Journal of Hydrology, 39(1978)355—364

[3]

AN EXAMPLE OF THE USE OF FACTOR ANALYSIS AND CLUSTER ANALYSIS IN GROUNDWATER CHEMISTRY INTERPRETATION

R.P. ASHLEY and J.W. LLOYD

Department of Geological Sciences, University of Birmingham, Birmingham B15 2TT (Great Britain)

(Received January 12, 1978; revised and accepted April 14, 1978)

ABSTRACT

Ashley, R.P. and Lloyd, J.W., 1978. An example of the use of factor analysis and cluster analysis in groundwater chemistry interpretation. J. Hydrol., 39: 355—364.

Factor analysis and cluster analysis, as applied to two widely differing sets of groundwater hydrochemical data, appear to be moderately successful as statistical tools for revealing hydrochemical and hydrogeological features, including patterns of groundwater flow.

They possess advantages over the traditional graphical methods of solving similar problems, principally in their systematic nature, and they can generate inter-parameter relationships that may be overlooked in the less sophisticated traditional methods.

Their disadvantages are that they are easily prone to misuse and misinterpretation due to their complexity. Further, there is a need for the user to have adequate statistical knowledge.

INTRODUCTION

One of the major problems of hydrochemical investigations in hydrogeology is the ease with which large quantities of data are generated. In a comprehensive study, particularly a regional study, every water sample collected and analysed should be of some use but the shear number can frequently cause confusion and error both for the interpreter and for those to whom he presents his conclusions.

Several methods of data analysis have been devised to simplify interpretation and presentation. Histogram displays have long been used but trilinear diagrams (Piper, 1944; Durov, 1948) have become more commonplace because of their greater versability. Such diagrams are usually presented in conjunction with hydrochemical distribution maps. However, the trilinear methods have three main drawbacks: they ignore many parameters which are otherwise suitable for hydrochemical studies; they may be ambiguous in some respects due to the limitations of confining six variables to two dimensions; they use only relative percentages of different ionic concentrations, rather than absolute concentrations. Schoeller's (1956) method depicts ab-

solute values but is limited both in parameter types and in number of examples that can be displayed.

In view of the limitations of the existing methods and increasing number of chemical parameters now being measured in groundwater studies there is aneed for more wide ranging statistical analysis of data. In other scientific fields multivariate analysis is proving fruitful. From the methods available, factor and cluster analysis (with the aid of digital computers) have obvious attractions in groundwater chemistry in that large quantities of data can be processed rapidly and systematically. In the case of factor analysis the display of the major features of chemical variation and groundwater flow in a region may be possible while cluster analysis should allow the grouping of waters according to their chemical composition.

There has been very little previous work on applying factor analysis and cluster analysis specifically to hydrochemical data, although other statistical methods have been applied e.g. multiple regression by Khan et al. (1972) and discriminant function analysis by Drake and Harmon (1973).

It is not the purpose of this paper to explain the theory of cluster of factor analysis as these are well described elsewhere (Davies, 1973; Jöreskog et al., 1976). However, from the beginning attention is drawn to the considerable discussion, which may be deemed pertinent to groundwater chemistry, as to the proper place of factor analysis in hydrology and related disciplines (see Wallis, 1968). Some researchers advocate the use of factors themselves as meaningful entities (Matalas and Reiher, 1967) and use factor scores as measurements of a variable with a real meaning. Other researchers, however, advocate the use of factor analysis simply as a numerical method of discovering those variables which are more important than others for representing parameter variations (the "antifactor analysis" of Wallis, 1968). As a tool for discovering or demonstrating hydrochemical processes, factor analysis has been used by Dawdy and Feth (1967). Apparently, however, neither factor not cluster analysis has yet been used systematically to process large quantities of groundwater chemical data.

The purpose of the work described in this paper is to study by example the potential of factor and cluster analysis as tools for hydrochemical investigation and to examine their advantages and disadvantages compared to traditional methods. To carry out the investigation hydrochemical analyses from the Santiago alluvial basin in Chile (Moreno et al., 1970) and from the Carboniferous Limestone of the Derbyshire Dome in England (Edmunds, 1971) were used as raw data upon which the factor and cluster analyses could be tested. These two areas were selected as a large amount of raw data is available from each and they represent very different groundwater environments.

GROUNDWATER ENVIRONMENTS

Santiago alluvial basin

The Santiago Basin in Chile is an intermontane (1800 km^2) basin of Pleisto-

cene and Tertiary alluvial deposits, principally sandy gravels, sandy clays and silts, up to 450 m thick. It is underlain and surrounded by structurally disturbed "basement" rocks of early Tertiary and older epochs (Fig.1).

The principal recharge to the groundwater of the basin is from the influent Maipo, Mapocho and Colina rivers and from infiltration by direct precipitation, which ranges between 300 and 550 mm per annum.

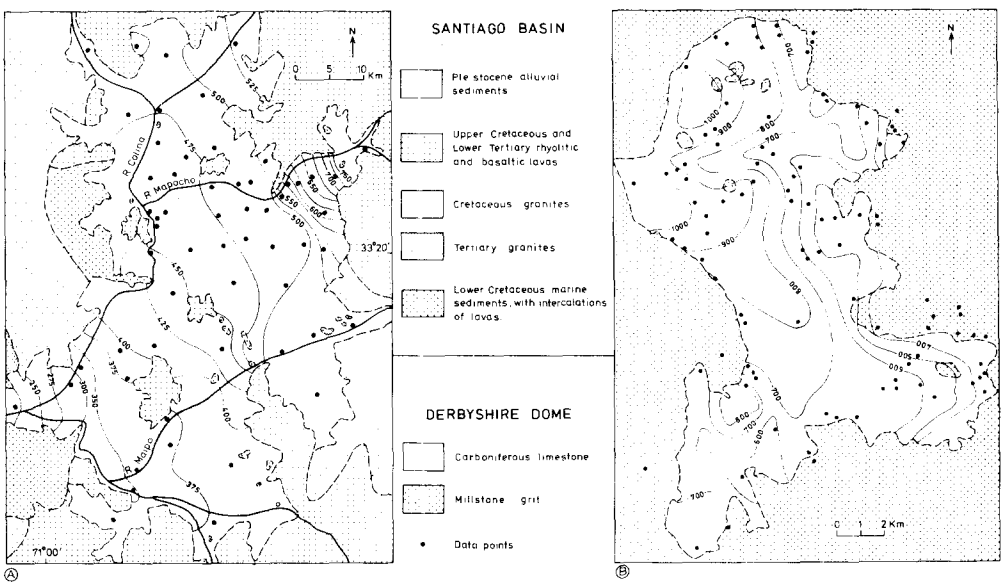


Fig.1. Geology and water table distribution in (A) the Santiago Basin and (B) the Derbyshire Dome.

Groundwater leaves the basin principally along the Maipo alluvial river channel in the southwest of the area, though abstraction of groundwater from wells is widespread. Cultivation with the aid of irrigation occurs over a wide area, and groundwater abstracted for this purpose may recharge the alluvium by infiltration

Hydrochemical analyses of a large numer of water samples taken from 266 wells scattered throughout the basin are available (Moreno et al., 1970) as well as analyses of river waters.

Samples have been analysed for the concentrations of the major ions (Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO_3^{-}, CO_3^{2-}, SO_4^{2-} and Cl^{-}) and also for pH and concentrations of silica, iron, NO_3^{-}, total dissolved solids, total hardness and non-carbonate hardness.

Derbyshire Dome

The Derbyshire Dome of England is a structure probably of Tertiary age, and approximately 600 km^2 of Carboniferous Limestone is exposed at its

centre. The surrounding rocks are predominantly of the Millstone Grit Series. The region has undergone extensive Permo-Triassic mineralisation and, as a result, is an area of economic interest (Fig.1B).

The information and data used in this paper were obtained principally from an investigation by Edmunds (1971) into the hydrochemistry of the groundwaters of this area, with particular emphasis on trace elements.

The principal recharge to the groundwater system is by infiltration from precipitation which ranges between 850 and 1250 mm per annum. Influent streams are also important. Groundwater discharges from the system by lateral flow into the Millstone Grit in the southeast and by effluent flow to springs and streams. Although the Derbyshire Dome consists of a karstic unconfined largely undeveloped groundwater system the hydrogeology is complicated by intrusive and extrusive igneous rocks within the limestones, which often give rise to perched water bodies. In addition, mining operations within the last couple of centuries for galena, sphalerite and copper ores, with their need for mine drainage, have altered the natural drainage pattern of the dome. Thermal springs are a feature of parts of the Derbyshire Dome, but chemical analyses of these were omitted from work described by this paper.

Water samples for chemical analysis were collected by Edmunds (1971) at about 100 localities, of which 19 were analysed in detail for a wide range of major and trace elements and 80 for a more restricted range of elements. The first 19 specimens were used by Edmunds (1971) to give an indication of those constituents which would be of sufficient interest to be studied in the larger sample of 80, rather than to indicate significant hydrochemical variations themselves.

The specimens were taken from spring and well waters, intended to be as representative as possible of the immediate provenance of the water: Millstone Grit, Carboniferous Limestone, perched water table on igneous rocks and mine drainage water.

METHODS OF DATA ANALYSIS

Factor analysis

The factor analysis used in this work was performed by a digital computer using a standard statistical program package. The data was analysed in *R*-mode and the results were output as:

(a) The loadings on all the variables (ionic concentrations, etc.) of the Varimax factors obtained from a principal component analysis of the raw data (those principal components with roots greater than 1 were selected as below this value the data provided nothing recognisable as worthwhile).

(b) The percentage of the total raw data variance accounted for by each Varimax factor.

(c) The scores (or "values") of each Varimax factor on each specimen.

Thus the values of each one of the Varimax factor scores could be treated

exactly as if they were measures of some hypothetical hydrochemical property of each specimen.

The factor analysis was performed three times: once using all the variables, once using anion concentrations only and once using cation concentrations only. Each time, the scores of the Varimax factors were plotted as a distribution map of the relevant area, starting with that factor which accounted for the highest percentage of the total data variance and hence might be expected to reveal the most significant hydrochemical variation.

Cluster analysis

The cluster analysis used in this work was performed in the same way as the factor analysis, by a digital computer using a standard statistical program package.

This time the data was analysed in *Q*-mode in order that similarities between specimens could be discovered rather than similarities between variables. The similarity coefficient used was the simple distance function and the clustering was performed by the Group Average (or Unweighted Pair-Group) method. To reduce the numbers of variables used in calculating the distance function, the calculation was performed on scores of Varimax factors extracted from the raw data.

The results were output in the form of a dendrogram showing the degree of similarity between specimens and groups of specimens. As with factor analysis, three cluster analyses were performed on each set of data: using all variables, using anion concentrations only and, finally, using cation concentrations only. In each case, specimens were plotted on maps of the relevant area, initially according to which of the first two major dendrogram groups the specimen belonged, but subsequently according to dendrogram subdivisions of these groups, continuing until no further pattern was discernible.

DISCUSSION OF ANALYSIS RESULTS OBTAINED

Factor analysis

The principal features of the groundwater flow in the Santiago Basin are readily displayed by the distribution map of total dissolved solids (Fig.2A). Closely matching this pattern is the distribution of the first factor extracted from the data using all the variables (Fig.2B). This factor has a high loading on total dissolved solids concentration (0.95) and so this relationship is only to be expected.

In a similar way the second factor extracted from all the variables and the first factors extracted from the anion and cation concentrations, considered separately, match the Na^+—K^+—Cl^-—SO_4^{2-} ion association. They do not reveal much additional detail of the groundwater flow, other than an indication of recharge to the alluvium (viz. a high ionic concentration and factor score)

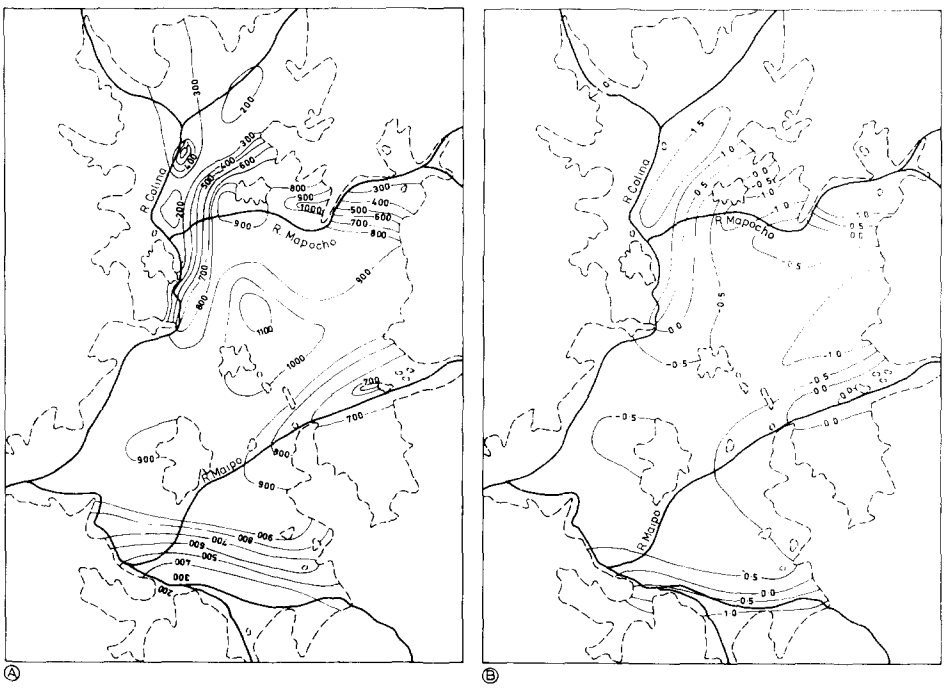


Fig.2. A. Distribution of total dissolved solids (mg/ml) in the Santiago Basin.
B. Distribution of first factor in factor analysis (high total dissolved solids weighting) in the Santiago Basin.

from the Mapocho River soon after it enters the basin (Fig.3A).

The third factor extracted from all the variables and the second factor from the anions only match CO_3^{2-} ion concentrations but, since this latter parameter has zero value almost everywhere, no conclusions may be drawn.

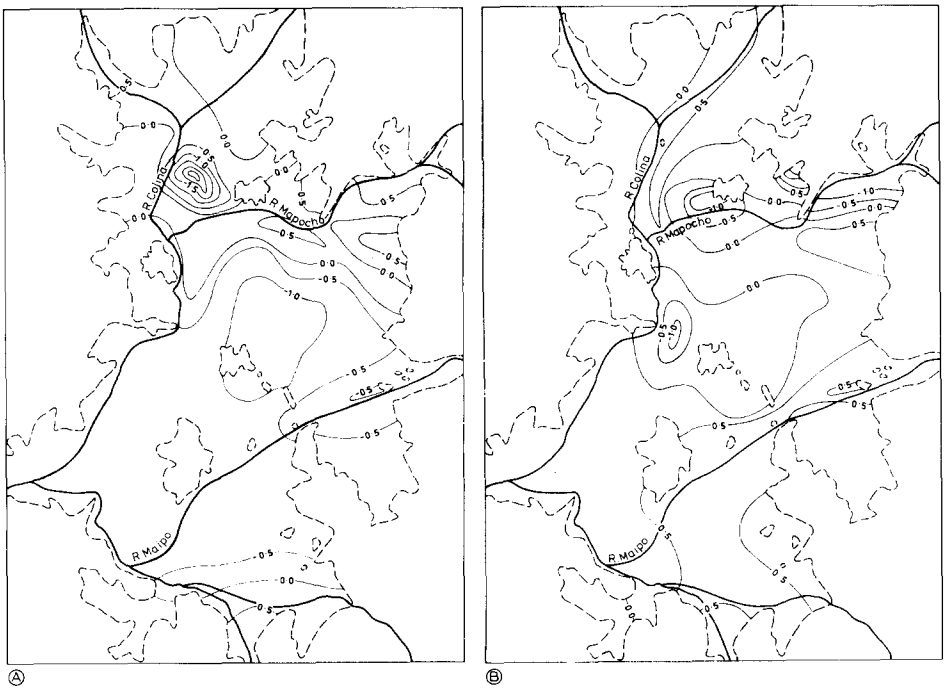
The distribution of the fourth factor extracted from all the variables (Fig.3B) and the second factor from the cation concentrations is more interesting. Both have a high loading on silica concentration, Extreme values of this factor, corresponding to high silica concentration, occur:

(a) Where the Mapocho River enters the basin after draining part of a Tertiary granite intrusion further east.

(b) Down gradient of exposed basement "islands" in the alluvium, i.e. where the groundwater is moving slowly.

(c) Where runoff to the basin presumably occurs from granitic intrusions along the northwestern and western margins of the basin.

The factor analysis process in the Santiago Basin case has therefore clearly correlated the first factor with the standard salinity distribution but has also highlighted the significance of one parameter (silica) which has provided some added information hitherto unrecognised about the flow distribution and style within the aquifer. In the process it has been necessary to examine



Ⓐ Ⓑ

Fig.3. Distribution of (A) second factor and (B) fourth factor.

only four parameters (factors) to obtain the information.

Factor analysis applied to data from the Derbyshire Dome is of special interest when concerned with the minor and trace constituents of the water. The major ions and factors accounting for the highest variance of the raw data showed the same correspondence as had been noted in the Santiago Basin results, though neither set of parameters revealed any clear flow patterns, possibly due to the tendency towards discrete flow paths in the limestone rather than flow in a continuous porous medium.

Edmunds (1971) selected Sr^{2+} (Fig.4A) and F^- ion concentration as being significant in revealing zones of mineralisation. The third factor extracted from all the variables (Fig.4B and the second extracted from anion concentrations (both having high loadings on Sr^{2+} and F^- ion concentrations) match the zones of mineralisation quite as well as if not better than the normal ion distributions.

Cluster analysis

Fig.5A and B shows the distribution of groundwaters in the Santiago Basin based on their major ion hydrochemistry according to a Durov diagram classification and a cluster analysis of all the variables, respectively. The figures identifying the groups of the cluster analysis are merely labels: each branch

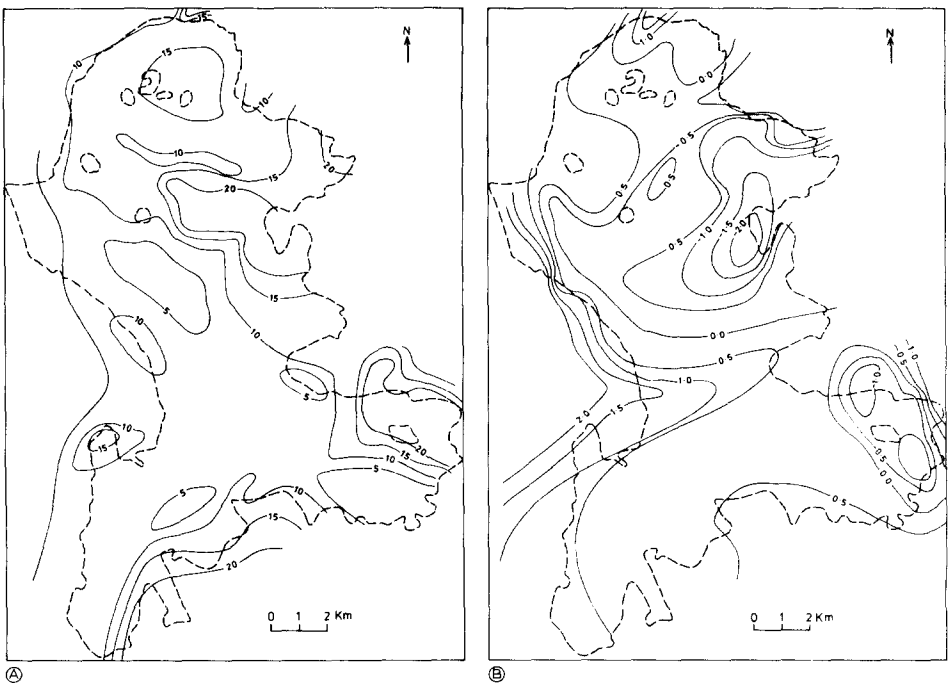


Fig.4. Distribution of (A) Sr^{2+} ion concentration (mg/l × 100) (after Edmunds) and (B) third factor in the Derbyshire Dome.

in the dendrogram being labelled *1* or *2*; subdivisions were labelled *11*, *12* and *21*, *22* etc., so that groups *21* and *22* are subgroups of the major group *2*. The cluster analysis shown here clearly reveals the relationship between groundwaters of different provenances more clearly than the Durov diagram. Cluster analyses based on anion or cation concentrations only revealed simplified versions of Fig.5B.

In the interpretation of the Derbyshire Dome waters neither the Durov diagram groupings nor the cluster analyses were successful in separating groups of the waters when using the major ions, though slight distinction does show between Millstone Grit and Carboniferous Limestone waters.

CONCLUSIONS

Factor analysis

Factor analysis as used here appears systematically to reveal basic hydrochemical features and/or flow patterns at least as successfully as traditional methods, which are usually more limited in the type of data they can use. In addition, factor analysis is successful in reducing large quantities of data to manageable form, and in indicating which of the many chemical parameters are of significance in a particular data set (e.g. silica for the Santiago Basin and F—Sr for the Derbyshire Dome).

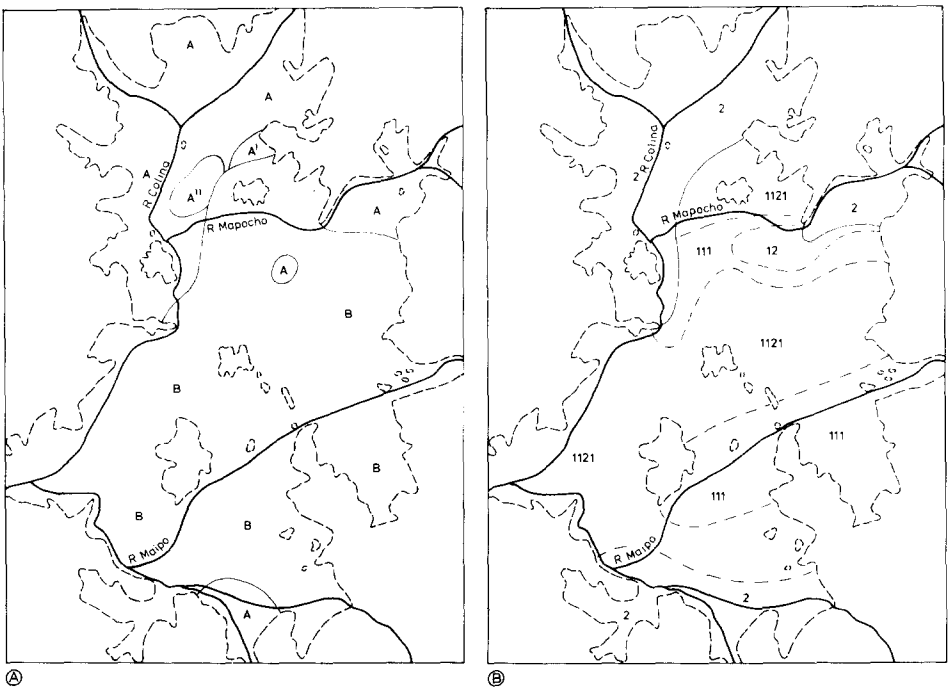


Fig. 5. Groupings of groundwaters in the Santiago Basin based on (A) a Durov classification and (B) a cluster analysis using all variables.

Factors have not been treated here as having any meaning themselves. If the meaning of a particular distribution of a factor is to be sought, then it is through the study of those variables (e.g. silica or Sr on which the factor has a high loading.

There seems to be no major advantage in subdividing the data into cation and anion concentrations.

Cluster analysis

The results of cluster analyses of the data from both the Santiago Basin and the Derbyshire Dome are inconclusive, though, in the case of the Santiago Basin as with factor analysis, they are apparently as successful as, if not more so, than traditional methods. As with factor analysis the advantages of cluster analysis are its systematic nature and its capacity for handling large quantities of data.

Cluster analysis appears to be more effective when all the major ion data is used than when cation and anion concentrations are separated.

ACKNOWLEDGEMENTS

The authors are grateful to the University of Birmingham Computer Centre

for the provision of the computer programs used in this work and for the advice on their use. The authors would also like to thank Mrs. Ashley for the preparation of most of the computer data cards used. The work was carried out under the auspicies of the Natural Environment Research Council.

REFERENCES

Davies, J.C., 1973. Statistics and Data Analysis in Geology. Wiley, New York, N.Y., 550 pp.

Dawdy, D.R. and Feth, J.H., 1967. Application of factor analysis in the study of chemistry of groundwater quality, Mojave River Valley, California. Water Resour. Res., 3(2): 505—510.

Drake, J.J. and Harmon, R.S., 1973. Hydrochemical environments of carbonate terrains. Water Resour. Res., 9(4): 949—957.

Durov, C.A., 1948. The geometrical method in hydrochemistry. Proc. Acad. Sci. U.S.S.R., 59(1): 87—96.

Edmunds, W.M., 1971. Hydrochemistry of ground waters in the Derbyshire Dome, with special reference to trace constituents. Inst. Geol. Sci. Rep. No.71/7, 62 pp.

Jöreskog, K.G., Klovan, J.E. and Reyment, R.A., 1976. Geological Factor Analysis. Elsevier, Amsterdam, 178 pp.

Khan, R.A., Ferrell, R.E. and Billings, G.K., 1972. The genesis of selected hydrochemical facies in Baton Rouge, Louisiana, groundwaters. Groundwater, 10(4): 14—20.

Matalas, N.C. and Reiher, B.J., 1967. Some comments on the use of factor analysis. Water Resour. Res., 3(1): 213—224.

Moreno, E.F., Urrutia, O.C. and Munoz, M.V., 1970. Hidrogeologia de la cuenca de Santiago. Inst. Invest. Geol., Chile, Publ. Especial No.3, 130 pp.

Piper, A.M., 1944. A graphic procedure in the geochemical interpretation of water analysis. Trans. Am. Geophys. Union, 25: 914—923.

Wallis, J.R., 1968. Factor analysis in hydrology — an agnostic view. Water Resour. Res., 4(3): 521—527.

Schoeller, H., 1956. Géochimie des eaux souterraines. Technip, Paris, 132 pp.