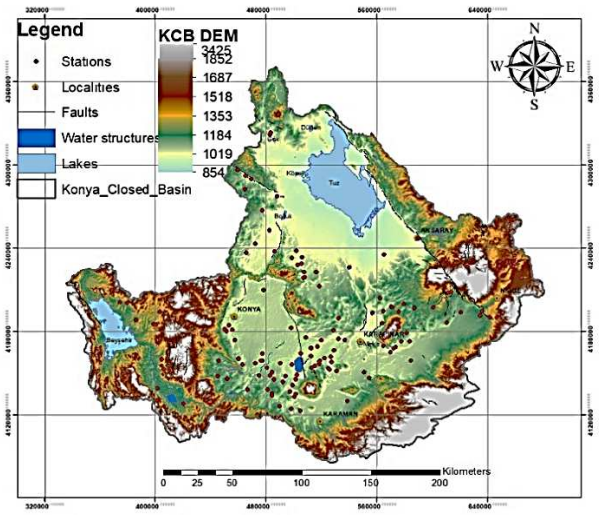
A. Study area

Konya Closed Basin is in Turkey's Central Anatolia Region between 36°51' and 39°29' northern latitudes and 31°36' and 34°52' eastern longitudes. With an area of approximately 5 million hectares, Konya Closed Basin represents 7% of the country territories. The basin is surrounded by the Sakarya and Kızılırmak basins to the north, the Kızılırmak and Seyhan basins to the east, the Eastern Mediterranean to the south, and the Antalya and Akarçay basins to the west (Figure 1).

The Taurus Mountains border the southern part of the Konya Closed Basin (KCB) in an arc shape. While the elevation of the plains in the interior regions varies between 850-1000 meters (covering approximately 65% of the entire basin), the elevations in the Taurus Mountains rise to 3900 meters (Figure 1).



(a)



(b)

Fig. 1. Kunya Closed Basin (KCB) location (a) and geographical map (b) (DEM - digital elevation model)

The majority of the water sources feeding the basin are streams and groundwater originating from the Taurus Mountains [5]. Due to the inefficient use of the water bodies within the basin and improper irrigation methods, surface water is depleting, and groundwater is being used extensively for agriculture, causing groundwater levels to drop to critical levels. From a hydrological point of view, in Konya Closed Basin exist 92 surface water bodies (natural lakes, ponds, reservoirs, etc.), and 18 groundwater bodies.

Known as Turkey's grain silo, the Konya Closed Basin is also one of the world's 200 most important ecological regions in terms of biological diversity. An important agricultural and economic production area, the Konya Closed Basin also hosts 15 significant plant areas and 6 significant bird areas [13].

B. Data series

There are 163 water wells (stations) in Konya Closed Basin where chemical properties are measured (Figure 1b). These measurements were conducted by the State Hydraulic Works (DSİ). The three chemical properties examined are pH, Sodium (Na), and Bicarbonate (HCO_3). The values measured at the beginning of the season (first six months) and the end of the season (last six months) spanning 2015 -

2022 period. Groundwater samples were collected from different depths which vary from a minimum of 25 m to a maximum of 200 m. The samples are analyzed in DSI specialized laboratory and for this reason they are expected to be reliable and free of gross errors. As we already explained, sometimes the wells dry up, or the level reaches critical values, and it is no longer possible to take samples for underground water analysis. In line with this, we decided to investigate only the wells for which we have data recorded for beginning (MB) and end of seasons (MS), for each year. It should also be mentioned that for 2018, there are no HCO_3 records for MB.

C. Methods

The methodology used in this paper is based on four parts (Figure 2).

To create a data base operational files were created using MS Excel, separately for beginning of season (MB) and end of season (MS). Usually, EDA (Exploratory Data Analysis) is a first step before additional statistical analysis or modeling are conducted. This analysis is used to visualize data and understand it, identify patterns, and establish correlations between variables.

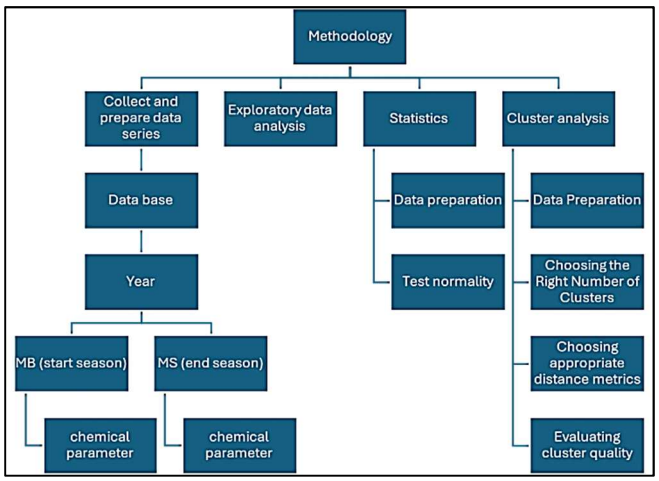


Fig. 2. Methodology proposed

Additionally, we investigated the three chemical parameters in relation to WHO (drinking water) and FAO (water for irrigation) limits accepted (Table I). The values of Na and - HCO_3 are in mEq/l (milliequivalents per liter).

TABLE I. STANDARDS LIMITS

Chemical parameters	Limits	
	WHO (2006)	***FAO (1970)***
pH	6.5 - 9.5	5.0 - 8.4
Na (mEq/l)	< 8.7	< 40
HCO_3 (mEq/l)	< 16.39	< 10

Statistics methods refers to descriptive statistics, test of normality and trend determination. We used Real statistic under MS Excel. Q-Q plot, Shapiro and D'Agostino-Pearson tests are used to investigate the normality of data sets. Note that the D'Agostino-Pearson test is not used if the data is less than 20.

In this study, we chose to apply on data sets the K-Means method of clustering, an iterative algorithm that finds a predetermined number (k) of clusters by minimizing the mean distance between geometric points. The algorithm follows several steps: 1) specify the number of clusters; 2) randomly assign data points to clusters; 3) compute cluster means: for each cluster, compute the average value for each of the data point; 4) allocate each data point to the closest cluster center; 5) repeat steps 3 and 4 until the algorithm converges. For a clustering algorithm, it is utterly important to determine the optimal number of clusters. We conducted K-means several times, exploring different clustering evaluation criteria, such as silhouette analysis or gap statistic, and we finally chose to evaluate the optimal number of clusters using the Calinski-Harabasz clustering evaluation criterion [14].

Before performing cluster analysis, we ensured that the data is clean and free from errors, by handling missing values. Due to the completely random nature of the missing data, our approach assumed a complete case analysis, which involves using only data points with complete information, thus any data points that contain missing values were removed. We also needed to get the data sets in the right format. For each year considered, we set a table of raw data, where each row represents a data point, and the columns represent the clustering variables, pH, Na and HCO_3, respectively. Because we conducted K-means numerous times with different variables, a data point had the MB or MS values of all variables or only of two of them, (pH, Na) or (pH, HCO_3).

Silhouette analysis is used to evaluate the optimal number of clusters. A critical step in cluster analysis is to choose the appropriate distance metrics, as it determines how similarity is calculated between data points. For this study, we selected the Manhattan distance (the sum of absolute differences between the coordinates of two points), also known as city block distance, as it is appropriate for numerical data with different scales and robust to outliers.

To interpret the cluster cohesion and separation, we used 2D and 3D scatter plots, that help us visualize the data points and their allocated cluster labels. We also employed the silhouette coefficient, which measures how well each data point fits within its allocated cluster compared to other clusters. It ranges from -1 to +1, with higher values indicating better-defined and well-separated clusters. If most points have a high silhouette coefficient, then the clustering solution is appropriate. If many points have a low or negative silhouette coefficient, then the solution might have too many or too few clusters. There is no convention on the definition of a "high" or "low" silhouette coefficient, and, for this study, we used "high" to mean greater than 0.5 and "low" to mean less than 0.2. We performed cluster analysis with the Statistics and Machine Learning Toolbox in MATLAB R2015a, which provides several clustering algorithms and tools.

III. Results and Discussion

A. EDA Analysis

EDA (Exploratory Data Analysis) is an analysis technique used to understand the characteristics and structure of a data set before applying statistical models or machine learning algorithms. EDA focuses on exploring data, typically through graphical visualizations and descriptive

statistics, to uncover patterns, relationships, anomalies, or hidden trends

Generally, the pH variation for all well stations and for all years are in the recommended limits of WHO and FAO (Figure 3).

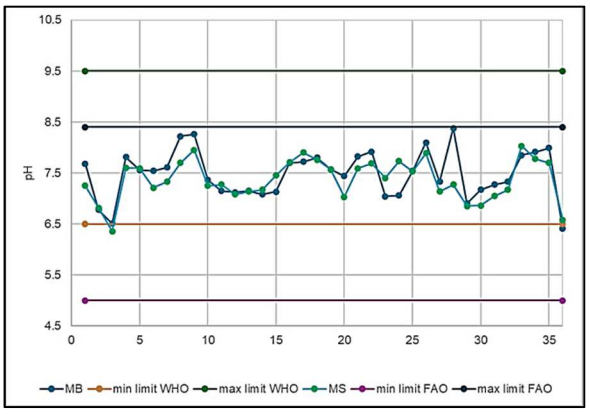


Fig.3. pH variation for 2018

However, there are some values that are out of limits (Table II, last rows).

TABLE II. PH STATISTICAL RESULTS

Specification pH	2015		2016		2017		2018		2019		2020		2021		2022	
	MB	MS	MB	MS	MB	MS	MB	MS	MB	MS	MB	MS	MB	MS	MB	MS
Descriptive statistics																
Mean	7.4	7.4	7.3	7.4	7.3	7.4	7.5	7.4	7.5	7.5	7.6	7.6	7.4	7.4	7.4	7.4
Standard Error	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.0
Median	7.4	7.4	7.3	7.4	7.3	7.6	7.5	7.4	7.5	7.6	7.7	7.7	7.5	7.4	7.4	7.4
Mode	7.0	7.4	7.3	7.7	7.1	7.7	7.5	7.3	7.3	8.0	7.7	7.7	7.9	7.4	7.2	7.3
Standard Deviation	0.3	0.3	0.4	0.4	0.5	0.5	0.5	0.4	0.9	0.4	0.4	0.4	0.3	0.3	0.4	0.4
Sample Variance	0.1	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.9	0.2	0.1	0.1	0.1	0.1	0.1	0.1
Kurtosis	-1.2	-0.9	0.6	0.4	7.4	2.5	-0.2	-0.1	11.3	1.2	0.1	0.1	-0.3	1.2	-0.2	0.3
Skewness	0.2	-0.1	-0.3	-0.7	-1.8	-1.5	-0.2	-0.5	-1.8	-0.7	-0.4	-0.4	-0.6	-0.8	-0.4	-0.6
Range	1.0	1.1	2.1	1.9	2.9	2.2	2.0	1.7	6.7	2.0	1.8	1.8	1.3	1.5	1.6	1.7
Maximum	7.9	7.9	8.2	8.0	8.1	8.1	8.4	8.0	10.1	8.2	8.4	8.4	7.9	8.0	8.0	8.0
Minimum	7.0	6.8	6.0	6.2	5.2	5.9	6.4	6.4	3.4	6.3	6.7	6.7	6.6	6.5	6.4	6.3
Count	17	17	97	97	63	63	36	36	36	36	57	57	31	31	79	79
AAD	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.3	0.5	0.3	0.3	0.3	0.3	0.3	0.3	0.3
MAD	0.3	0.3	0.2	0.3	0.2	0.3	0.4	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.2
IQR	0.5	0.5	0.5	0.6	0.5	0.6	0.7	0.6	0.5	0.4	0.5	0.5	0.5	0.4	0.5	0.4
Shapiro-Wilk Test																
W-stat	x	x	0.98	0.96	0.85	0.88	0.98	0.97	0.73	0.95	0.97	0.98	0.95	0.94	0.97	0.97
p-value	x	x	0.11	0.00	0.00	0.00	0.88	0.37	0.00	0.15	0.24	0.37	0.17	0.10	0.08	0.04
alpha	x	x	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
normal	x	x	yes	no	no	no	yes	yes	no	yes	yes	yes	yes	yes	yes	no
d'Agostino-Pearson																
DA-stat	x	x	2.87	7.86	39.9	23.7	0.34	1.67	32.8	5.15	2.40	1.75	2.38	5.84	2.45	5.40
p-value	x	x	0.24	0.02	0.00	0.00	0.84	*0.43*	0.00	0.08	0.30	0.42	0.30	0.05	0.29	0.07
alpha	x	x	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
normal	x	x	yes	no	no	no	yes	yes	no	yes	yes	yes	yes	yes	yes	yes
outliers																
no of outliers	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no	no
no of records out of WHO and/or FAO limits																

no of records	no	no	1	2	1	3	1	1	4	no	no	no	no	no	1	1
no of records	no	no	no	no	no	1	no	no	2	no	no	no	no	no	no	no

B. *Statistical Analysis*

Statistical analysis for pH is presented in Table II. From the above table the following can be observed: (i) the timeseries data are an average around 7.4; (ii) 64% form all data series are normal; (iii) for 2015 the normality test was not performed, since the series contains 17 records; (iv) no outliers are observed. All Q-Q plots diagrams are consistent with the results of the normality tests. For 2015 MB Q-Q plot reveals that this series is not normal (Fig. 4). The 2015 MS series is normal.

The calculations for the other two parameters were carried out in the same way. Concerning the Na, for 2019 MB there are no records registered. Statistical analysis reveals the following results: (i) average of Na is around 2mEq/l; (ii) all series are not normal; (iii) no outliers are observed.

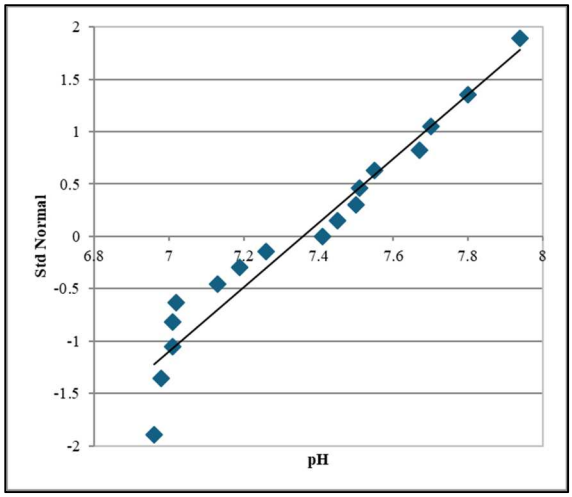


Fig. 4. Q-Q plot for 2015 MB series

Regarding bicarbonate (HCO_3), we must note that there are no registrations for the year 2019 MB. The series is not normal. The average of this parameter is between 8.87 and 8.94, which means that in general the values registered are within the standard limits for drinking and irrigation water. Some outliers were identified. They are removed from the data series in the following analyses.

C. *Clusters Analysis*

Cluster analysis aims to identify homogeneous subgroups of examples within a set of groups that minimize variance among groups while maximizing between-group variation.

The following models were tested: (i) pH MB-MS, (ii) pH-Na MB; (iii) pH-Na MS; (iv) pH- HCO_3 MB; (v) pH-HCO_3 MS; (vi) pH- Na-HCO_3 MB; (vii) pH- Na- HCO_3 MS. The silhouette plot in Fig. 5 shows that for model (vii) the data is split into four clusters of approximately equal size. All the points in the clusters have high silhouette values, indicating that the clusters are well separated.

Generally, 2-4 clusters are identified on models used. We decided to exemplify the results only for the year 2018. For the model (vii) presented in Fig.7, 4 clusters are identified (from 97 data, 35% are including in cluster 1, 37% in cluster 3, 23% in cluster 4 and the rest in cluster 2). Table III shows the average of all chemical parameters for each cluster. pH values increase from cluster 1 to 4.

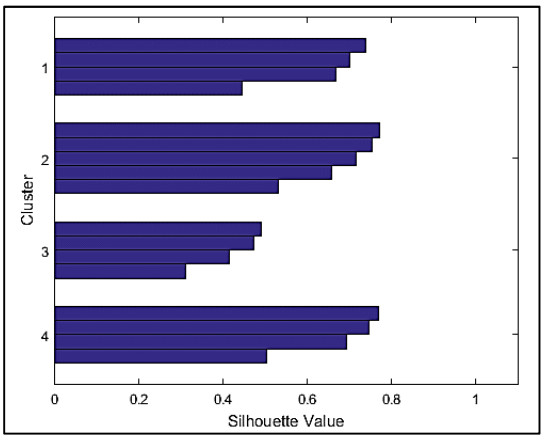


Fig.5. Silhouette value for 2015 series

For example, Fig 6 shows the model (i) for 2018, and Fig. 7 shows the model (vii) for 2018.

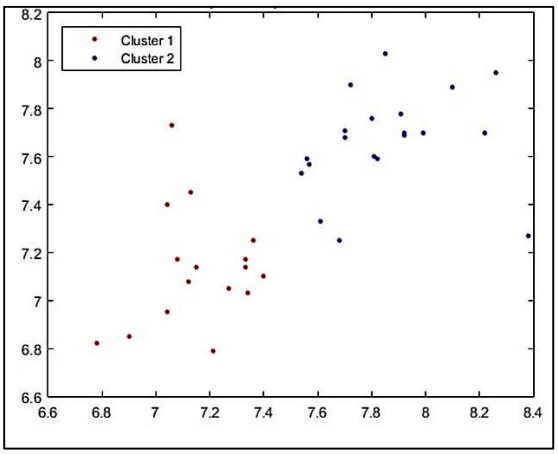


Fig.6. Model (i) results for 2018.

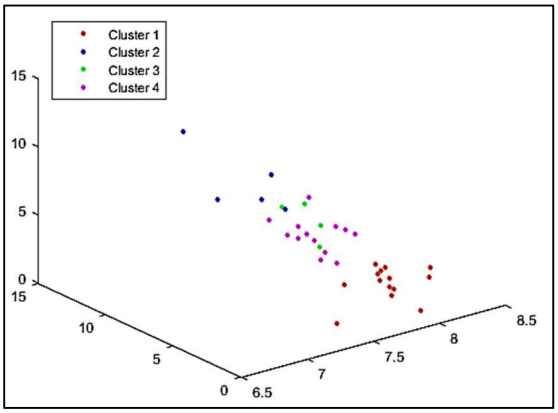


Fig.7. Model (vii) pH- Na- HCO3 MS results for 2018Cluster 3 represents 36 records with the lowest value of Na and HCO3 (Table III). Cluster 1 represents 34 records with the highest value of HCO3 and cluster 2 represents 4 records with the highest value of Na.

TABLE III. MODEL (VII) FOR 2018

Average	Cluster 1	Cluster 2	Cluster 3	Cluster 4
pH	7.06	7.14	7.56	7.57
Na (mEq/l)	1.76	41.79	0.93	5.70
HCO_3 (mEq/l)	9.34	8.91	4.19	5.99

As can be seen from Table III, the Na average value of the records exceed 2 to 10 times the limit value of sodium for drinking water at all 4 wells. The standard value of Na for irrigation water exceeds 2 times the one of a single well.

IV. CONCLUSION

The present study uses a combination of classical statistical analysis and K-means methodology. Generally, all the data series are not normal. The international standards for pH in drinking water and irrigation water are generally respected, with some exceptions. The same results are obtained for Na and HCO_3. Simulation results show that the K-means algorithm can efficiently investigate the chemical data measured on Konya Closed Basin. This method allows us to detect the water wells where anomalies of the chemically investigated parameters are recorded. For example, in 2018, for model (vii), cluster 2 contains sodium values that far exceed the international standard for drinking water (<8.7mEq/l). The cluster results have the potential to assist local governments establish more focused and sustainable policies for managing groundwater supplies. More research studies need to be conducted for better water classification.

ACKNOWLEDGMENT

This research was supported by the Transilvania University of Brasov. Authors thanks to DSI from Turkey to provided data sets.

REFERENCES

1. YK. Ali, JA. Hussain, A. Kamal and KZ. Faisal, "Physio-chemical properties of groundwater and their environmental hazardous impact: Case study of Southwestern Saudi Arabia", Journal of King Saud University - Science, vol. 33(2):101292, 2021

2. Z. Gong, X. Tian, L. Fu, H. Niu, Z. Xia, Z. Ma, J. Chen and Y. Zhou, "Chemical Characteristics and Controlling Factorsof Groundwater in Chahannur Basin. Water", vol. 15(8):1524, 2023

3. C. Li, B-H. Men and S-Y. Yin, "Analysis of Groundwater Chemical Characteristics and Spatiotemporal Evolution Trends of Influencing Factors in Southern Beijing Plain", Front. Environ. Sci. vol. 10, 913542, 2022, doi: 10.3389/fenvs.2022.913542.

4. V. Demir, E. Uray, O Orhan, A. Yavariabdi and H. Kusetogullari, "Trend Analysis of Ground-Water Levels and The Effect of Effective Soil Stress Change: The Case Study of Konya Closed Basin, European", Journal of Science and Technology, vol. 24, 2021, pp. 515-522

5. F Başçiftçi, S. Durduran and C. Inal, "Mapping Ground Water Level With Geographic Information System (GIS) in Konya Closed Basin", Electronic Journal of Map Technologies, vol. 5(2), 2013, pp. -15

6. K.K. Yılmaz and M.T. Yılmaz, "Evaluation of Groundwater Storage changes at Konya Closed Basin Turkey using GRACE based and in situ measurements", presented at the EGU General Assembly 2016 (17 - 22 Nisan 2016), Vienna, Austria, Available: https://hdl.handle.net/11511/77958

7. M. El-Rawy, H. Fathi, F. Abdalla, F. Alshehri and H. Eldeeb, "An Integrated Principal Component and Hierarchical Cluster Analysis Approach for Groundwater Quality Assessment in Jazan, Saudi Arabia", Water, vol. 15, 1466, 2023, doi: 10.3390/w15081466.

8. A.E Marín Celestino, D.A. Martínez Cruz, E.M. Otazo Sánchez, F. Gavi Reyes and D. Vásquez Soto, "Groundwater Quality Assessment: An Improved Approach to K-Means Clustering, Principal Component Analysis and Spatial Analysis: A Case Study", Water, vol. 10, 437, 2018, doi: 10.3390/w10040437

9. G. Shenbagalakshmi, A. Shenbagarajan, S. Thavasi, M. Gomathy Nayagam and R. Venkatesh, "Determination of water quality indicator using deep hierarchical cluster analysis", Urban Climate 49:101468, 2023, doi: 10.1016/j.uclim.2023.101468

10. A. Shihab and A. Hassan, "Cluster analysis classification of groundwater quality in wells within and around Mosul city, Iraq", Journal of Environmental Hydrology, vol. 14(24), 2006

11. A.P.Windarto, M.N. Hasan Siregar, W. Suharso, B. Fachri, A. Supriyatna, I. Carolina, Y. Efendi and D. Toresa, "Analysis of the K-Means Algorithm on Clean Water Customers Based on the Province", J Phys: Conf Ser, vol. 1255, 012001, 2019, doi: 10.1088/1742-6596/1255/1/012001

12. M.H. Eid, M. Eissa, E.A. Mohamed, H.S. Ramadan, G. Czuppon, A. Kovács and P. Szűcs, "Application of stable isotopes, mixing models, and K-means cluster analysis to detect recharge and salinity origins in Siwa Oasis, Egypt" Groundwater for Sustainable Development, vol. 25, 101124, 2024, doi: 10.1016/j.gsd.2024.101124

13. N.K. Özür and M. Ataol, "Assessment on the use of Corine data in Turkey", Journal of Institute of Social Sciences, vol. 9(2), 2018, pp. 110-130

14. T. Caliński, J. Harabasz, “A dendrite method for cluster analysis”, Communications in Statistics - Theory and Methods, vol. 3(1), 1974, pp.1-2