University of Birjand

Water Harvesting Research Vol. 4, No. 1, Spring & Summer 2021, Original Paper p. 68-75

10.22077/JWHR.2021.4632.1044

Designing a quality monitoring network of Gonabad Aquifer using principal component analysis (PCA) method

Samira Rahnama[a], Abbas Khashei Siuki[b]*, Ali Shahidi[c] and Ali Mohammad Noferesti[d]

[a]PhD Student, Department of Water Science and Engineering, Faculty of Agriculture, University of Birjand, Birjand, Iran.
[b]Professor, Department of Water Science and Engineering, Faculty of Agriculture, University of Birjand, Birjand, Iran.
[c]Associate Professor, Department of Water Science and Engineering, Faculty of Agriculture, University of Birjand, Birjand, Iran.
[d]Department of Water Science and Engineering, Faculty of Agriculture, University of Birjand, Birjand, Iran.

* Corresponding Author, E-mail address: abbaskhashei@birjand.ac.ir
Received: 26 August 2021/ **Revised**: 7 September 2021/ **Accepted**: 9 September 2021

Abstract

In order to efficiently manage groundwater resources, determination of the main sampling points is very important to reduce sample size and save time and cost. Principal Component Analysis (PCA) is one of the data reduction techniques that has an important role in identifying insignificant data. In this research, 22 wells of Gonabad plain with a statistical length of 10 years (2007-2016) were used. In the studied area, the annual average of 11 quality parameters of Ca, Mg, Na, EC, TDS, Cl, SAR, HCO_3, SO_4, TH, pH groundwater was investigated by using this technique to determine the quality effective wells in the aquifer of this plain. Using PCA, the relative importance of each well was calculated between 0 (for completely ineffective well) to 1 (for the very effective wells). The results showed that among the 22 wells in the study area, 7 wells were identified as the quality effective wells of Gonabad plain, which had a good dispersion in the region and could play an important role in reducing sampling costs.

Keywords: Effective well, Gonabad plain, Groundwater, Principal Component Analysis.

1- Introduction

Groundwater is one of the most important sources of water for drinking, industrial and agricultural use, especially in arid and semi-arid regions (Nguyan et al., 2013). For this reason, continuous monitoring of groundwater will play an important role in its management. Groundwater monitoring network design is usually done to monitor the groundwater level, groundwater quality or both, which has an important role in managing the operation of the aquifer. In the groundwater-monitoring network, in order to save cost, time and increasing the sampling accuracy, the main wells should be used, which are the same wells that are effective in monitoring (NouriGheidari, 2013). Monitoring methods are performed by both statistical and geological ones (Helena et al., 2000). Advanced statistical methods such as geostatistics and modern methods such as neural network and genetic algorithm are used in statistical methods including simulation methods, analysis of variance and probabilistic methods. However, geological methods are based on the quantity and quality of geological information and groundwater (Lucas and Jauzein, 2008). Principal Component Analysis (PCA) is a mathematically optimal method for reducing data volume and converting primary variables to limited components (Jolliffe, 2002) that has been used to monitor the groundwater network. According to the degree of correlation or covariance between wells or measuring points (Ouyang, 2005), principal

component analysis defines principal or latent components. In this way, after identifying the components causing the most variance changes, the variables having the highest correlation coefficient with the principal components can be extracted. Principal component analysis is widely used in surface water and groundwater (Siyue, 2009), some of which are mentioned below.

Gurunathan and Ravichandran (1994) used PCA method to identify the quality of Italian open aquifers. In this study, they introduced evaporation, irrigation cycle and bedrock material as the main variables. Ouyang (2005) used PCA methods to assess the water quality of the St. John's River Monitoring Network in Florida, USA. According to the obtained results, calcium, magnesium, alkalinity, total nitrogen, soluble nitrate and nitrite are among the effective parameters in assessing the water quality of this river. Taguas et al. (2008) used principal component analysis to investigate the relationship between instantaneous discharge and daily one. NouriGheidari (2013) identified wells effective in determining the groundwater level of Gheidar plain using principal component analysis. Based on the results, it was found that by removing wells whose relative importance is less than 0.5, the coefficient of change of groundwater does not change much compared to when all wells are used. Vonberg et al. (2014) examined the factors affecting the increase of atrazine using principal component analysis in 60 wells located in shallow aquifers in Germany; in this study, they showed that agricultural use was the most important source of atrazine in these areas. Akbarzadeh et al. (2016) optimized the groundwater quality-monitoring network of Mashhad aquifer using spatial-temporal modeling. The results showed that out of 287 wells in the region, 111 wells were sufficient as stations to monitor the quality of groundwater resources in the aquifer of Mashhad. BabaeiHessar et al. (2016) identified wells effective in determining groundwater level in the Urmia plain. In this study, minor wells were identified using principal component analysis. Considering the results, it was found that by removing minor wells, the number of which was about half of the total wells, the coefficient of variation of the water level was reduced by 50% and the error of determining the water level was less than 15%. In a study, Kavusi et al. (2019) used PSO (Particle Swarm Optimization) algorithm to determine the optimal number and position of observation wells. The obtained results showed that the number of observation wells was equal to 28 rings (42 observation wells) which indicated a 55% reduction in the number of piezometers compared to the initial state. In a study, Farpoor et al. (2019) numerically simulated the trend of chromium changes in the Birjand plain aquifer. The outcomes revealed that changes in the concentration of this pollutant depended on fluctuations in groundwater level and increasing the water level could reduce the amount of chromium in the aquifer. In the research conducted by Khashei Siuki et al. (2021) in the field of chrome monitoring network design of Birjand plain aquifer, which was done using PCA method, it was shown that out of 25 wells in the study area, 15 wells could be introduced as chromium groundwater index well in Birjand plain.

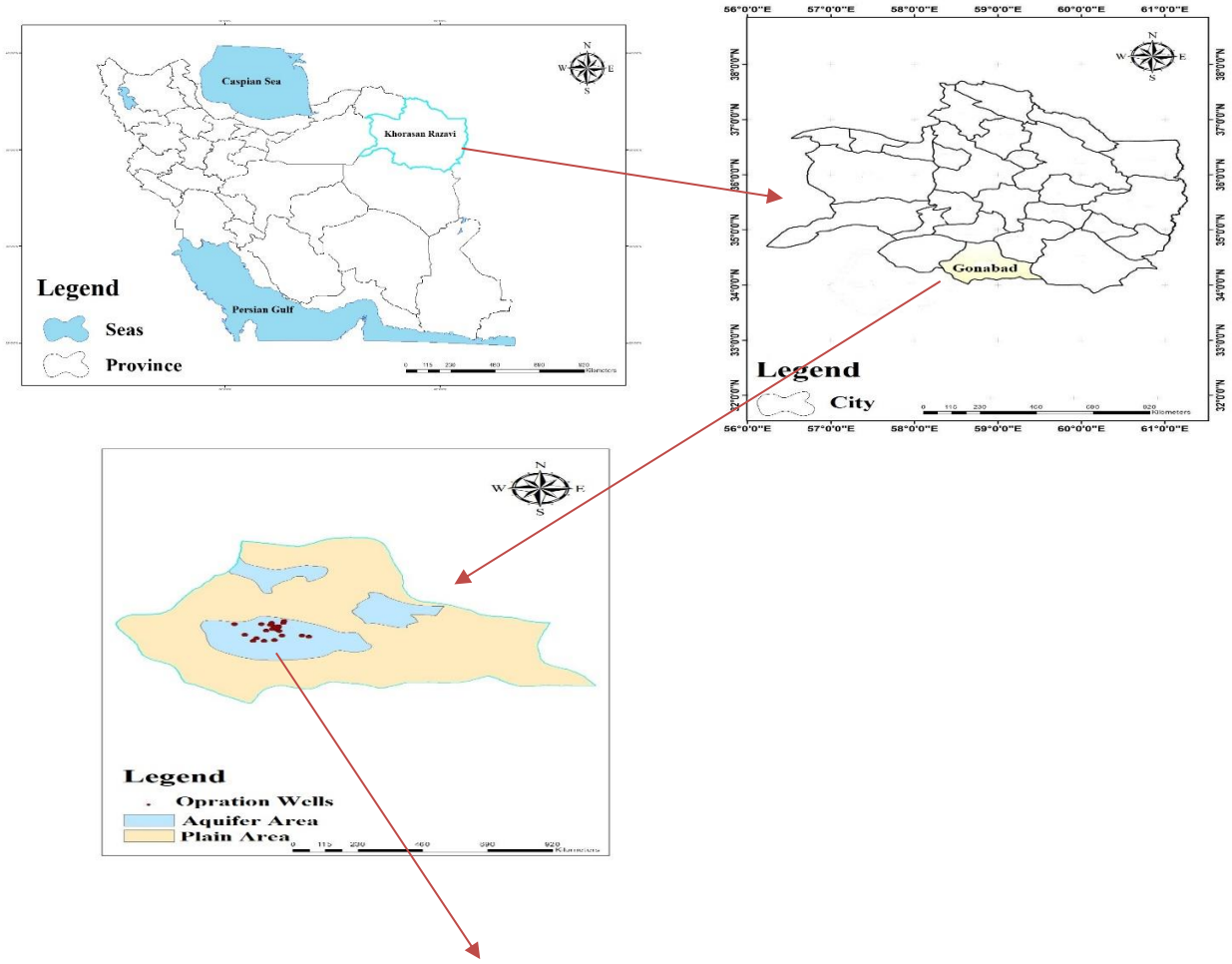
So far, studies have been conducted in the field of design of groundwater or surface water quality monitoring network, the main purpose of which was to select the location of the sampling station and reduce their number without increasing the measurement error. Various methods such as Khodaverdi et al. (2020), Kavusi et al. (2019) and Rezaei et al. (2015) have presented for this purpose that most of these studies are try to optimize the monitoring network. On the other hand, in the field of using PCA method, researchers such as Rahnama and Sayari (2019), Alves et al. (2018) and Zhao et al. (2012) have done and in these studies and the other ones, PCA method has often been used to identify the most important factors changing water quality. Therefore, the study of research records conducted in Iran and other regions shows that less attention has been paid to the identification of qualitative sampling wells. Accordingly, in this study, principal component analysis was used as a data reduction method to determine the relative importance of qualitative wells in Gonabad plain and finally effective qualitative

sampling wells in the aquifer of this plain were determined by PCA method.

2- Materials and methods

Area of study

The watershed studied in eastern Iran is located in the southern part of Khorasan Razavi province. Geographically, the plain under study is located in the range of 58 º19' to 59 º 01' east longitudes and 34º 03' to 34º 22' north latitude. The average annual rainfall in this plain is 128.9 mm per year. Figure (1) shows the geographical location of the study area and the distribution of water wells in the area. In this study, the statistics of 22 wells supervised by the Ministry of Energy have been used. To analyze the main components, the annual data of calcium (Ca^{2+}), magnesium (Mg^{2+}), sodium (Na^{+}), electrical conductivity (EC), total soluble solids (TDS), chlorine (Cl^{-}), sodium adsorption ratio (SAR), bicarbonate (HCO_3^{-}), sulfate (SO_4^{2-}), hardness (TH), acidity (pH) of groundwater of these wells, which was recorded from 2007 to 2016, have been used.



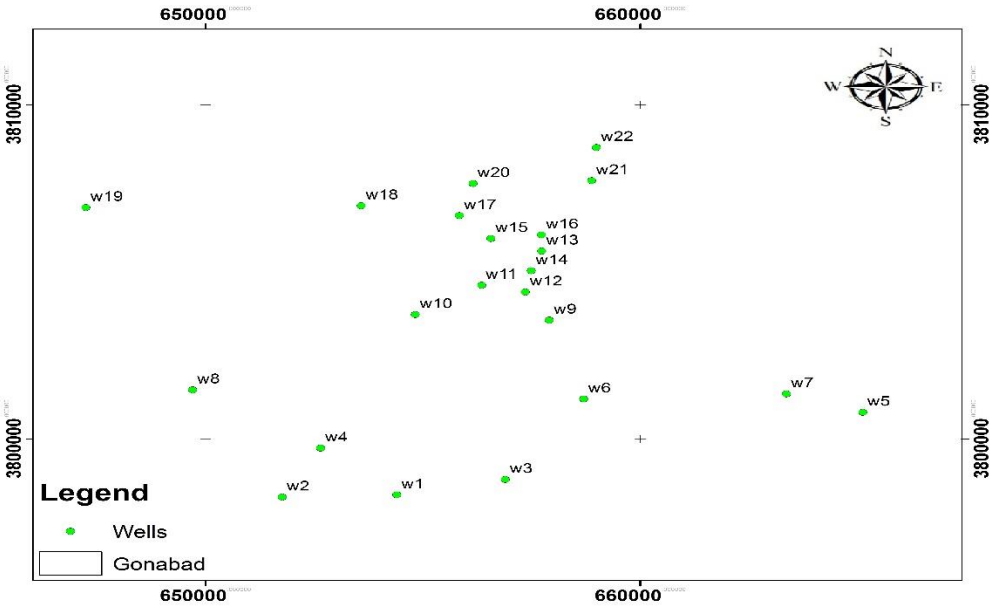


Fig. 1- Location of the study area and well

Principal Component Analysis (PCA)

PCA is a multivariate statistical method that can be used to reduce the complexity of analyzing the initial variables of the problem in cases where we have a large amount of information and plan to better interpret them (Noori et al., 2009). The purpose of principal component analysis is to reduce the size of the data (reducing the number of variables) by summarizing a large number of primary variables into a smaller number of principal components. For this purpose, the variance in multivariate data is broken down into components; as the first principal component (PC_1) justifies as much variance in the data as possible. The second principal component (PC_2) justifies the largest possible variance after the first component and to the end. The extracted components have two characteristics: a) they are perpendicular to each other (independent of each other), b) the total number of principal components is equal to the number of primary variables. The important thing about principal components is that these components are obtained by rotating the coordinate axes in the direction of maximum variance so that the angle between the coordinate axes will not change after the period (Bazrafshan and Hejabi 2017).

To calculate the importance of each well, the correlation coefficient between the main components and the observed data is used. The correlation coefficient of the well with the principal component is obtained from Equation (1).

$$Cor\left(z_j, x_i\right) = a_{i,j+1}^{T} \lambda_j a_{i,j} \quad (1)$$

Where, ai, j is the i element of the main component of j. The higher the coefficient, the higher the relative importance of the well is (Sanchez-Martos et al. 2001).

Generally, the number of wells (p) should be less than or at most equal to the number of observations (n) (which is the same as the number of statistical years) (Petersen 2001). In this study, to calculate the relative importance of each well for each well, 10 wells (equal to 10 statistical years) with the closest neighborhood to the desired well were identified. For example, to monitor well W_{14}, the SAR parameter uses 10 adjacent wells, namely W_{12}, W_9, W_{11}, W_{10}, W_{13}, W16, W_{15}, W_{17}, W_{20} and W_{18}. Therefore, there will be a 10 × 10 matrix for each well. It should be noted that the monitoring does not use the well data itself, but only 10 adjacent wells. Then, for each well, the principal component analysis was performed once to determine the correlation coefficient of each well with the principal component. In the selection of effective wells, wells with a correlation coefficient of less than 0.75 were eliminated (Ouyang, 2005). Thus, for each principal component analysis, a number of wells were identified as effective wells, and finally, the number of times each well participated in the analysis, as well as the number of times it was identified as an effective well was also determined. Equation (2) was used to

determine the relative importance of each well.

$$\text{Relative importan} = \frac{N}{n} \quad (2)$$

This ratio indicates the importance of each well compared to other wells, in which N is the number of times that each well is considered as an effective well and n is the number of times that each well has participated in the analysis. The greater the relative importance of a well, the greater the impact on monitoring. In order to investigate the effect of removing each well from the calculations, two criteria of coefficient of variation and monitoring error were used. Using Equation (3), the amount of monitoring error for removing ineffective wells at a given threshold can be calculated by comparing the average of wells at that threshold with the average of all wells (Gurunathan and Ravichandran, (1994).

$$Error = \frac{(m_n - m_0)}{m_0} \times 100 \quad (3)$$

Where m_n is the average value for the removal of wells calculated according to the relative importance, m_0 is the average value of all existing wells. In this research, to implement this method, SPSS statistical software. Ver 19 was used.

3- Results and discussion

As Table 1 shows, for the SAR parameter, the two components PC_1 and PC_2 have more variance. According to Table (1), wells W_{12}, W_9, W_{11}, W_{10}, W_{13}, W_{16}, W_{15} and W_{20}, which have a correlation coefficient above 0.75, can be selected as effective wells in monitoring well W_{14}. This analysis is performed to evaluate the effect of all wells and determine the number of times each well is affected. Then, according to the specific number of times a well is involved in the analysis of the main components, the relative importance of each well is calculated. Table (2) shows the relative importance of each well for the study area. The higher the rank, the more important it is.

If the threshold is equal to 0, 0.2, 0.4, 0.6, 0.8 and 1 for ranking wells (NouriGheidari, 2013), at threshold 0, all wells are included in the analysis and at threshold 1, wells with a rank of 1 are accepted. According to Table (2) about the SAR parameter, there are two wells with a rank of 1. This means that of all the wells, wells W_{10} and W_{16} are the most important and as many as participated in the analysis, are known as effective wells. The number of effective wells for each threshold for the SAR parameter will be 22, 18, 14, 11, 7 and 2, respectively.

Table 1- Correlation Coefficient Matrix of W_{14} Well Monitoring (SAR)

Wells (Wi)	Principal Component (PCj)	
	PC_1	PC_2
W_{12}	0.818	0.387
W_9	0.877	-0.253
W_{11}	0.895	-0.049
W_{10}	0.909	-0.194
W_{13}	0.639	-0.316
W_{16}	0.858	-0.384
W_{15}	0.852	-0.272
W_{17}	0.564	0.678
W_{20}	0.902	0.245
W_{18}	0.728	0.392

Table 2- Wells ranking based on principal component analysis (SAR)

Well	Number of times it has been identified as an effective well.	Number of times they participated in the analysis	Rank	Well	Number of times it has been identified as an effective well.	Number of times they participated in the analysis	Rank
W_{10}	16	16	1.00	W_1	4	8	0.50
W_{16}	12	12	1.00	W_4	3	6	0.52
W_{11}	19	20	0.95	W_{17}	5	12	0.42

W_{15}	12	13	0.92	W_2	2	6	0.33
W_{20}	11	12	0.92	W_{14}	5	15	0.33
W_9	13	15	0.87	W_3	2	8	0.25
W_7	4	5	0.80	W_{21}	1	5	0.20
W_{22}	3	4	0.75	W_6	1	8	0.13
W_{12}	15	21	0.71	W_{18}	1	9	0.11
W_5	2	3	0.67	W_{13}	1	14	0.07
W_8	3	5	0.60	W_{19}	3	4	0.00

To investigate the amount of error in selecting the number of effective wells, researchers have proposed to calculate the coefficient of variation for the remaining wells at each threshold and compare it with the coefficient of variation for all wells (NouriGheidari, 2013). This method is acceptable if it is assumed that the coefficient of variation increases with the removal of inefficient wells. In this way, the threshold at which the least difference in the coefficient of variation occurs is selected. Of course, it is important to note that the coefficient of variation does not always increase with the removal of ineffective wells (as in the present study) and can increase or decrease depending on the nature of the information of the removed wells. Consequently, in the present study, in order to select an acceptable threshold level, in addition to calculating the coefficient of variation, the average is used in calculating the error instead of the coefficient of variation (BabaeiHessar et al., 2016). Figure (2) shows the average coefficient of variation versus the threshold (SAR parameter). In this way, the coefficient of variation for each statistical year was calculated and after estimating its average value, was plotted against the threshold. According to Figure (2), at the threshold of 0.2, the coefficient of variation decreases. Considering that the average value of all wells (zero threshold) is equal to 9.28. At the threshold of 0.2 with the removal of two wells W_6, W_{18}, W_{13} and W_{19} in which the average values are higher than the average values of all wells and the coefficient of change suddenly decreases significantly. In addition, at the threshold of 0.4 to 1, with the elimination of wells that have a higher average, the coefficient of change has decreased. Elimination of insignificant wells continues until the average coefficient of variation is not high. To ensure the results, the error value was also calculated for each threshold. As shown in Figure (3), the threshold error rate of 0.6 is equal to 1.04 percent, i.e. by removing 11 less important wells; the groundwater level estimation error increases 1.04 percent compared to the situation where all wells are used. The error value increases from the threshold of 0.6 and above as well. Therefore, based on the monitoring error, the optimum threshold is 0.6. Therefore, 11 wells that remain at this threshold can be considered as effective wells in monitoring the SAR parameter of Gonabad plain.

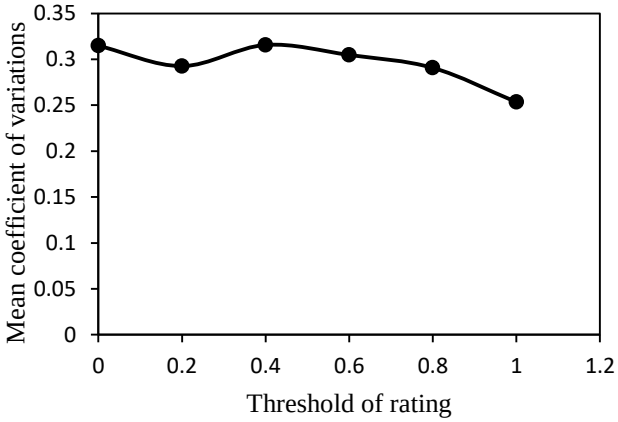


Fig. 2- Average coefficient of variations against the threshold of rating parameter SAR

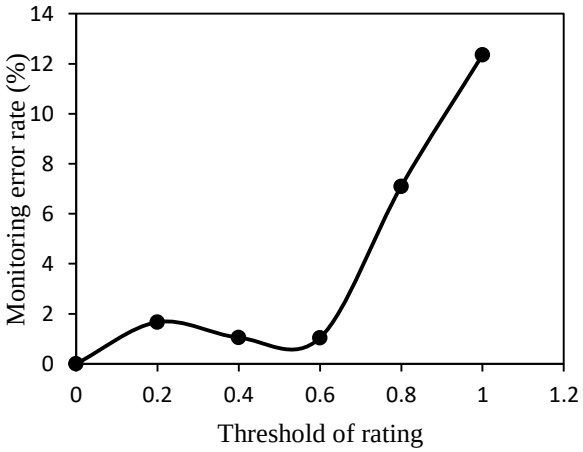


Fig. 3- Monitoring error rate against the rating threshold parameter SAR

The same steps were repeated for other qualitative parameters studied in this study.

Finally, 18, 15, 19, 11, 16, 14, 14, 15, 15, and 18 wells were respectively selected as effective wells in the parameters of Ca, Cl, EC, Mg, Na, pH, HCO_3, SO_4, TDS and TH that could play an important role in reducing sampling costs. Finally, after examining the importance of each well, common wells in all quality parameters studied in this study, namely wells W_4, W_9, W_{12}, W_{16}, W_7, W_8 and W_{10} were introduced as effective wells in terms of all parameters. The study conducted by BabaeiHessar et al. (2016) showed that by removing less significant wells in determining the groundwater depth of the Urmia plain, which was about half of the total number of wells (12 wells), the coefficient of variation was reduced by 50% and the error of determination The water table was less than 15%. NouriGheidari (2013) showed that by removing wells whose relative importance was less than 0.5, the error in determining the groundwater level of the Qeydar plain did not increase much and its value would be less than 13%.

4- Conclusion

In many studies, it is important to identify important sampling points in terms of reducing sample size and saving time and money. Principal component analysis is one of the methods that can be used to summarize data and reduce sampling points. In this way, by identifying more wells that are important and eliminating inefficient wells, it is possible to save a lot of money and time. In this study, groundwater qualitative parameters of different wells in Gonabad plain were investigated. Then, using the analysis of the main components, the relative importance of each well in each of the studied qualitative parameters of groundwater in this plain was calculated. By performing principal component analysis, the relative importance of each well was calculated between 0 (for inefficient wells) to 1 (for fully effective wells). The results showed that in general, out of 22 wells in the study area, 7 wells were introduced as groundwater qualitative index wells in Gonabad plain, which had a good distribution in the region.

5- Acknowledgments

We would like to thank the dear professors and friends who helped us write this research.

6- Conflicts of Interest

No potential conflict of interest was reported by the authors.

7- References

Akbarzadeh, M., Ghahraman, B., and Davary, K. (2016). Optimization of Groundwater Quality Monitoring Network in Mashhad City Aquifer Using Spatial-Temporal Modeling. *Journal of Iran- Water Resources Research*, 12(1): 133-144.

Alves, J. P. H. A., Fonseca, L. C., Chielle, R. S. A. and Macedo, L. C. B. (2018). Monitoring water quality of the Sergipe River basin: an evaluation using multivariate data analysis. Revista Brasileire de Recursos Hidricos Brazilian. *J. Water. Resour*, 23, 1-12.

Babaeihessar, S., Hamdami, Q. and Ghasemieh, H. (2016). Identify the Effective Wells in Determination of Groundwater Depth in Urmia Plain Using Principle Component Analysis. *J. Water. Soil*, 31, 10-50.

Bazrafshan, J. and Hejabi, S. (2017). Drought Monitoring Methods. University of Tehran Press. 224 pp.

Farpoor, F., Ramezani, Y. and Akbarpour, A. (2019). Numerical Simulation of Chromium Changes Trend in Aquifer of Birjand plain. Iran. *J. Irrig. Drain*, 12(5): 1203-1216.

Jolliffe, I. T. (2002). Principal Component Analysis. Springer series in statics, ISBN 978-0-387-95442-4.

Helena, B., Pardop, R., Vega, M., Barrado, E., Manuel, J., and Fernandez, L. (2000). Temporal evolution of groundwater composition in an alluvial aquifer by principal component analysis. *Journal of Water Research*, 34(3): 807-816.

Gurunathan, K., and Ravichandran, S. (1994). Analysis of water quality data using a multivariate statistical technique- a case study. IAHS Pub, 219.

Kavusi, M., Khasheisiuki, A., Porrezabilondi, M. and Najafi, M. H. (2019). Application of New LSSVM-PSO Optimization-Simulation Model in Designing Optimal

Groundwater Level Network Monitoring. Iran. *J. Ecohydro*., 5, 1306-1319.

Khashei Siuki, A., Shahidi, A. and Rahnama, S. (2021). Comparison of Birjand aquifer chromium monitoring network using principal component analysis (PCA) and entropy theory. Environ. *Water Eng*., 7(2), 209–220. DOI: 10.22034/jewe.2020.254396.1448

Khodaverdi, M., Hashemi, S. R., Khasheisiuki, A. and Porrezabilondi, M. (2020). Optimal Design of Groundwater-Quality Sampling Networks with MOPSO-GS (Case Study: Neyshabour Plain). J. Water. *Irrig. Manag*. (J. Agric.), 9, 199-210.

Lucas, L. and Jauzein, M. (2008). Use of principal component analysis to profile temporal and spatial variations of chlorinated solvent concentration in groundwater. *Environmental Pollution*, 151: 205-212.

Nguyan, T. T., Nakagawa, A. K., Amaaguchi, H. and Gilbuena, R. (2013). Temporal chenges in the hydrochemical facies of groundwater quality in tow main aquifers in Hanoi. Vietnam, DOI: 10.5675/ICWRER_2013.

Noori, R., Abdoli, M. A., AmeriGhasrodashti, A. and JaliliGhazizade, M. (2009). Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: A case study of Mashhad. *Environmental Progress & Sustainable Energy*, 28(2): 249-58.

NouriGheidari, M. H. (2013). Determintion of Effective Wells to Monitor the Ground Water Level Using the Principal Components Analysis. Journal of Sciences and Technology of Agriculture and Natural Resources, *Water and Soil Sciences*, 17(64): 149-158.

Ouyang, Y. (2005). Evaluation of river water quality monitoring stations by principal component analysis. *Water research*. 39: 2621-2635

Petersen, W. (2001). Process identification by principal component analysis of river water-quality data. *Ecological Modelling. Model*.138: 193-213.

Rahnama, S. and Sayari, N. (2019). Survey and Trends of Chemical Water Quality Parameters of Tajan River Water Quality Using Principal Component Analysis and Aqua Chem Software. Human. *Enviro*, 48, 13-25.

Rezaei, E., Khasheisuki, A. and Shahidi, A. (2015). Design of Groundwater Level Monitoring Network, Using the Model of Least Squares Support Vector Machine (LS-SVM). Iran. *J. Soil. Water. Res*., 45, 389-396.

Sanchez- Martos, F., Jimenez- Espinosa, R. and Pulido- Bosch, A. (2001). Mapping groundwater quality variables using PCA and geostatistics: a case study of Bajo Andarax, southeastern Spain. *Hydro. Sci. J*., 46, 227- 242.

Siyue, L. (2009). Water quality in the upper Han River, China: The impacts of land use/land cover in riparian buffer zone. *Hazardous Materials*, 165(1): 317-324.

Taguas, E., Ayuso, L., Pena, A., Yuan, Y., Sanchez, M., Giraldez, V. and Pérez, R. (2008). Testing the relationship between instantaneous peak flow and mean daily flow in a Mediterranean Area Southeast Spain, *Catena*. 75(2): 129– 137.

Vonberg, D., Vanderborght, J., Cremer, N., Pütz, T., Herbst, M. and Vereecken, H. (2014). 20 years of long-term atrazine monitoring in a shallow aquifer in western Germany. *Water Research*, 50: 294–306.

Zhao, Y., Xia, X. H., Yang, Z. F. and Wang, F. (2012). Assessment of water quality in Baiyangdian Lake using multivariate statistical techniques. *Proc. Enviro. Sci*., 13, 1213-1226.