Chapter 22
Groundwater Quality Assessment Using Principal Component and Cluster Analysis

Ahmed Garba, Ahmed Muhd Idris, and Jibrin Gambo

22.1 Introduction

The quality of groundwater is a very serious concern today, either naturally occurring processes or human actions may have a significant influence on the water quality of aquifer that may confine its usage. Understanding the groundwater characteristics is essential for groundwater management. Groundwater quality is shaped by natural and human activities such as climate, geology of the surrounding and agricultural practices. A poor quality of water threatens human health and as a result affects economic development and social prosperity (American Public Health Association [APHA] 2005).

Groundwater usually contains very low levels of trace metals depending upon the composition and the dissolution of the rock which is in interaction with the aquifer (American Society for Testing and Material 2004). The use of groundwater contaminated with trace metals may present environmental and public health risk in the city, depending on the contamination status (Bartram and Balance 1996).

The application of statistical techniques for the study of groundwater data provides an authentic understanding of aquifer and ecological condition and leads to the recognition of the potential sources that determine the groundwater system. This of course will serve as a tool for authentic and good management of water resources and speedy results on pollution problems (Belkhiri et al. 2011).

A. Garba (✉) · A. M. Idris
Department of Science Laboratory Technology, Binyaminu Usman Polytechnic, Hadejia, Nigeria
e-mail: gaahmed@bupoly.edu.ng

J. Gambo
School of General Studies, Binyaminu Usman Polytechnic, Hadejia, Nigeria

M. Sherif et al. (eds.), *Water Resources Management and Sustainability*, Water Science and Technology Library 121, https://doi.org/10.1007/978-3-031-24506-0_22

22.2 Principal Component Analysis (PCA)

Large datasets are increasingly widespread in many disciplines. In order to interpret such datasets, methods are required to drastically reduce their dimensionality in an interpretable way, such that most of the information in the data is preserved. Many techniques have been developed for this purpose, but principal component analysis (PCA) is one of the oldest and most widely used. Its idea is simple—reduces the dimensionality of a dataset, while preserving as much 'variability' (i.e. statistical information) as possible. Although it is used, and has sometimes been reinvented, in many different disciplines it is, at heart, a statistical technique and hence much of its development has been by statisticians.

This means that 'preserving as much variability as possible' translates into finding new variables that are linear functions of those in the original dataset, that successively maximize variance and that are uncorrelated with each other. Finding such new variables, the principal components (PCs), reduces to solving an eigenvalue/eigenvector problem. The earliest literature on PCA dates from Pearson and Hotelling, but it was not until electronic computers became widely available decades later that it was computationally feasible to use it on datasets that were not trivially small. Since then its use has burgeoned and a large number of variants have been developed in many different disciplines. Substantial books have been written on the subject and there are even whole books on variants of PCA for special types of data. In the formal definition of PCA will be given, in a standard context, together with a derivation showing that it can be obtained as the solution to an eigen problem or, alternatively, from the singular value decomposition (SVD) of the (centred) data matrix. PCA can be based on either the covariance matrix or the correlation matrix. The choice between these analyses will be discussed. In either case, the new variables (the PCs) depend on the dataset, rather than being pre-defined basis functions, and so are adaptive in the broad sense. The main uses of PCA are descriptive, rather than inferential; an example will illustrate this.

PCA stands for Principal Component Analysis. It is one of the popular and unsupervised algorithms that has been used across several applications like data analysis, data compression, de-noising, reducing the dimension of data and a lot more. PCA analysis helps you reduce or eliminate similar data in the line of comparison that does not even contribute a bit to decision making. You have to be clear that PCA analysis reduces dimensionality without any data loss. Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of "summary indices" that can be more easily visualized and analyzed.

Principal component analysis today is one of the most popular multivariate statistical techniques. It has been widely used in the areas of pattern recognition and signals processing and is a statistical method under the broad title of factor analysis.

PCA forms the basis of multivariate data analysis based on projection methods. The most important use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers. This overview may uncover the relationships between observations and variables, and among the variables. PCA is a very flexible tool and allows analysis of datasets that may contain, for example, multicollinearity, missing values, categorical data, and imprecise measurements. The goal is to extract the important information from the data and to express this information as a set of summary indices called principal components.

Principal Component Analysis helps you find out the most common dimensions of your project and makes result analysis easier. Consider a scenario where you deal with a project with significant variables and dimensions. Not all these variables will be critical. Some may be the primary key variables, whereas others are not. So, the Principal Component Method of factor analysis gives you a calculative way of eliminating a few extra less important variables, thereby maintaining the transparency of all information. Is this possible? Yes, this is possible. Principal Component Analysis is thus called a dimensionality-reduction method. With reduced data and dimensions, you can easily explore and visualize the algorithms without wasting your valuable time. Therefore, PCA statistics is the science of analyzing all the dimensions and reducing them as much as possible while preserving the exact information.

The PCA method permits the connection between the variables to be discovered, whereby the dimensionality of the data sets can be reduced. It is a strong method for approaching pattern recognition. PCA illustrated virtually the significant parameters, which depict the whole data set rendering data reduction with minimum loss of original information. In PCA Eigen values are normally used to determine the number of components (PCs) that are retained for further analysis. Number of components is equal to the number of variables in PCA.

22.2.1 Properties of Principal Component

Technically, a principal component can be defined as a linear combination of optimally weighted observed variables. The output of PCA are these principal components, the number of which is less than or equal to the number of original variables. Less, in case when we wish to discard or reduce the dimensions in our dataset. PCs possess some useful properties which are listed below.

PCs are essentially the linear combinations of the original variables, the weights vector in this combination is actually the eigenvector found which in turn satisfies the principle of least squares.

The PCs are orthogonal, as already discussed.

The variation present in the PCs decreases as we move from the 1st PC to the last one, hence the importance.

22.2.2 Guidelines When to Use the Principal Component Method of Factor Analysis?

Sometimes, you may be clueless about when to employ the techniques of PCA analysis. If this is your case, the following guidelines will help you.

You'd like to reduce the number of dimensions in your factor analysis. Yet you can't decide upon the variable. Don't worry. The principal component method of factor analysis will help you.

If you want to categorize the dependent and independent variables in your data, this algorithm will be your choice of consideration.

Also, if you want to eliminate the noise components in your dimension analysis, PCA is the best computation method.

22.2.3 Application of Principal Component Analysis

You can find a few of PCA applications listed below.

PCA techniques aid data cleaning and data preprocessing techniques.

You can monitor multi-dimensional data (can visualize in 2D or 3D dimensions) over any platform using the Principal Component Method of factor analysis.

PCA helps you compress the information and transmit the same using effective PCA analysis techniques. All these information processing techniques are without any loss in quality.

This statistic is the science of analyzing different dimensions and can also be applied in several platforms like face recognition, image identification, pattern identification, and a lot more.

PCA in machine learning technique helps in simplifying complex business algorithms.

Since Principal Component Analysis minimizes the more significant variance of dimensions, you can easily denoise the information and completely omit the noise and external factors.

22.2.4 PCA Applicability to Groundwater Geochemistry Data

PCA is a multivariate statistical procedure designed to classify variables based on their correlations with each other. The goal of PCA and other factor analysis procedures is to consolidate a large number of observed variables into a smaller number of factors that can be more readily interpreted. In the case of groundwater, concentrations of different constituents may be correlated based on underlying physical and chemical processes such as dissociation, ionic substitution or carbonate equilibrium

Table 22.1 Principal component analysis result

	VR1	VR2	VR3	VR4	VR5
pH	– 0.103	– 0.266	– 0.153	– 0.037	– 0.58
Temp	0.066	– 0.057	0.005	0.877	0.181
EC	– 0.292	– 0.729	0.112	0.046	– 0.176
T/H	0.871	– 0.187	– 0.180	– 0.234	0.153
TDS	0.869	0.169	0.341	0.237	0.034
Ca	0.362	0.201	– 0.661	0.069	0.395
Mg	– 0.015	0.497	0.665	0.005	0.288
Cl^-	0.549	– 0.048	0.621	0.305	0.204
NO_3	0.891	0.244	– 0.003	0.063	– 0.179
Fe	– 0.242	0.607	0.109	– 0.483	– 0.178
Cu	0.811	0.285	0.317	0.203	– 0.045
Zn	– 0.144	– 0.027	– 0.08	– 0.043	0.906
Mn	0.346	0.037	0.780	0.050	0.001
Cr	– 0.154	– 0.859	– 0.093	0.116	– 0.127
Pb	– 0.03	0.168	– 0.128	– 0.782	0.289
Eigen values	4.54	2.615	1.974	1.624	1.051
% of variance	30.267	17.431	13.161	10.826	7.006
Cumulative %	30.267	47.698	60.858	71.684	78.690

reactions. PCA helps to classify correlated variables into groups more easily interpreted as these underlying processes. The number of factors for a particular dataset is based on the amount of non-random variation that explains the underlying processes. The more factors extracted, the greater is the cumulative amount of variation in the original data (Table 22.1).

The results indicated that a total five factors were extracted and rotated using the varimax which accounts for about 78% of the total variance, which can be considered used to identify the principal source of variation in the hydrochemistry. The factor loading was grouped according to the criteria of Liu et al. (2003) by which strong, moderate and weak loadings corresponds to absolute loading values of > 0.75, 0.75–0.50, and 0.50–0.30 respectively (Muhaya et al. 2021).

The first component loading accounted for 30.27% of the total variance showed higher loading for TH, TDS, NO_3^-, Cu and Cl^-, with significant contribution from Ca and Mn. It is reasonable to observe a strong positive loading of TH, TDS, NO_3^-, with Cl^-, which possibly results due to leachate from domestic waste release in some part of the area. The higher Cl^- and TDS is an indication that there is a possibility that the groundwater is polluted by sewage, or waste from refuge dumping site. This factor can be labeled as the hardness and anthropogenic.

Second factor (VR2) described 17.43% of total variance, constituting higher loadings for Fe, and higher negative loading for Cr and EC, with significant contribution from Mg. These parameters may be considered to results from the ionic dissolution during groundwater migration. This factor can be labeled as a natural process leading to heavy metal pollution. VR3 accounted for about 13.16% of total variance and constitute loading on Mn, Mg, Cl, and moderate negative loading on Ca with significance contribution from TDS and Cu. The major variables constituting this factor is related to the hydro chemical variables might be from minerals in the groundwater.

Fourth Factor loading (VR4) for dry season accounted for 10.83% of the total variance and represents a higher a loading for temperature and negative higher loading for Pb with significance contribution from Cl^-.

Fifth Factor loading (VR5) for the dry season accounted for 7.01% of the total variance which represent higher loading for Zn with significance contribution from Ca.

22.3 Cluster Analysis

The assumptions of cluster analysis techniques include homoscedasticity (equal variance) and normal distribution of the variables. However, an equal weighing of all the variables requires long transformation and standardization (z-scores) of the data. Comparisons based on multiple parameters from different samples are made and the samples are grouped according to their 'similarity' to each other. The classification of samples according to their parameters is termed Q-mode classification. This approach is commonly applied to water-chemistry investigations in order to define groups of samples that have similar chemical and physical characteristics. This is because rarely is a single parameter sufficient to distinguish between different water types. Individual samples are compared with the specified similarity/dissimilarity and linkage methods are then grouped into clusters. The linkage rule used here is Ward's method. Linkage rules iteratively link nearby points (samples) by using the similarity matrix. The initial cluster is formed by linkage of the two samples with the greatest similarity. Ward's method is distinct from all the other methods because it uses an analysis of variance (ANOVA) approach to evaluate the distances between clusters. Ward's method is used to calculate the error sum of squares, which is the sum of the distances from each individual to the center of its parent group. These form smaller distinct clusters than those formed by other methods.

Cluster analysis has been carried out to substitute the geo-interpretation of hydrogeochemical data. Cluster analysis has been useful in studying the similar pair of groups of chemical constituents of water. The similarity/dissimilarity measurements and linkage methods used for clustering greatly affect the outcome of the Hierarchical Cluster Analysis (HCA) results.

After a careful examination of the available combination of 197 similarity/dissimilarity measurements, it was found that using Euclidean distance (straight line distance between two points in c-dimensional space defined by c variables) as similarity measurement, together with Ward's method for linkage, produced the most distinctive groups. In these groups each member within the group is more similar to its fellow members than to any other member from outside the group. The HCA technique does not provide a statistical test of group differences; however, there are tests that can be applied externally for this purpose. It is also possible in HCA results that one single sample that does not belong to any of the groups is placed in a group by itself. This unusual sample is considered as residue. The values of chemical constituents were subjected to hierarchical cluster analysis. Based on the indices of correlation coefficients, similar pairs groups of chemical constituents have been linked. Then the next most similar pairs of groups and so on, until all the chemical constituents have been clustered in a dendrogram by an averaging method.

Cluster analysis (CA) is a statistical tool to sort the rightful groups of data according to their similarities to each other. It include broad suite of techniques designed to find groups of similar items within a data set. Hierarchical methods usually give rise to a graphical output called a dendrogram or tree that demonstrates this hierarchical clustering pattern. It categorizes the objects into the stratum on the basis of similarities within a class and dissimilarities between different classes. The result of CA helps in translating the data and suggest pattern.

22.3.1 Requirements of Clustering in Data Mining

The following points throw light on why clustering is required in data mining:

Scalability—We need highly scalable clustering algorithms to deal with large databases.

Ability to deal with different kinds of attributes—Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.

Discovery of clusters with attribute shape—The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.

High dimensionality—The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.

Ability to deal with noisy data—Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

Interpretability—The clustering results should be interpretable, comprehensible, and usable.

22.3.1.1 Clustering Methods

Clustering methods can be classified into the following categories:

- Partitioning Method
- Hierarchical Method
- Density-Based Method
- Grid-Based Method
- Model-Based Method
- Constraint-Based Method.

22.3.1.2 Applications of Cluster Analysis

Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.

Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.

In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.

Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.

Clustering also helps in classifying documents on the web for information discovery.

Clustering is also used in outlier detection applications such as detection of credit card fraud.

As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

22.3.1.3 Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

22.3.1.4 Clustering Groundwater Geochemistry Parameters

Classification of wells according to their water quality can provide useful information for the users. Complex processes control the distribution of water quality parameters in groundwater, which typically has a large range of chemical composition. The ground water quality depends not only on natural factors such as the lithology of the aquifer, the quality of recharge water and the type of interaction between water and aquifer, but also on human activities, which can alter these groundwater systems either by polluting them or by changing the hydrological cycle. Sophisticated data analysis techniques are required to interpret groundwater quality effectively Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

Cluster analysis is used to discover the correlation among the different sampling points. Cluster analysis yields either cluster or groups on the basis of similarity or dissimalarity of variables. Therefore, the cluster analysis is to discover a system of organized observations where a number of groups/variables share properties in common. From the dendrogram (Fig. 22.1), the outcome of cluster analysis lead to the grouping of the boreholes into four (4) distinct groups or clusters based on their similarities (Table 22.2).

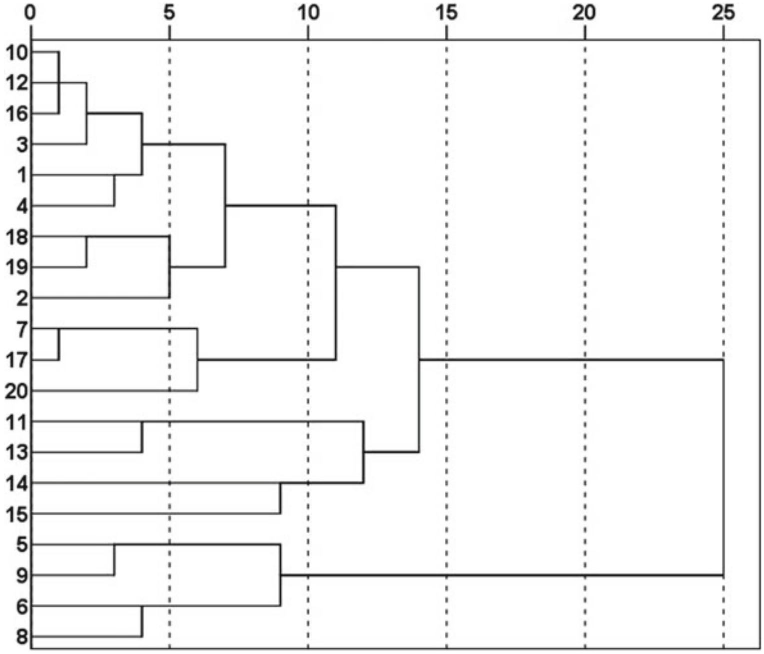

Fig. 22.1 Dendrogram of water sampling sites

Table 22.2 Mean values of geochemistry data

Cluster	pH	Temp	EC μS/cm	TH mg/L	TDS mg/L	Ca mg/L	Mg mg/L	Cl^- mg/L	NO_3^- mg/L	Fe mg/L	Cu mg/L	Zn mg/L	Mn mg/L	Cr mg/L	Pb mg/L
1 (n = 9)	6.53	27.8	235	154	207	***53.8***	30.4	1.98	7.35	0.49	0.48	0.34	0.75	0.11	1.57
2 (n = 3)	6.25	27.9	597	151	210	28	28	4.57	1.25	***0.53***	0.20	***0.87***	0.98	0.17	***1.70***
3 (n = 4)	6.38	27.8	286	***301***	***588***	44.8	***32.5***	***6.79***	***29.7***	0.29	***3.53***	0.05	***5.56***	0.08	1.39
4 (n = 4)	6.76	27.8	***1116***	145	91	19.7	20.1	0.28	2.78	0.26	0.01	0.03	0.93	***0.18***	0.99

The underlined values indicated “the value is the lowest among all the clusters”

Samples from cluster 1 is composed of nine (9) boreholes 1, 2, 3, 4, 10, 12, 16, 18 and 19, and constitute 45% of the water samples, and are characterized by high concentration of Ca, and low concentration of Mn and EC in all clusters. Sample from cluster 2 is composed of three (3) boreholes 7, 17 and 20 and constitute 15% of the water samples, and are characterized by high concentration of Fe, Zn and Pb and low concentration of NO_3^- in all clusters. Samples from cluster 3 is composed of four (4) boreholes 11, 13, 14 and 15 and constituted 20% of the water samples, and are characterized by high concentration of T/H, TDS, Mg, Cl^-, NO_3^-, Cu, Mn, and low concentration of Cr in all clusters. The higher values of EC, Total Hardness in some boreholes is obvious because of the solvent action as the water comes in contact with soil and rock is capable of dissolving Ca, Mg and other ions that promote EC and hardnessConductivity or specific conductance is a measure of the ability of water to conduct an electric current. It is sensitive to variations in dissolved solids, mostly mineral salts. The degree to which these dissociate into ions, the amount of electrical charge on each ion, ion mobility and the temperature of the solution all have influence on conductivity.

Samples from cluster 4 is composed of four (4) boreholes 5, 6, 8 and 9 and constituted of 20% of the water samples and are characterized by high values of EC and high concentration Cr and low values of T/H, TDS and low concentration Ca, Mg, Cl^-, Fe, Cu, Zn and Pb in all clusters.

22.4 Conclusion

This study demonstrated the usefulness of the PCA and CA in the interpretation of groundwater quality data by providing useful information on the possible sources that influence the water system and gives guide on effective management of the water resource. PCA was employed to look into the source of each water quality parameters and generated five factors/components with 78.69% total variance, indicating the major variations are related to human action and natural processes. Cluster analysis results sorted the 20 boreholes into four statistically significant clusters based on their similarities. The concentrations of Mn, Cr, and Pb determined were above the maximum permissible limit (Nigerian Standard for Drinking Water Quality (NSDWQ) 2007; Orisakwe 2013). Numerous researchers have reported on adverse effects on human health due to exposure to some of these trace elements in drinking water, for Cr (Ramadan and Haruna 2019; Shrestha and Kazama 2007), Pb (Shrestha and Kazama 2007; Vetrimurugan et al. 2017).

Acknowledgements This is to acknowledge the Managements of the Tertiary Education Trust Fund (TETFund) Abuja, Nigeria and Binyaminu Usman Polytechnic, Hadejia, Jigawa State for sponsorship and approving to attend this conference in cash and kind, I am very much grateful.

References

American Public Health Association [APHA] (2005) Standard methods for the examination of water and waste water, 20th edn. Washington, DC, pp 65–68

American Society for Testing and Material (2004) Annual book of ASTM standards, water and environmental technology. 11.01, Water (i). ASTM International, West Conshohocken, PA, pp 79–170

Bartram J, Balance R (1996) Water quality monitoring: a practical guide to the design of fresh water quality and monitoring programmes. Chapman and Hall, London, pp 34–39

Belkhiri L, Boudoukha A, Mouni L (2011) A multivariate statistical analysis of groundwater chemistry data. Int J Environ Res 5(2):537–544

Liu CW, Lin K, Kuo YM (2003) Application of factor analysis in the assignment of groundwater quality in blackfoot disease area in Taiwan. Sci Total Environ 313(1–3):77–89

Muhaya BB, Mulongo SC, Kunyonga CZ, Mpomanga WA, Kalonda ME (2021) Assessment of trace metal levels of groundwater in Lubumbashi, Kampemba and Kamalondo communes of Lubumbashi city, Democratic Republic of Congo. J Environ Sci Eng A 10(1):9–25. https://doi.org/10.17265/2162-5298/2021.01.002

Nigerian Standard for Drinking Water Quality (NSDWQ) (2007) Nigerian industrial standard. NIS 554:2007 ICs 13.060.20

Orisakwe KU (2013) Challenge detection analysis of land uses in Hadejia town, Nigeria. Int J Appl Sci Technol 3(3):60–168

Ramadan JA, Haruna AI (2019) Health risk assessment from exposure to heavy metals in surface and groundwater resources with Barkin Ladi, North Central Nigeria. J Geosci Environ Prot 7(2):1–21. https://doi.org/10.4236/gep.2019.72001

Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ Model Softw 22:464–475

Vetrimurugan E, Brindha K, Elango L, Ndwandwe OM (2017) Human exposure risk to heavy metals through groundwater used for drinking water in an intensively irrigated river delta. Appl Water Sci 7:3267–3280. https://doi.org/10.1007/s13201-016-0472-6