

# APLICACION DEL MICROORDENADOR A ESTUDIOS DE CALIDAD DE AGUAS

INTERPRETACION REAL DE ALGUNOS TRATAMIENTOS

ANTONIO CASTILLO MARTIN

Dpto. de Hidrogeología y de Investigaciones Geológicas del CSIC. Facultad de Ciencias. Universidad de Granada.

Obtenida una tabla de datos físico-químico-biológicos de aguas (de M variables x N observaciones), queda siempre el problema, tanto mayor cuanto más numerosa es la serie de datos a extraer de la misma el máximo de información posible. Para ello, cada vez más, se impone la utilización del microordenador con técnicas y tratamientos matemáticos multivariados.

En esta línea, se presentan tres tipos de análisis matemáticos (Correlación y Regresión, Cluster y Factorial) de especial interés y aplicabilidad a estudios de analítica de aguas en general. Tras señalar muy brevemente los fundamentos de los mismos, se exponen, comentan e interpretan algunas situaciones reales a partir de los correspondientes gráficos de salida dibujados por el microordenador.

La información y el apoyo suministrado por estas "herramientas" matemáticas es del máximo interés en aquellos trabajos o actividades donde se generen importantes series de datos analíticos (estaciones depuradoras, control de tratamientos en fábrica...), a cuyos profesionales va dirigido especialmente este artículo.

## ANALISIS DE CORRELACION Y REGRESION

### FUNDAMENTO

La utilidad principal de este tipo de análisis es la de poner de manifiesto

to, gráfica y analíticamente, relaciones entre variables y/o parámetros. El análisis de regresión considera únicamente la distribución de frecuencias de una variable (dependiente), suponiendo conocida a la otra (variable independiente). El análisis de correlación, por el contrario, considera la variación con-

junto de las dos variables seleccionadas. Así pues, dicho análisis determina el grado de asociación de las variables a través del conocido coeficiente de correlación, mientras que el análisis de regresión lo que obtiene es una ecuación que permite predecir valores de la variable dependiente a partir de los aportados por la variable independiente.

La ecuación de la recta de regresión que debe emplearse a efectos inductivos ha de calcular la pendiente y la ordenada en el origen a partir de la media, varianza y covarianza de la población; si ello se hace así, es sumamente recomendable introducir también el cálculo del denominado error estándar de estimación, a fin de conocer el margen de tolerancia, junto con el grado de fiabilidad, de que se dispone en los correspondientes cálculos inductivos de la variable dependiente.

### EJEMPLO DE APLICACION. INTERPRETACION

En la figura 1 se presenta una salida

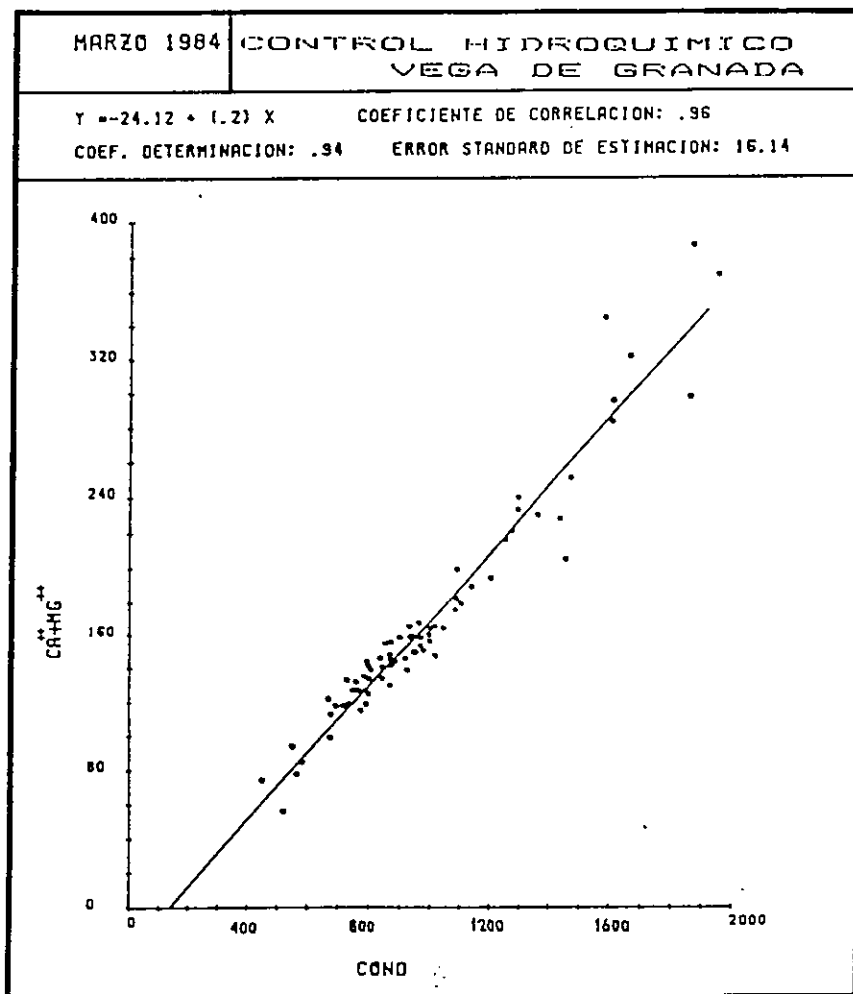
por *plotter* de un análisis de correlación y regresión de las variables conductividad-iones de calcio + magnesio de un caso real. Dicha representación fue obtenida por un *plotter* Watanabe adaptado a un microordenador Apple II Europlus, a través del programa correspondiente; el programa, denominado Correlación vertical, fue elaborado por S. Fernández Lorca (ETSI de Minas, Madrid). Como puede observarse en la salida de la figura 1, dicho programa, aparte de representar gráficamente la nube de puntos y de ajustar la recta teórica, calcula la ecuación de la recta de regresión, el coeficiente de correlación, el coeficiente de determinación y el error estándar de estimación.

La representación que se comenta formó parte de un conjunto de 28 gráficos de correlación y regresión (1 y 2), realizados para dos campañas de análisis de aguas subterráneas de un mismo sistema acuífero. A partir de las correspondientes matrices de correlación entre variables, obtenidas en los procesos previos de cálculo de los programas de análisis Cluster y Factorial (que se comentan más adelante), se elaboró un dendrograma de alta correlación entre variables, una de cuyas ramas es la que equivale a la representación expuesta en la figura 1. No es oportuno comentar aquí las consideraciones que fueron extraídas de este amplio tratamiento de correlación y regresión, aunque sí creemos interesante apuntar las consideraciones que se deducen de la representación de la figura 1, siguiendo un camino deductivo que pueda utilizarse para cualquiera otras representaciones análogas.

La exclusiva observación de la representación gráfica aludida pone de manifiesto que existe una excelente correlación entre las dos variables consideradas; ello se apoya matemáticamente en un coeficiente de correlación de 0,96, así como en un coeficiente de determinación de 0,94 (que indica que el 94 por 100 de la varianza muestral se halla explicada por el ajuste teórico obtenido); finalmente, y en consonancia con los anteriores valores, el error estándar de estimación hallado, de sólo 16,14 mg/l, confirma una predicción de la variable Y (variable dependiente), al 95 por 100 de certeza, aplicando un intervalo de con-

Fig. 1. Análisis de Correlación y Regresión para la relación conductividad-iones de calcio y magnesio (programa CORRELACION VERTICAL, microordenador Apple II Europlus).

\*\*\* ANALISIS DE CORRELACION Y REGRESION \*\*\*



fianza sobre el valor teórico calculado de  $\pm 32,28$  mg/l.

Información complementaria puede obtenerse de la distribución espacial de la nube de puntos, que pone de relieve un segmento de alta densidad puntual acotado por un rango de conductividad de 700 a 1.000  $\mu$  mhos/cm y por un intervalo correspondiente de contenido en iones de calcio + magnesio de 110 a 170 mg/l. El hecho de que la recta de regresión ajustada no intersekte el origen de coordenadas, mostrando una conductividad residual (en ausencia de iones de calcio y magnesio) de tan sólo 120  $\mu$  mhos/cm, indica que aunque la dureza aglutina al peso mayoritario de la salinidad total, existe una cierta influencia de aportes, fundamentalmente de tipo antrópico, que no tienen correlación geoquímica con los de los sulfatos y bicarbonatos (aniones

complementarios de los cationes de calcio y magnesio).

## ANALISIS CLUSTER

### FUNDAMENTO

El análisis Cluster (también denominado de enjambres, racimos...) constituye otra herramienta matemática, de fácil manejo, de suma utilidad para la clasificación tanto de elementos (análisis cualitativo) como de variables (análisis cuantitativo). En estudios de analítica de aguas (3,4...) permite clasificar elementos o variables de cualquier conjunto de determinaciones (Cluster de variables) o de observaciones (Cluster de casos) atendiendo a diversos criterios de distancia (distancias euclídeas, coeficientes de correlación...).

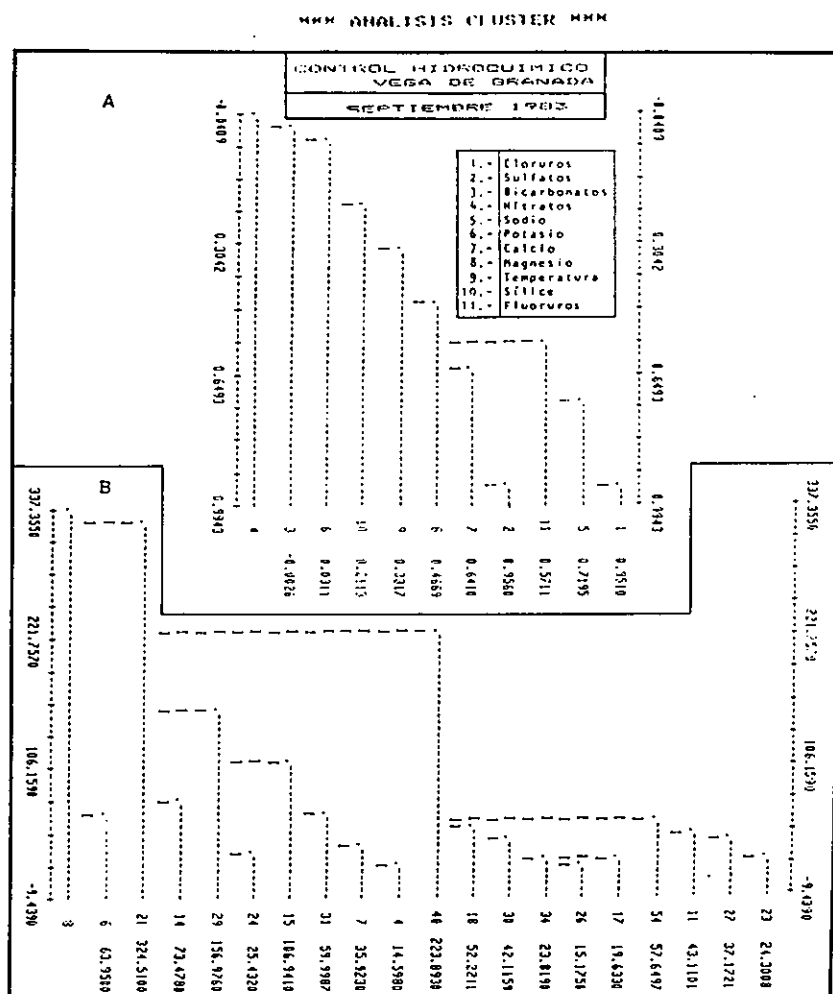
En definitiva, considerada una matriz de datos analíticos  $X = (x_{ij})$ , donde  $x_{ij}$  es el valor de la variable  $x_j$  en el  $i$ -ésimo individuo, y cada una de las filas y columnas formadas por  $M$  y  $N$  observaciones y variables, respectivamente, el Cluster de casos, o de variables, lo que hace es determinar  $M$  o  $N$  subconjuntos, según corresponda, de tal forma que cada individuo o variable pertenezca a una y sólo una de estas clases. Todo ello se realiza en base a una medida de distancia (según el criterio elegido) entre los elementos que representan a cada una de las observaciones o variables correspondientes.

### EJEMPLO DE APLICACION. INTERPRETACION

En la figura 2 se presentan dos salidas por impresora, correspondientes a un análisis Cluster de variables según coeficientes de correlación (A) y a un análisis Cluster de casos según distancias euclídeas (B). Dichos dendrogramas, que es como se denominan genéricamente estas representaciones, fueron obtenidos por una impresora Data General IPS-5.000 conectada a un microordenador Data General MPT-100, a partir del programa correspondiente; el programa, denominado Cluster, fue adaptado por A. Padilla de Davis (5). Una vez introducida la matriz MN de datos originales ( $M$  variables x  $N$  observaciones), este programa ofrece la posibilidad de elaborar el Cluster de las variables analíticas introducidas o el Cluster de las observaciones; asimismo, está concebido para trabajar con distancias euclídeas o con coeficientes de correlación como criterios de distancia (6 y 7).

El dendrograma superior (A) muestra la relación existente, en un caso real, entre un conjunto de 11 variables físico-químicas. La representación subraya la alta "familiaridad" de las variables cloruros (n° 1)-sodio (n° 5) y sulfatos (n° 2)-calcio (n° 7), familias que se unen a mayor distancia con los fluoruros (n° 11) y con el magnesio (n° 8), y, más lejos aún, entre sí (a un coeficiente de correlación de 0,57). Según se ha descrito, y puede observarse en la figura 2, ambas familias iniciales componen conjuntamente

Fig. 2. Cluster de variables (A) y de casos (B), según coeficientes de correlación y distancias euclídeas, respectivamente (programa CLUSTER, microordenador Data General MPT-100).



un bloque de correlación del que quedan desconectadas el resto de las variables físico-químicas consideradas en el tratamiento, las cuales, a su vez, no aparecen significativamente correlacionadas entre sí. En este sentido, merece destacarse la "afamiliaridad", o carácter sectorio, de los nitratos (n° 4), aportados antrópicamente, de los bicarbonatos (n° 3), sujetos a modificaciones por cambios en la presión parcial de  $CO_2$ , y del potasio (n° 6), aportado antrópicamente y fuertemente retenido por el suelo.

El dendrograma inferior (B) muestra la relación existente entre las observaciones (análisis), utilizando en este caso, como criterio de familiaridad, las distancias euclídeas. Como puede observarse, se diferencian claramente tres agrupaciones o familias de análisis. La primera esta-

ría compuesta por los números 23 al 40, la segunda por los números 4 al 21, y la última por los números 6 y 8. El dendrograma complementario, realizado con los coeficientes de correlación, puede utilizarse para definir subfamilias de puntos de alta correlación y baja distancia, de gran interés en estudios geoquímicos (en el más amplio de los sentidos) e inductivos.

## ANALISIS FACTORIAL

### FUNDAMENTO

El análisis Factorial es un tratamiento matemático multivariable muy utilizado con diversas finalidades en distintos campos de la ciencia. En estudios hidroquímicos se dispone ya de una relativa experiencia de su aplicación (8, 9, 10...).

Se puede decir, en pocas palabras, que es una técnica que permite reducir las M variables consideradas, de un conjunto de N observaciones, a un número reducido de nuevas variables, las cuales soportan el mayor peso de la varianza muestral. A estas nuevas variables es a las que se denomina Factores o Componentes. El tratamiento ulterior permite hacer representaciones, tanto de variables originales como de observaciones, en base a los Factores o Componentes deseados, lo que normalmente suministra información adicional no alcanzable con técnicas descriptivas.

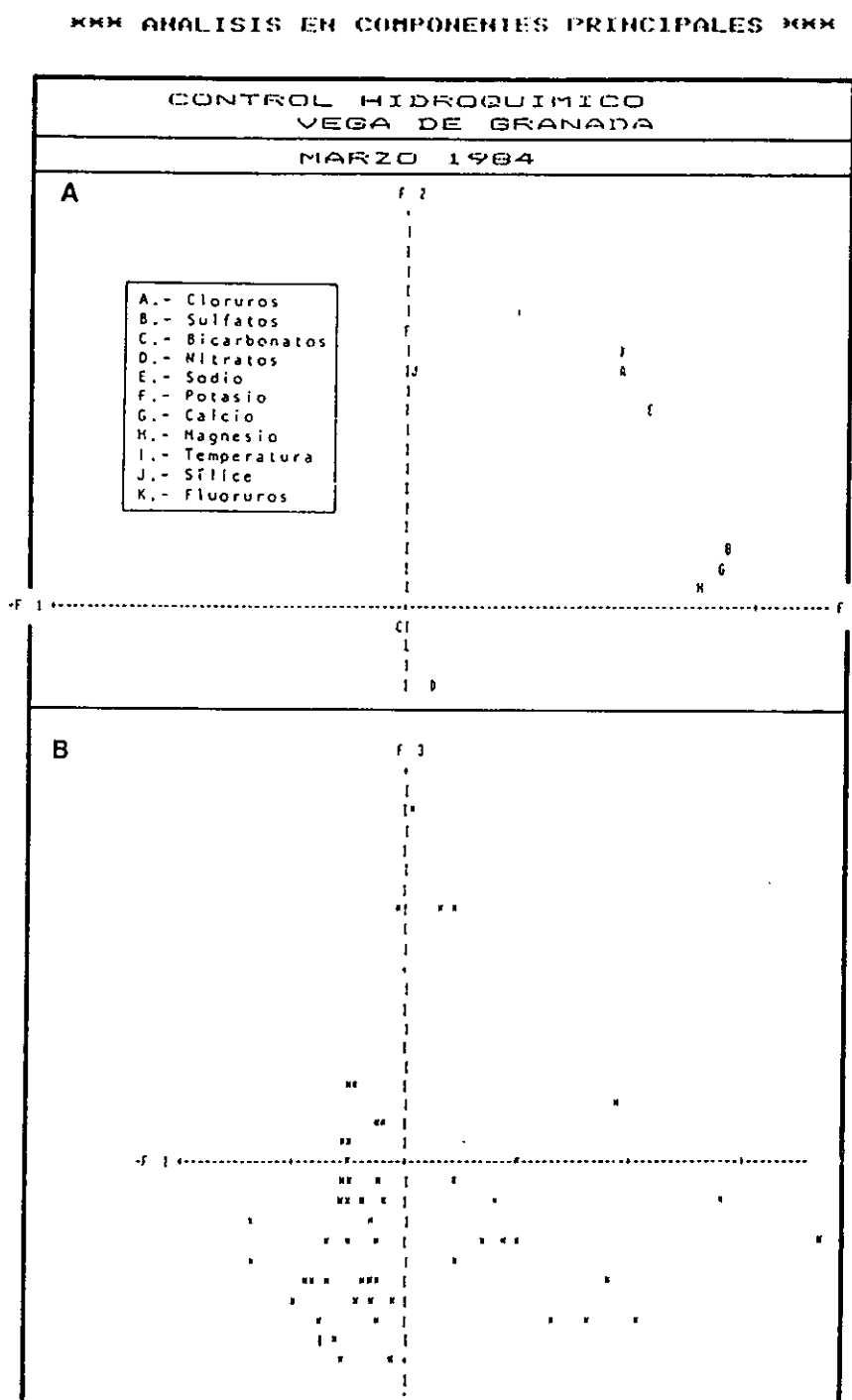
### EJEMPLO DE APLICACION. INTERPRETACION

En la figura 3 se presentan dos salidas por impresora, correspondientes a un mapa de posición de variables hidroquímicas realizado con los factores 1-2 (A), y a un mapa de posición de observaciones para los factores 1-3 (B). Dichos mapas de posición, como se denominan usualmente estas representaciones, fueron obtenidos por una impresora Data General IPS-5.000 conectada a un microordenador Data General MPT-100, a partir del programa correspondiente; el programa, denominado FACTOR, fue, asimismo, adaptado por A. Padilla de Davis (5) (6 y 7).

En el mapa de posición superior de la figura 3 (A) puede observarse, para un conjunto de análisis de aguas, cómo los iones sulfato (B), calcio (G) y magnesio (H) forman un mismo agrupamiento, el cual explica la mayor parte de la varianza acumulada en el factor 1. Otra agrupación es la integrada por los iones fluoruro (K), cloruro (A) y sodio (E), cuyo peso en la varianza se distribuye entre los factores 1 y 2. Quedan descolgados de una vinculación apreciable al peso de varianza de los factores 1 y 2, y, por tanto, de los procesos geoquímicos mayoritarios, la temperatura (I), los bicarbonatos (C) y los nitratos (D).

El mapa de posición inferior (B) realiza, por el contrario, una discriminación selectiva del conjunto de análisis obtenidos, en este caso, a partir de los factores 1-3. Como puede observarse, existe una alta densidad de agrupamiento en el

Fig. 3. Mapas de posición de variables (A) y de observaciones (B) para un tratamiento Factorial (programa FACTOR, microordenador Data General MPT-100).



cuadrante inferior izquierdo, cuyos elementos responden a semejantes procesos geoquímicos. La información más interesante es, lógicamente, la aportada por los elementos desagrupados; así, la discriminación producida en sentido vertical pone de manifiesto puntos de agua con estructuras composicionales modifi-

cadas por aportes fundamentalmente de origen antrópico (contaminación); la discriminación obtenida en sentido horizontal resalta, por el contrario, a los puntos de agua fuertemente influenciados por aportes salinos de naturaleza evaporítica (2).

## AGRADECIMIENTOS

Este trabajo no hubiera podido realizarse sin la gentileza y el apoyo de D. Alberto Padilla y D. Sergio Fernández Lorca, autores de los programas de microordenador comentados, los cuales fueron aplicados a estudios hidroquímicos llevados a cabo por el autor.

## BIBLIOGRAFIA

- (1) Fernández Lorca, S. y Castillo Martín, A. Estimaciones de parámetros físico-químicos en aguas por correlación simple. Aplicación al sistema hidrogeológico de La Vega de Granada. I Cong. de Geoquímica. 61-62. Soria (1985).
- (2) Castillo Martín, A. Estudio hidroquímico del acuífero de La Vega de Granada. Ed. IGME y Univ. de Granada. 600 pág. Granada (1986).
- (3) Ibañez, V. y Sanchís, E. Utilización de modelos estadísticos multi-

variables en hidrogeoquímica. III Simp. de Hidrogeología. IX: 263-275. Madrid (1983).

- (4) Moral, S.; Pulido Bosch, A. y Valenzuela, P. Aplicación de los análisis Cluster al estudio de características físico-químicas de aguas subterráneas. Est. Geol. (1984).
- (5) Davis, J.C. Statistics and Data Analysis in Geology. John Wiley and Sons. 550 pág. (1973).

- (6) Padilla, A. Aplicación del microordenador a la hidrogeología. Tesis Licenc. Univ. Granada. 81 pág. Granada (1984).

- (7) Padilla, A.; Pulido Bosch, A. y Benavente, J. Programas de elaboración y tratamiento de datos físico-químicos de interés en hidrogeología. II Simp. sobre el Agua en Andalucía. I: 437-446. Granada (1986).

- (8) Mudry, J.; Kiraly, L. y Muller, I. Analyse multivariée du chimisme de quelques sources karstiques du Jura Suisse et Franc-Comtois. Bull. Cent. d'Hydrog. 3: 183-221 (1979).

- (9) Bakalowicz, M. Contribution de la géochimie des eaux à la connaissance de l'aquifère karstique et de la karstification. Thèse Doct. Univ. Paris. 269 pág. (1979).

- (10) Benavente, J.; Da Costa Almeida, C.; Castillo Martín, A. y Oliveira, M. Algunos ejemplos de aplicación del análisis Factorial al tratamiento de datos hidroquímicos. III Simp. de Hidrogeología. IX: 297-310. Madrid (1983).