J KAU: Earth Sci., vol. 6, pp 135-146 (1413 A.H. / 1993 A.D.)

Identification of Chemical Processes Using Principal Component Analysis

MAHMOUD S. ALYAMANI* and TIM C. ATKINSON**
* *Hydrogeology Department, Faculty of Earth Sciences, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia;*
** *School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom*

ABSTRACT. Principal component analysis (PCA) is useful technique for interpreting commonly collected groundwater quality data and relating the data to specific hydrochemical processes. Thus, multivariate analysis in hydrochemistry allows avoidance limitations that are associated with classical methods, such as Durov's and trilinear diagrams.

Six components which represent different chemical processes were identified and their relative areal impact determined. Comparisons made among the results presented and the findings of earlier studies high-light the descriptive capabilities of principal component analysis as an effective exploratory tool in hydrochemical investigations.

Introduction

Due to the complexity of the chemical evolution of groundwaters and the substantially large amount of basic information available, investigators are often unable to obtain a clear picture of the system under study. The basic behavioral model must be known as the necessary framework upon which more sophisticated interpretations or detailed explanations are to be built. It is in those first stages that multivariate analysis comes into play as a rather essential tool, particularly in days where high speed computers are available.

Principal component analysis (PCA) or factor analysis (FA), as stated by Harman (1976) "does give a simple interpretation for a given data and affords an elemental description of a certain set of variables analysed". In the hydrochemistry field, these techniques are widely used because they have several advantages over traditional graphical approaches by Dalton and Upchurch (1978), Dawdy and Feth (1967), and Lawrence and Upchurch (1982).

The PCA is a technique for grouping together different geochemical variables according to their degree of co-variation among samples. Therefore, PCA may be more useful in identifying, or confirming, geochemical processes. As described below there are numerous variables which together make up the groundwater quality. These variables may have greater or lesser importance in the overall groundwater chemistry. Some of them may be primarily controlled by one hydrochemical process, others by different processes. Therefore, some variables may show co-variation as a group and be distinct in this pattern from other groups. When examining the causes of the variation in the groundwater quality it is useful to reduce the number of variables not by eliminating redundant or insignificant ones, but rather by replacing the whole matrix of variables by a matrix made up of a more limited number of new variables which will stand in place of the original ones. These new variables can be found by using PCA.

The purpose of this study is to demonstrate the usefulness of PCA as a tool for recognition of chemical processes from commonly collected groundwater quality data which allows delineation of areas where groundwater is affected by the different processes that take place in the study area.

Study Area

The area of particular reference for this study is the country around Wadi Binhashbal. The location of Wadi Binhashbal, and a number of other towns and villages are shown in Fig. 1.

The water samples described in this work are obtained from existing wells that penetrate a shallow aquifer (alluvial deposits and weathering and fractured crystalline rocks). For several years the area has been subject to intensive pumping and poor management of water resources. In addition, to the above problems the structural setting of the area has lead to an increase in water salinity.

Procedures

The 114 groundwater samples described herein were collected from existing wells during a three week period in December 1986. Nearly all of the sampled wells were private domestic supply wells. Chemical data results obtained were used in the analyses. The variables used are Na^+, K^+, Mg^{2+}, Ca^{2+}, HCO_3^-, Cl^-, SO_4^{2-}, electrical conductivity (μmhos/cm), pH, dissolved oxygen (DO), partial pressure of carbon dioxide (P_{CO_2}), groundwater temperature (°C), and the saturation indices of calcite, dolomite, and gypsum. All chemical constituents mentioned above were in ppm except where noted. Data used in the PCA were in the unit specified above.

The wide range in the units and values among the variables would cause differences in one variable to have a larger effect than other variables of lower concentration if the absolute values are used. To overcome this problem it is necessary to remove the effects of scale differences in describing the variables, and so the variables are put in standard form using the following standardisation formula;

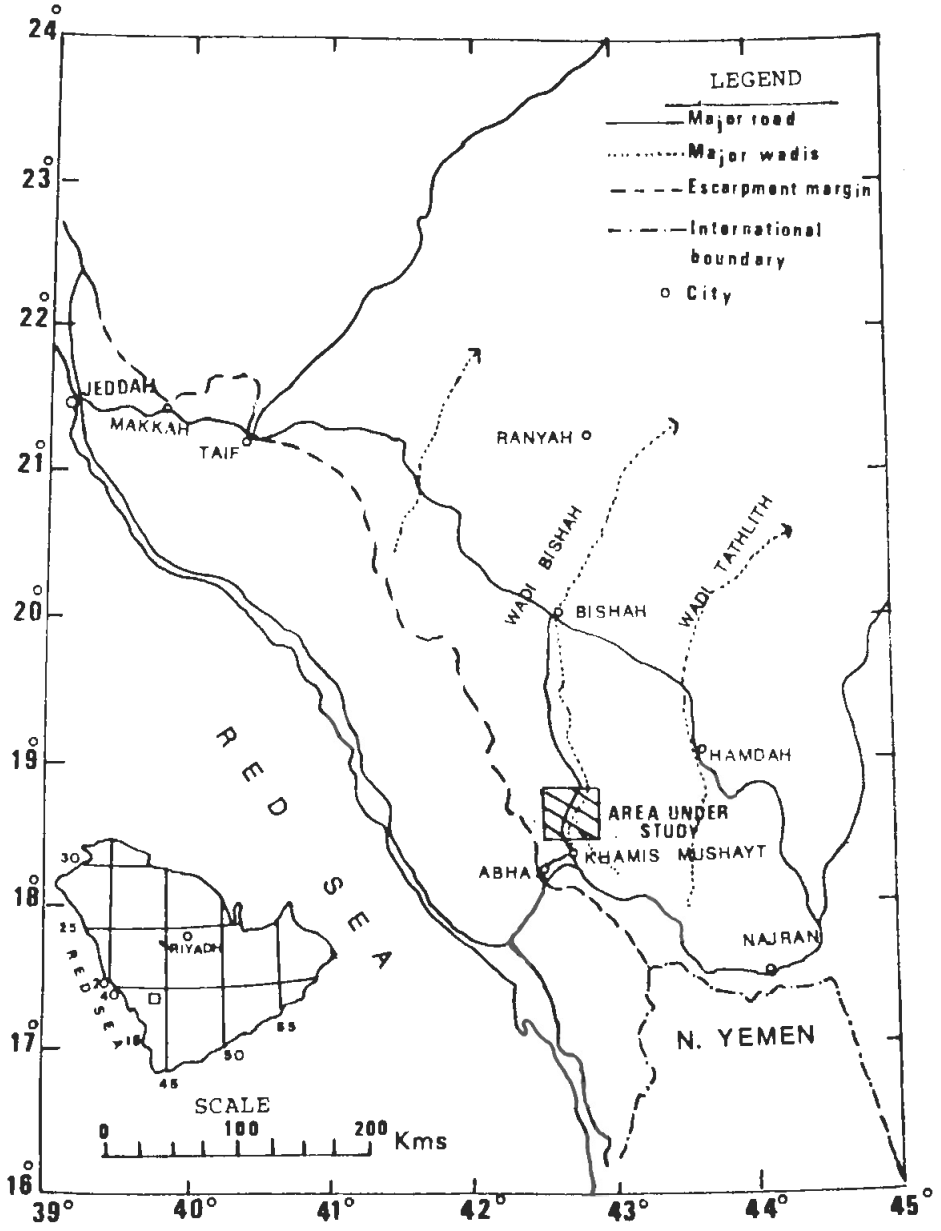

FIG. 1. Location map of study area.

$$Z = \frac{X - M}{s}$$

in which X is the observed variable with mean, M, and standard deviation, s; and finally Z is the standard variable without unit but with zero mean and unit standard deviation. It is important to notice that $(X–M)$ represents deviation form the mean value.

The standardised value then has equal weights for each variable and thus all standardised variables have the same metric.

As pointed out by Natusch and Hopke (1983), PCA used the matrix of correlation coefficients [between a set] of variables to identify some underlying pattern of relationships such that the data may be rearranged or reduced to a smaller set of components according to their observed interrelations. The varimax rotation was adopted in the present investigation. This method is one of the most commonly used techniques for orthogonal rotation and results in components for which the variable loadings are maximised. The variable loadings are the degrees of association between each variable and each component. The correlations between the variables and a component are known as the component loadings and the squares of their values indicate the proportion of the variance in the individual variable that can be encountered in the component (Johnston 1980). The sum of the squared variable loadings indicates the total variance accounted for by the component and is known as the eigenvalue of that component.

The component scores were calculated for each sample. They represent the values for the observations on the new variables, reflecting their values on the original variables and the contribution of each component (new variable) makes to the variance of these. The component scores can be obtained by the following formula;

$$S_{ik} = \Sigma\ D_{ij}\ L_{jk}$$

where D_{ij} is the standardised value for observation i on variable j;
L_{jk} is the loading of variable j on component K; and
S_{ik} is the score of observation i on component K; and summation is overall (n) variables.

The PCA was performed using the Statistical Package for Social Sciences Programs (SPSS-X).

Results and Discussion

The output results for the chemical data are given in Table 1. Four components had eigenvalue >1 while the other two components (V and VI) are of eigenvalue <1. However, the "ski-slope" diagram (Fig. 2) suggests that six components should be retained, where the experimenter selects components upto the point at which the diagram shows a break of slope.

TABLE 1. Variables and component loading after varimax rotation for the groundwater samples during recharge period, (Dec. 1986).

Variables	Component Loading					
	Comp. I	Comp. II	Comp. III	Comp. IV	Comp. V	Comp. VI
EC	0.945	0.142	0.040	0.077	0.046	0.060
Cl	0.945	0.142	0.040	– 0.077	0.046	0.060
Ca	0.914	0.122	0.124	– 0.038	0.088	– 0.100
SI (gypsum)	0.908	0.131	0.074	– 0.104	– 0.022	0.010
Mg	0.896	0.127	0.119	– 0.130	0.057	0.120
Na	0.895	0.103	0.017	0.040	– 0.028	0.330
SO_4	0.886	0.095	0.028	0.074	0.013	0.240
pH	– 0.057	0.968	– 0.172	0.090	– 0.060	0.070
P_{CO_2}	0.013	– 0.912	0.378	– 0.086	0.078	– 0.050
SI (dolomite)	0.421	0.879	0.170	– 0.074	0.062	0.020
SI (calcite)	0.438	0.847	0.169	– 0.065	0.032	– 0.050
HCO_3	0.123	– 0.147	0.976	– 0.038	0.008	0.050
Diss. Oxyg.	– 0.089	0.048	– 0.042	0.976	– 0.170	0.050
Temp.	0.023	– 0.040	0.011	– 0.165	0.914	– 0.010
K	0.583	0.085	0.086	0.101	0.033	0.790
Eigenvalues	6.59	3.39	1.23	1.06	0.89	0.85
PCT of VAR	43.9	22.6	8.2	7.1	5.9	5.7

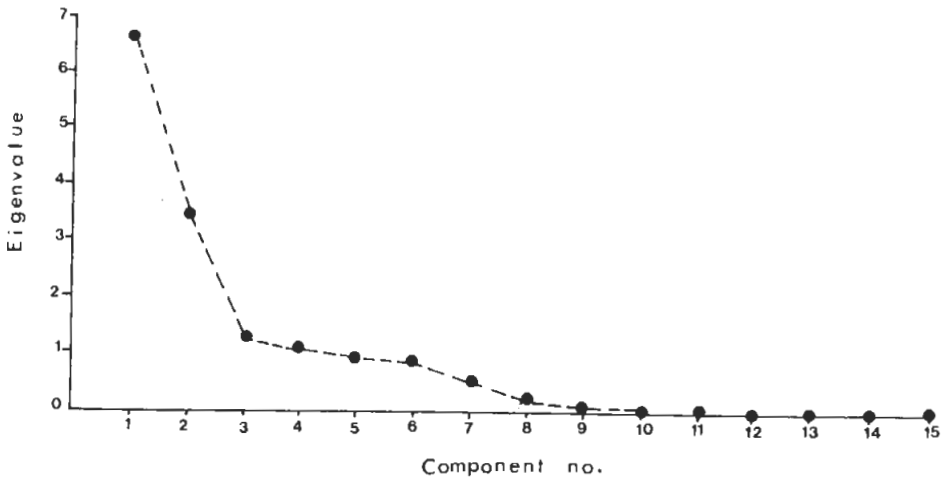

FIG. 2. Number of components plotted against the proportion of variance it extracted (eigenvalue).

The first four components account for 81.8% of the total variance while the last two components (V and VI) represent 11.6% of the total variance.

The percent total variance as shown in Table 1, represents the amount of the total variance in the system accounted for each component and the sum of these percentages provides the quantity of the total variance reflecting the components extracted in the system.

Inspection of the component loading (Table 1) shows that each of the first six components has a geochemically significant combination of variables. These components will be individually interpreted;

Component I accounts for 43.9% of the total variance. It has high loading for electrical conductivity, chloride, calcium, magnesium, sodium, and sulphate plus saturation index of gypsum. High values of these constituents most likely result from evaporation. All are positively correlated with component I weightings. Thus, component I represents the degree of overall mineralisation of the groundwater. This in turn is affected by mineral dissolution and evaporation, of which evaporation is believed to be the most important. On the other hand, low weightings on component I may also represent mixing of older concentration water with dilute recharge. Potassium has a secondary loading on this component, but is primarily associated with component VI.

Component II accounts for 22.6% of the total variance and has a high loading of pH, partial pressure of carbon dioxide (P_{CO_2}) and the saturation indices of calcite and dolomite. These variables both control and reflect the dissolution and precipitation of carbonate mineral. The sign values of the component loadings indicate that pH and P_{CO_2} values are inversely proportional, which is expected since when P_{CO_2} decrease the pH increases. Surprisingly, HCO_3 is not significantly loaded on this component.

Component III accounts for 8.2% of the total variance and has a high loading of bicarbonate (HCO_3) and moderate one of P_{CO_2}. It probably reflects the initial dissolution of minerals by groundwater. The low variance associated with this component may reflects buffering of the HCO_3 content by CO_2 exchange between groundwater and the atmosphere, and perhaps the precipitation of calcite and dolomite. Geochemically, both components II and III are associated with the CO_2-H_2CO_3-HCO_3-H_2O chemistry of the groundwater.

Components IV, V, VI represent single variables, dissolved oxygen, temperature and potassium respectively. The fact that potassium shows only moderate association with the "mineralisation" component I and high association with an orthogonal component of its own confirms that its concentration is largely independently controlled. Although K^+ must be released into the groundwater by weathering of K^- feldspar, it is removed, perhaps by ion exchange or uptake as a plant nutrient.

By examining the component scores, the spatial importance of the new variables and its distribution can be mapped. The spatial distribution of the first major component obtained for the chemical data was used to construct Fig. 3.

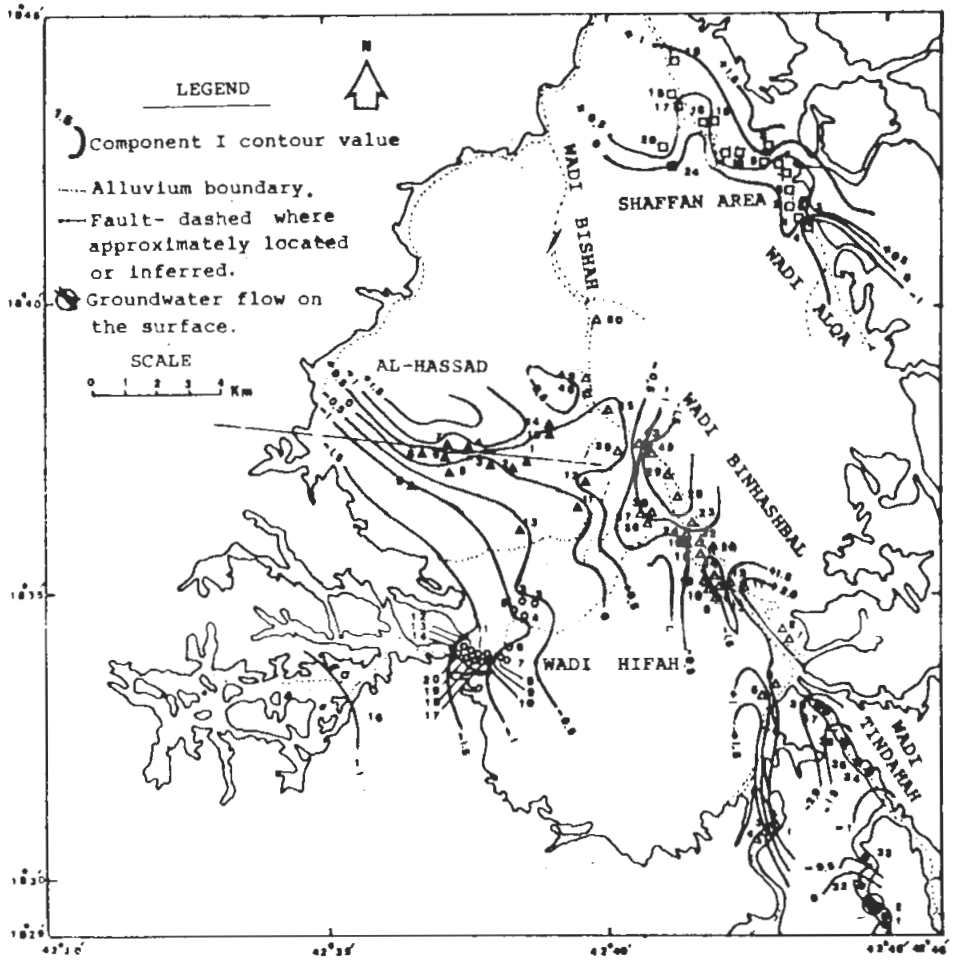

FIG. 3. Component I distribution in the area.

In term of the distribution of high scores (+ve), the distribution of the component matches that of electrical conductivity map (Fig. 4), confirming the distribution of low and high salinity waters outlined previously. However, the contour pattern is somewhat simpler than that of EC.

When compared with the water table map (Fig. 5) it shows that the degree of mineralisation reflected by component I increases generally in a down-gradient direction. Thus as water flows through the ground it becomes progressively more mineralised. Local anomalies in this process occur in areas where evaporation is artificially enhanced by abstraction followed by irrigation return. A natural anomaly, previously noted, occur at the E-W fault and dike in Al-Hassad area.

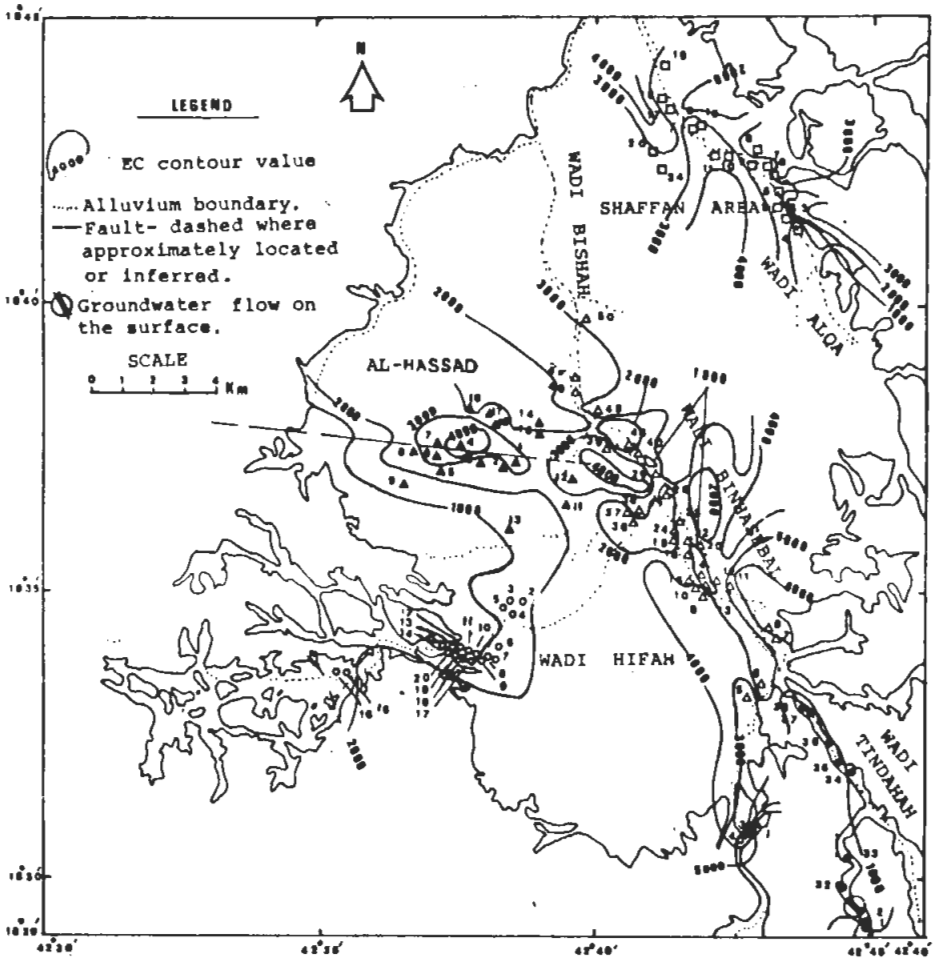

FIG. 4. Map showing electrical conductivity (EC) distribution in the study area.

A generalised conceptual model was constructed (Fig. 6) showing the locations of the main geochemical processes represented by the first four principal components obtained. Component I is interpreted mainly in terms of evaporation which occurs in two main locations during irrigation, and by direct evaporation from groundwater ponds and shallow water table. The third category of saline water, that trapped in weathered hollows and basins in the bedrock is harder to understand. It may represent groundwater affected by light recharge events during droughts, when evaporated salts may be flushed from soil and unsaturated zone. These more saline waters might be flushed out later by heavy, fresh recharge, surviving only in the lower parts of closed basins in the bedrock. Such basins may be weathered volumes of bedrock, as in Al-Hassad, or irregular rock-head filled by alluvium as shown in Fig. 6. Compo-

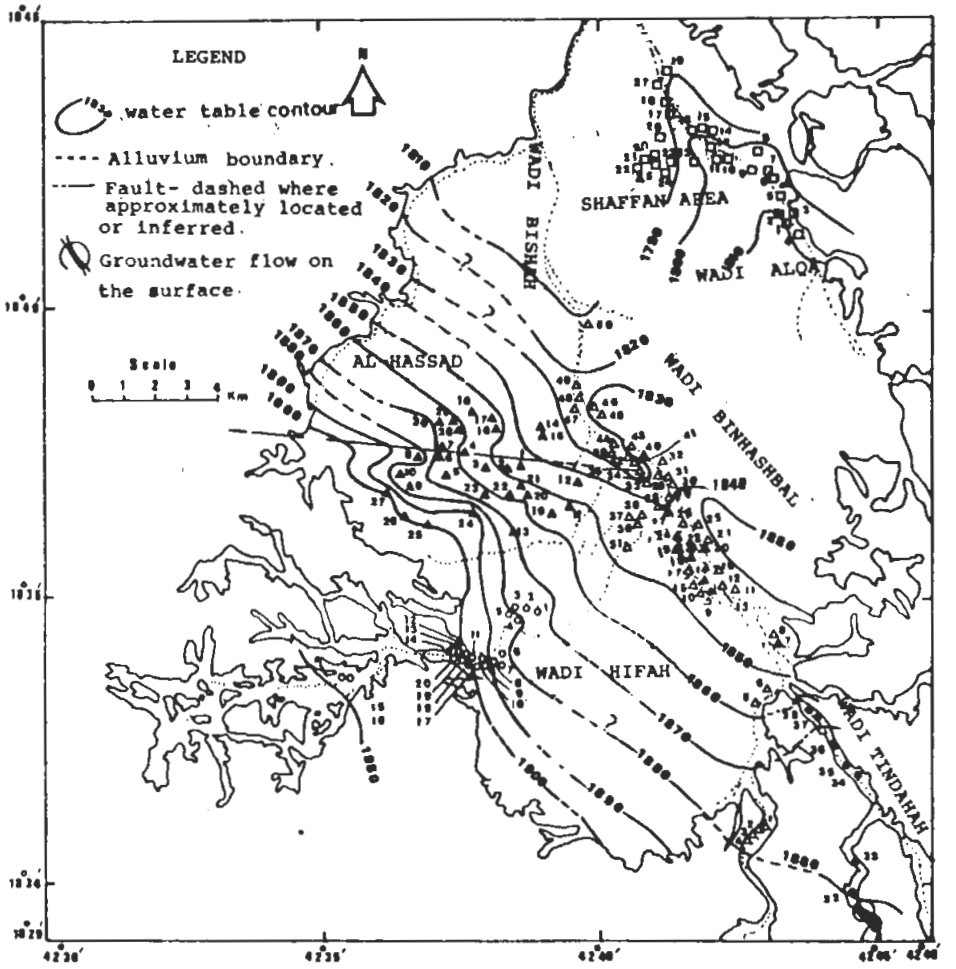

FIG. 5. Water table contour map.

nent II and III relate to weathering, dissolution and precipitation of carbonate minerals. These processes occur throughout the aquifer, although precipitation of calcite is probably only significant in zones of very shallow water table and evaporation. Because of the shallow depth to the water table, almost all waters are well oxygenated (component IV), implying a very restricted depth of groundwater circulation.

Conclusion

The results of PCA have rather clarified several aspects of groundwater condition from the hydrochemical point of view, and have led to the following conclusions that just three components (I, II, and III) can describe the processes that are taking place

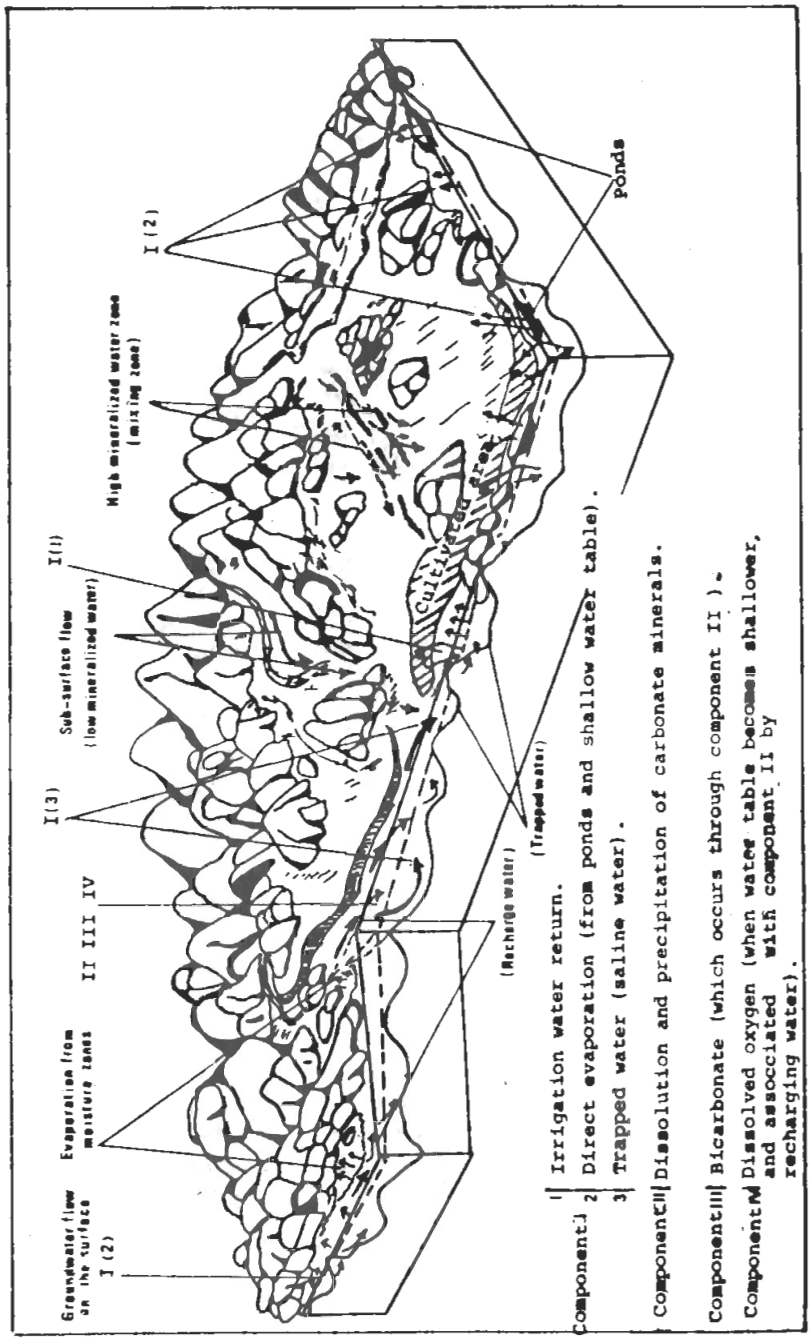

FIG. 6. Mineralised processes represented by PCA results.

in the hydrochemistry system. They represent two dominant groups of processes: degree of mineralisation, which is in turn controlled mainly by evaporation, and buffering of the bicarbonate system which controls weathering and secondary mineral precipitation.

A part from the anomalous behaviour of potassium, which may be due to its status as a nutrient, there is little evidence from the PCA to support ion exchange as an important process. Also the PCA technique delineates the area influenced by each chemical process when the component scores are mapped. Further, the example presented in this work indicates a reasonably good agreement between the conclusions drawn by the classical methods such as EC contour map and those obtained from the application of PCA.

Finally, it might be proposed that a priori realization of the hydrochemical and hydrological mechanism operating on the system are required to interpret the PCA results.

References

Dalton, M.G. and **Upchurch, S.B.** (1978) Interpretation of hydrochemical facies by factor analysis, *Groundwater*, No. **10:** 228-233.

Dawdy, D.R. and **Feth, J.H.** (1967) Applications of factor analysis in studies of chemistry and groundwater quality, Mojave River Valley, California, *Water Resour. Res.* **3**(2): pp. 505-510.

Harman, H.H. (1976) *Modern factor analysis*, 3rd. Ed. The University of Chicago Press, 487p.

Johnston, R.J. (1980) *Multivariate statistical analysis in geography*, Longman Inc., New York, 280p.

Lawrence, F.W. and **Upchurch, S.B.** (1982) Identification of recharge areas using geochemical factor analysis, *Groundwater*, **20**(6): 680-687.

Natusch, D.F.S. and **Hopke, P.K.** (1983) *Analytical aspects of environmental chemistry*, John Wiley and Sons Inc., New York.

التعرف على العمليات الكيميائيـــة باستخـــدام تحليل المركبات الأساسيـــة

محمود سعيد اليماني* و تم اتكنسون**

*قسم جيولوجيا المياه ، كلية علوم الأرض ، جامعة الملك عبد العزيز ، جـــدة ، المملكة العربية السعودية

**مدرسة العلوم البيئية ، جامعة إيست إنجليا ، نوريش ، المملكة المتحدة

المستخلص . يعتبر تحليل المركبات الأساسية من الوسائل المفيدة لتفسير المعلومات العامة المجمعة لنوعية المياه الجوفية وعلاقتها بالعمليات الكيميائية التي تحدث .

وكـذلـك التحليل ذو المتغيرات العديدة في دراسة كيمياء المـاء يسمح عادة باستبعاد المحدودية المصاحبة في تطبيق الطرق التقليدية مثال ذلك طريقة الثلاثي الخطي وطريقة ديوروف .

في الدراسة الحالية تم التعرف على ستة مركبات أساسية والتي تمثل عمليات كيميائية مختلفة كما تم تحديد الآثار النسبية لها في المنطقة ، وقد عملت مقارنة بين هذه النتائج ومع ما عمـل عن دراسـات مبكـرة بالطرق التقليدية وقد وجد أن الامكانيات الوصفية التي استخلصت بوساطة تحليل المركبات الأساسية يعتبر أداة استكشافية مؤثرة في التحقيقات التي تتم في كيمياء المـاء .