Use of Principal Component Analysis in hydro-geochemical characterization of an aquifer along the Mediterranean coast of Turkey

Nurkan Karahanoglu

Department of Geological Engineering, Middle East Technical University
06 531, Ankara, TURKEY

This paper presents a research about statistical evaluation of water quality samples by using the method of Principal Component Analysis. Chemical analyses of the water samples collected from an aquifer in the Mediterranean coast of Turkey were used as input data in which the major anions and cations, pH, Electrical conductivity, Transmissivity and Lithology were considered as variables. Statistical analyses were aimed at investigating interrelations and similarities between the variables, studying hydro-geochemistry of the groundwater system and determining the most effective components in the multivariate data set. Additionally the effect of time was searched by using reduced data sets in order to lead the analysis towards saltwater intrusion in the region. As a result of the analyses correlation coefficients were calculated for different data sets and their transient variations were evaluated. Principal component analysis delineated two groups of water samples as those coming from saltwater and freshwater origins.

Introduction

Starting with Ghyben-Herzberg in late 1800s, the problem of seawater intrusion has been introduced into the literature and in recent years it became one of the major problems in coastal aquifers. Due to gentle behavior of the dynamic equilibrium between freshwater and saltwater these aquifers are under continual danger of seawater contamination. Increased water demand in resort regions triggers movement of seawater to intrude into aquifers and change freshwater quality. This requires optimal management of coastal aquifers which are to be guided by quality monitoring.

Hydrogeochemical investigations of coastal aquifers could yield valuable information about relations between seawater and freshwater in such regions. Careful analysis of the physical and chemical quality parameters and their spatial and temporal distributions is quite helpful in defining the level of seawater contamination. Analysis of water samples, collected from coastal aquifers could well be used to detect probable trends of deteriorating water quality. This brings interpretation of a multivariate dataset which includes distribution of the physical and chemical variables for multitude number of water samples. Classical methods, like Scholler graphs and Piper diagrams, were successfully used in interpreting quality datasets, however, such methods can no longer be applicable for analyzing multivariate datasets where both the physical and chemical variables are included for large numbers of water samples. Multivariate statistical methods, take their roles in performing such tasks and they are found to be very powerful.

Several studies can be mentioned where statistical methods have been extensively applied to analyze multivariate data sets in different fields of the earth sciences (Davis, 1986; Rock, 1988; Helsel and Hirsch, 1992, Brown, 1993; Istok et.al., 1993; Wen and Kung, 1993). In recent years several statistical applications have been noted as focused on the topics related with the environmental problems (Geiss et.al., 1991; Melloul and Collin, 1992; Xhoffer et.al., 1991; Sodestrom, 1992; Pohlmann, 1993; Zitko, 1994; Melloul, 1995).

Xhoffer and others (1991) accomplished a principal component analysis to study characterization of individual aerosol particles over the North Sea and the English channel. The samples were clustered by using the principal component analysis as, the marine-derived samples, CaSO4-rich samples and those related to high silica and sulphur abundance. Geiss and others (1991) performed a multivariate correlation analysis and compared time series of heavy metal concentrations at two points of a river.

They concluded that correlation analysis proved to be a useful tool for investigating internal and external connections of time and distance series. Melloul and Collin (1992) pointed the vital and complementary role of the Principal Component analysis in evaluating multivariate data sets of groundwater samples, especially in identification of relevant groups of water and the factors that bring about a change in their quality. They conducted a hydrogeochemical research in water quality factor identification and used the principal components to identify relevant water chemistry types and the factors that cause a change in water quality in the Dan metropolitan region of the coastal plain aquifer of Israel. Brown (1993) applied correlation, principal component and multiple regression analyses to laboratory chemical and petrographic data to assess the usefulness of these techniques in evaluating physical and hydraulic properties of carbonate rock aquifers in Pennsylvania. Zitko (1994) proved the usefulness of the principal component analysis in the examination of enzyme induction and contaminants, heavy metals in sediments and mussels, metals in sediments and corals, survey of organochlorine compounds in fish. He concluded that principal component analysis could provide a deeper insight into the structure of the data and help reaching conclusions those were not immediately obvious. Melloul (1995) developed a methodology to investigate the Nubian sandstone aquifer beneath the Sinai Peninsula (in Egypt) and the Negev Desert (in Israel) by using the principal component analysis. He emphasized that the method successfully utilized the principal component analysis in the description of a complex flow system, in the delineation of optimal operational zones, and in the characterization of groundwater flow paths.

In view of these studies, hydro-geochemistry of a coastal aquifer in Turkey has been investigated by using the principal component analysis in order to define saltwater and freshwater interactions in the region. Erzin groundwater basin is located along the Mediterranean coast (Fig. 1) and supplies freshwater to continuously increasing water demand of the irrigation. In respect to its potential the aquifer needs to be thoroughly investigated to clarify recent indications of water quality deterioration. For this purpose chemical analysis of the water samples, collected from three springs and 161 water wells during different observation periods, and some physical parameters of the aquifer were studied (Karahanoglu et al., 1995). A multivariate data set was formed with variables of major anion and cation concentrations, pH and electrical conductivity (EC) values, transmissivity and type of screened lithologies. In this form the data matrix collects spatial and temporal variations related to unknown number of combinations between physical and chemical variables, which are responsible from this realization. It would be almost impossible to interpret such a data set and uncover hidden interrelations between so many variables without using statistical techniques.

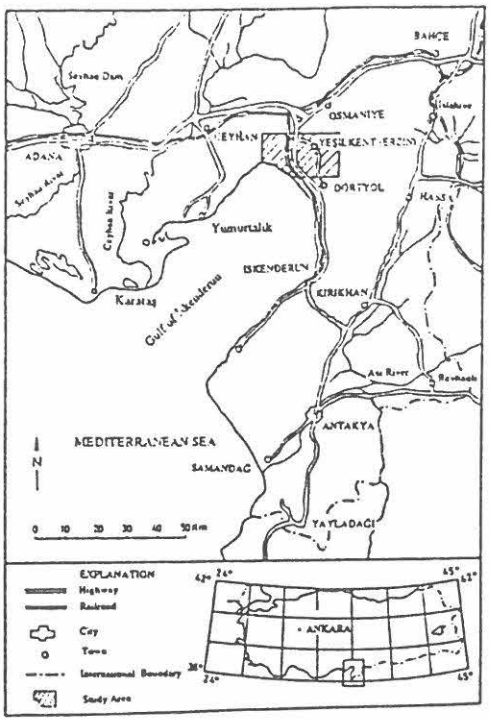

Fig. 1. Location map of the study area.

Hydrogeological setting of the study area

Owing to its great potential, Erzin aquifer has been extensively studied for the geological and hydrogeological aspects (Türkmen et.al., 1974; Doyuran, 1982). The basin was simulated for the flow mechanism (Karahanoglu et al., 1986) and a numerical model was proposed for the coupled mechanism of fluid flow and saltwater intrusion (Emekli et al., 1996).

The Erzin groundwater basin is formed of Tertiary and Quaternary deposits and stores water under unconfined conditions. Geological map of the area shows that impervious ophiolitic series form the basement and boundary along the eastern part. The northern boundary is occupied by well-cemented and well-compacted Miocene sandstone, conglomerate and marl alternations. These impervious units are unconformably covered by a thick conglomerate deposit with highly fractured and jointed nature. The estimated thickness of the conglomerate is about 1500 m and it forms the major aquifer where the hydraulic conductivities range between 10-50 m/day. In the northern and northwestern part of the Erzin plain olivine basalts are located and they show excellent storage and conduit properties. These basalts form the most productive aquifer in the basin and the hydraulic conductivities range between 50-150 m/day. The conglomerate and the basalt aquifers are estimated to be hydraulically connected (Doyuran, 1982) and these units supply water by means of 161 wells and three springs throughout the basin (Fig. 2).

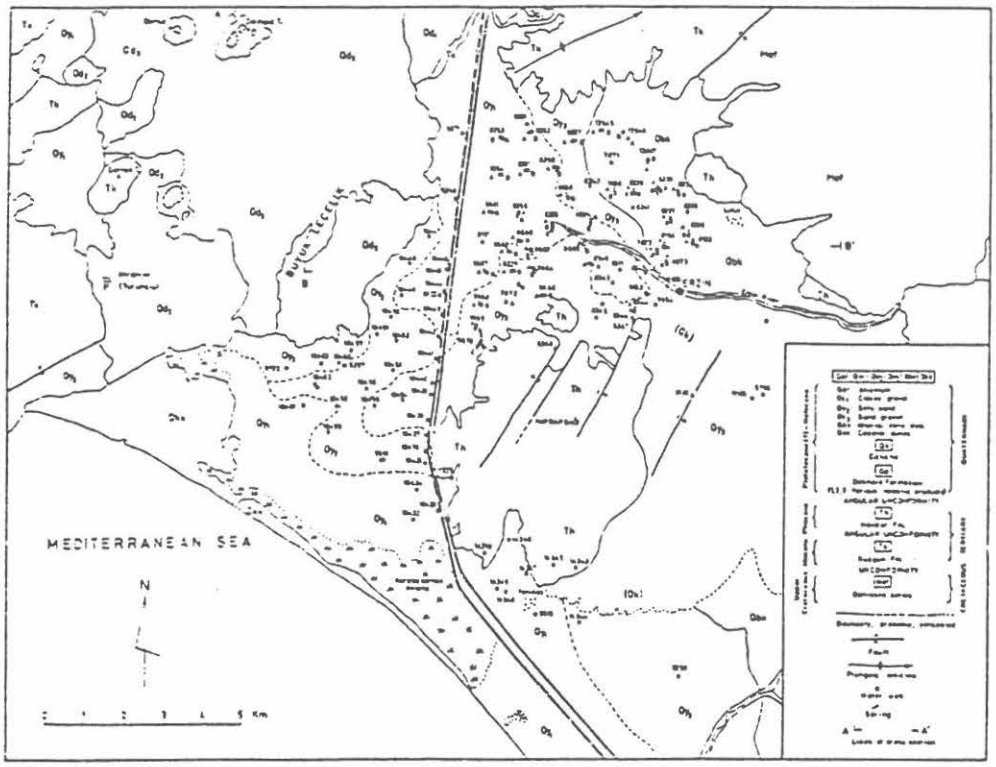

Fig. 2. Geological Map of the Erzin Plain Groundwater Basin and the water wells.

Method of Study

Samples have been collected from the water wells and springs and they have been analyzed for the major anions and cations as well as their pH and EC (electrical conductivity) values. Transmissivity coefficients and type of the screened lithologies were also collected for some of the wells. In this form a data matrix was formed by using chemical analysis of the samples and some of the physical parameters of the Erzin groundwater basin. This yields a multivariate data set with 316 number of observations for the variables of pH, EC, NaK, CaMg, CO3, HCO3, Cl, SO4, % Na, Trans (Transmissivity) and Lith (Lithology). In this form, the raw data includes all kind of information related to spatial and temporal variations which might result from unknown number of combinations between the variables. In order to interpret such a complex data set Principal Component Analysis was applied in order to uncover hidden interrelations between so many variables.

In the analysis the data matrix is manipulated to evaluate cumulative variances, the eigenvalues and the eigenvectors for extracting the principal components. These principal components (new set of variables) are formed by linear combination of the original variables according to their loadings on the cumulative variance. In the procedure they are determined by using the correlation matrices, and the coefficients of each principal component are evaluated accordingly. Hence the method aims at determining minimum number of such components that would account for most of the variance in the raw data (Davis, 1986).

Analyses begun with calculating the correlation coefficients in order to detect similarities between the variables. For this purpose the data set was separated into subgroups and the coefficients were calculated accordingly. Table 1 shows the correlations between the variables compiled for 1964-68

Table 1. Correlation coefficients for 1964-1968 data set.

	pH	EC	NaK	CaMg	CO3	HCO3	Cl	SO4	%Na
pH	1.0								
EC	-0.016	1.0							
NaK	0.004	0.77	1.0						
CaMg	-0.025	0.65	0.13	1.0					
CO3	0.14	-0.04	-0.08	0.03	1.0				
HCO3	-0.04	0.31	0.008	0.42	-0.05	1.0			
Cl	-0.002	0.82	0.79	0.39	-0.03	-0.19	1.0		
SO4	-0.003	0.60	0.44	0.50	-0.06	0.01	0.35	1.0	
%Na	0.06	0.40	0.77	-0.16	-0.12	0.0025	0.44	0.18	1.0

observation. The analysis was repeated for a reduced data set (80 sample points) with transmissivity and lithology in order to search for a probable interlink between the physical and the chemical parameters. These values indicate that there exists a close similarity between electrical conductance and the ions that show tendencies to form salts. EC is affected positively by most of the variables except CO3; NaK correlates very well with Cl and % Na as expected and it has correlation with SO4; CaMg correlates with SO4 and it has fair correlation with HCO3 and Cl. Fair similarity between Cl and SO4 should also be mentioned. The reverse relation between lithology and transmissivity is due to order of the symbols used in digitizing different lithologies and it indicates that transmissivity decreases with increasing lithology symbol (Table 2). From the correlation matrix, it is also seen that the chemical constituents have very low correlation with lithology and transmissivity. However, a weak similarity can be detected for CaMg and HCO3 variables, stating that the water samples taken from wells with greater transmissivities (basalt aquifer), contain more CaMg and HCO3 ions. The analysis was repeated for May 1993, October 1993 and June 1994 observations and similar correlations were obtained. Table 3 delineates time change of some of the correlation coefficients on the EC and NaK basis. It is interesting that on EC basis the correlation coefficients with NaK,

Table 2. Correlation coefficients for the selected samples with transmissivity and screened lithology.

	pH	EC	NaK	CaMg	HCO3	Cl	SO4	%Na	Trans	Lith
pH	1.0									
EC	-0.27	1.0								
NaK	-0.13	0.84	1.0							
CaMg	-0.31	0.71	0.24	1.0						
HCO3	-0.22	0.17	-0.16	0.48	1.0					
Cl	-0.18	0.89	0.89	0.47	-0.21	1.0				
SO4	-0.24	0.65	0.55	0.47	-0.09	0.48	1.0			
%Na	-0.05	0.51	0.80	-0.11	-0.08	0.54	0.32	1.0		
Trans	-0.05	0.18	0.008	0.29	0.33	0.06	0.04	-0.12	1.0	
Lith	0.02	-0.15	0.02	-0.28	-0.31	-0.06	0.08	0.17	-0.64	1.0

CaMg, Cl, and %Na increase, whereas the coefficients with HCO3 and SO4 decrease with time. This should point that the contribution of NaK and Cl ions (seawater constituents) to Electrical Conductivity (EC) of the water samples becomes more effective than that of HCO3 and SO4 ions (freshwater constituents). A similar trend is found for the coefficients computed on the NaK basis,

Table 3. Variation of correlation coefficients with time for different variable couples.

	NaK	CaMg	HCO3	Cl	SO4	%Na
EC						
1964-68	0.77	0.65	0.31	0.82	0.60	0.40
May 1993	0.71	0.92	0.19	0.82	0.18	0.52
Oct 1993	0.86	0.82	0.13	0.88	0.003	0.73
June 1994	0.89	0.83	-0.09	0.92	0.03	0.76
NaK						
1964-68		0.13	0.008	0.79	0.44	0.77
May 1993		0.38	-0.49	0.96	-0.31	0.96
Oct 1993		0.43	-0.26	0.99	-0.29	0.95
June 1994		-0.49	-0.51	0.99	-0.31	0.97

where the correlation coefficient with Cl increases to 0.99, however correlation with HCO3 and SO4 depict a decreasing trend from 1964-68 to June 1994. This could mean that due to seawater intrusion effect, the probability of NaK ion to form a compound with Cl becomes greater than the probability of forming NaK-HCO3 and NaK-SO4 compounds.

Principal Component Analysis

Principal component analysis formed two principal components for each data group by combining the original variables. Table 4 lists the principal components and the related contributions of the

Table 4. Principal component weights for different data sets.

Variable	1964-1968		May 1993		October 1993		June 1994		DATATL	
	PC-1	PC-2	PC-1	PC-2	PC-1	PC-2	PC-1	PC-2	PC-1	PC-2
EC	0.510	0.165	0.408	0.381	0.461	0.265	0.433	0.289	0.493	0.096
NaK	0.471	-0.301	0.498	-0.088	0.488	-0.101	0.480	-0.009	0.464	-0.183
CaMg	0.289	0.591	0.264	0.546	0.281	0.588	0.246	0.559	0.295	0.395
HCO3	-0.077	0.504	-0.190	0.533	-0.074	0.574	-0.237	0.539	0.013	0.463
Cl	0.457	-0.167	0.500	0.016	0.490	-0.078	0.479	0.040	0.462	-0.083
SO4	0.348	0.195	-0.134	0.466	-0.119	0.430	-0.155	0.538	0.358	0.038
%Na	0.312	-0.461	0.459	-0.219	0.459	-0.232	0.460	-0.144	0.322	-0.297
Lith									-0.051	-0.498
Trans									0.075	0.492

variables. Study of the table shows that the principal components (PC-1 and PC-2) are formed by different contributions. For 1964-68 data group, EC, NaK and Cl have more contributions to the first principal component (PC-1) whereas the second component (PC-2) is heavily weighted by CaMg and HCO3 ions. Transmissivity and lithology have no contribution to the first principal component (PC-1), whereas the transmissivity is found to have the maximum weight on the second principal component (PC-2) in addition to CaMg and HCO3 ions. May 1993, October 1993 and June 1994 data sets yield similar results with slight deviations. In general the first principal component (PC-1), is majorly contributed by EC, NaK, Cl and Na% and CaMg, HCO3 and SO4 repeatedly form the majority in the second principal component (PC-2).

Overall evaluation of the components delineates that some of the variables are interrelated. The first principal component (PC-1) collects the variables which are either seawater constituents (NaK, Cl, Na%) or those affected by seawater (EC); the second component (PC-2) shows that major contributions come from land-derived ions (CaMg, HCO3, SO4).

Fig. 3 is constructed to show scatter of the 1964-68 group data with respect to the first (PC-1) and the second principal components (PC-2). On the same graph the original variables are represented by lines, which reflect how each variable contributes to the components. Analysis of the variable lines indicates that there exist close relations between CaMg - HCO3 and NaK - Cl couples. SO4 and EC lie in between these implying that the electrical conductivity is affected by both of these couples. This plot supports to verify the above classification of the water samples, as two distinctive groups of CaMg and HCO3, and NaK and Cl variables are easily identified.

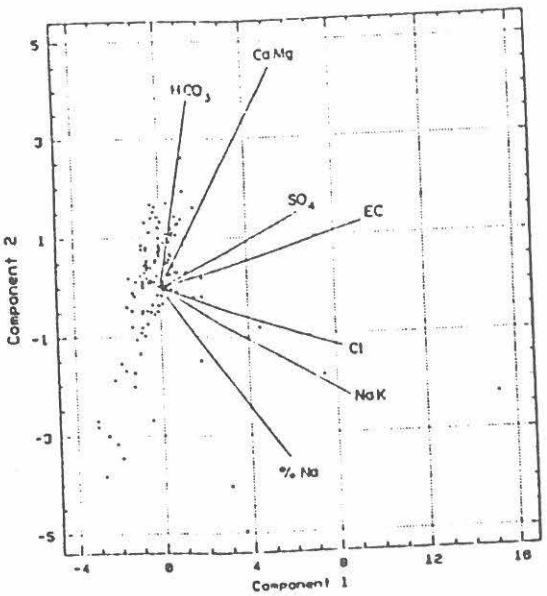

Fig. 3. Scatter of 1964-68 group data.

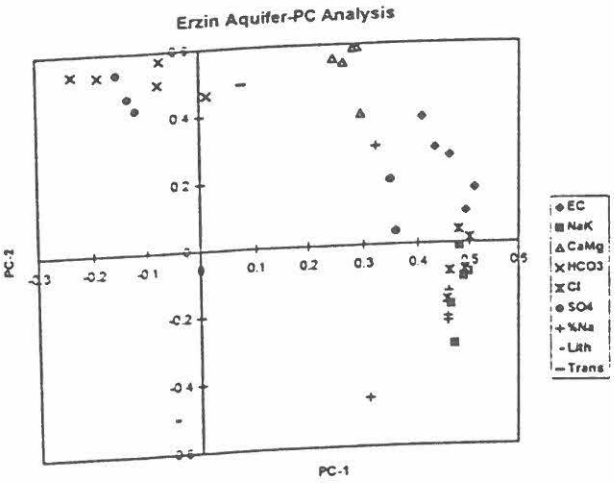

Fig. 4. Principal component weights.

A composite figure is formed to compare different data sets in terms of their component weights (Fig. 4). Principal component weights of the individual variables are grouped in this figure and sub-regions are defined. NaK, Cl, and CaMg variables define smaller areas due to their smaller variances in time. SO4 was dominantly controlled by the first principal component (PC-1) in 1964-68 group, whereas for the last three data groups it becomes majorly controlled by the second principal component (PC-2). Minor contribution of the first principal component on HCO3 variable loses its effect for the last three data groups. This simply means that HCO3 and SO4 ions become majorly controlled by the second principal component in recent observations. It is also observed that NaK - Cl and CaMg - HCO3 groups are located at the two extreme ends with EC sub-region in between. It is interesting that similar behavior is detected in the correlation analysis that verifies these relations.

Contributions of the lithology and transmissivity on the principal components were investigated by constructing a scatterplot with reduced data set (Fig. 5). It is found that majority of the sampling wells collect around the second component (PC-2) axis, and a few wells deviate from the major group. On the same graph the data points were replaced by corresponding lithologies and two different sub-groups were identified. These groups are formed by the wells tapping basalt formations and by those wells, which tap from conglomerate type lithologies. Wells screened to basalt have greater contributions to the second principal component since the basalt has higher coefficients of transmissivities (Fig. 3). The negative contribution of conglomerate-tapping wells to the second principal component is due to negative correlation between Lith and Trans variables (Table 4).

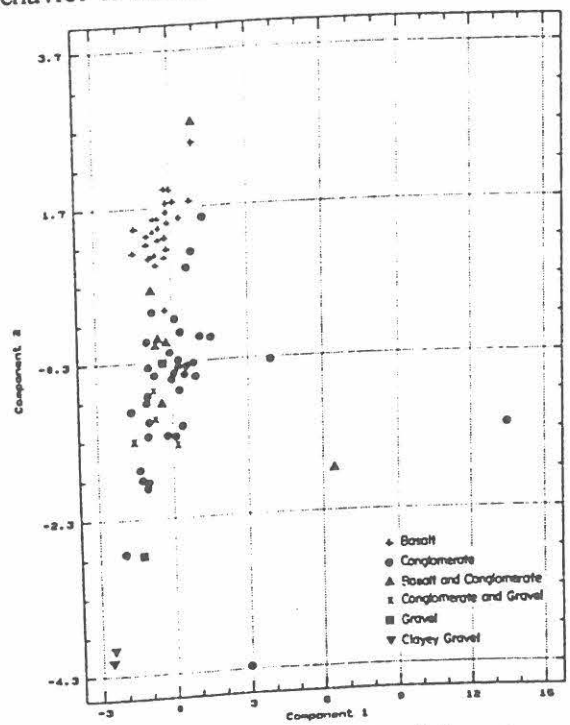

Fig. 5. Scatterplot for reduced dataset.

Conclusions

In this research the Erzin plain groundwater basin (Turkey) was extensively studied to investigate hydrogeochemistry of the aquifer and clarify interrelation between seawater and freshwater in the region. The research was accomplished by applying the correlation and the principal component analysis in order to identify seawater intrusion problem in the area. As a result of the investigation the following conclusions were drawn:

- Correlation analysis showed that some of the major ions have significant correlation and these correlations follow temporal change in the Erzin aquifer. This change is quite clear for NaK, Cl, HCO3 and SO4 ions. Correlation of the first two ions with the electrical conductance increases with time whereas the correlation of the next two follows a decreasing trend. This explains that NaK and Cl ions take the majority in controlling the electrical conductance of the water samples. CaMg is also determined as effective for controlling the electrical conductivity however its time change is not so clear. Among the variables pH and CO3 were determined to have random behavior and they have no correlation with the other variables. Screened lithologies and the transmissivities were determined to be strongly correlating. On the other hand the weak correlation of CaMg and HCO3 ions with lithology and transmissivity, indicated that these ions are contributed from the aquifer formation.

- Principal component analysis determined that two new variables (principal components) could explain variations in the water samples as effectively as the entire data set. These principal components were formed from the combination of EC, NaK, Cl, % Na for the first one and CaMg, HCO3 and SO4 for the second. This simply groups the original variables as a combination affected by seawater and a combination of freshwater constituents. The analysis also yielded that lithology and transmissivity had no contribution to the first principal component, whereas transmissivity had the greatest contribution to the second principal component. Principal component analysis clearly categorized two major clusters of water samples depending on the type of lithology. In the first cluster water samples from the basalt formation are collected, and in the second water samples from conglomerate type lithologies are gathered.

- Application of the above methods proved that great complexity of the water samples of the Erzin aquifer could be clarified by using multivariate data interpretation techniques. This research defined and extracted hidden interrelations between the original variables, which had important role on the present geochemical state of the aquifer. It is mathematically found that there exists an interrelation between seawater and freshwater in the region, and this is an early indication that the basin is being contaminated by seawater.

REFERENCES

BROWN, C.E., 1993: Use of principal components, correlation and stepwise multiple regression analyses to investigate selected physical and hydraulic properties of carbonate-rock aquifers. -J. Hydrology, 147, pp. 169-195.

DAVIS, J.C., 1986 : Statistics and data analysis in Geology. - 2nd edition, John Wiley and Sons, 646 pp. NewYork.

DOYURAN, V., 1982: Geological and hydrogeological features of Erzin and Dörtyol plains. - Bull. Geological Society of Turkey, 25, pp.151-160.

EMEKLİ, N., KARAHANOGLU, N., YAZICIGİL, H., DOYURAN, V., 1996: Numerical simulation of saltwater intrusion in Erzin groundwater basin, Turkey. - Water Environment Research, (in press).

GEISS, S., EINAX, J., DAMZER, K., 1991: Multivariate analysis and its application in environmental analysis. - Analytica Chimica Acta, 242, 1, pp.5-9.

HELSEL, D.R., & HIRSCH, R.M., 1992: Statistical methods in water resources. - Elsevier, 522 pp. Amsterdam.

ISTOK, J.D., SMYTH, J.D., FLINT, A.L., 1993: Multivariate geostatistical analysis of groundwater contamination: A case history. - Ground Water, 31, 1, pp. 63-74.

KARAHANOGLU, N., DOYURAN, V., & SUVAGONDHA, F., 1986: Finite element model for the Erzin (Hatay) groundwater basin. - Bull. Geological Society of Turkey, 26, pp.53 -60.

KARAHANOGLU, N., YAZICIGİL, H., DOYURAN, V., EMEKLİ, N., HALLAJİ, K., 1995: Finite element simulation of saltwater-freshwater interface in coastal aquifers. - Project Final Report, YBAG-0074, The Scientific and Technical Research Council of Turkey (in Turkish). 288 pp. Ankara.

MELLOUL, A., 1995: Use of principal component analysis for studying deep aquifers with scarce data - Application to the Nubian sandstone aquifer, Egypt and Israel. - Hydrogeology Journal, 3, 2, pp. 19-39.

MELLOUL, A., & COLLIN, M., 1992: The principal component statistical method as a complementary approach to geochemical methods in water-quality factor identification-An application to the coastal plain aquifer of Israel. - J. Hydrology, 140, 1-4, pp. 49-73.

POHLMANN, H., 1993: Geostatistical modeling of environmental data. - Catena, 20, 1-2, pp.191-198.

ROCK, N.M.S., 1988: Numerical geology; in Lecture notes in Earth Sciences. - Battacharji, S., Friedman, G.M., Nuegebouer, H.J., and Seilacher, A., (eds), Springer Verlag,, 427 pp. Heidelberg.

SODERSTROM, M., 1992: Geostatistical modeling of salinity as a basis for irrigation management and crop selection- A case study in central Tunisia. - Environmental Geology and Water Sciences, 20, 2, pp.85-92.

TÜRKMEN, G., ERTÜRK, A., & TÜRKMAN, M., 1974: Dörtyol-Erzin plain hydrogeological investigation report. - Turkish Hydraulics Works, Ankara, Turkey (in Turkish), 76 pp. Ankara.

WEN, X.H., & KUNG, C.S., 1993: Stochastic simulation of solute transport in heterogeneous formations: A comparison of parametric and nonparametric geostatistical approaches. - Ground Water, 31, 6, pp. 953-965.

XHOFFER, C., BERNARD, P.,VANGRIEKEN, R., & VANDERAUWERA, L., 1991: Chemical characterization and source apportionment of individual aerosol particles over the North Sea and the English Channel using multivariate techniques. - Environmental Science and Technology, 25, 8, pp.1470-1478.

ZITKO, V., 1994: PC analysis in the evaluation of environmental data. - Marine Pollution Bulletin, 28, 12, pp.718-722.