

TECNICAS DE OBTENCION DE COMPONENTES PRINCIPALES Y SU APLICACION A LOS PARAMETROS DE CALIDAD DE AGUAS

Por RAFAEL BIRLANGA

Ingeniero de C., C. y P.

RAFAEL HERAS

Prof. Ingeniero de C., C. y P.

Es frecuente, en el campo de los recursos hidráulicos, el tener que evaluar un cierto fenómeno que depende de un gran número de parámetros; como consecuencia de ello es difícil ver con claridad si todos los parámetros son igualmente representativos del fenómeno o, por el contrario, se puede obtener una representación formal de éste eligiendo de entre los parámetros de origen algunos de ellos, o combinaciones de los mismos, de forma que se alcancen niveles de confianza satisfactorios.

El primer problema es, por tanto, y partiendo de técnicas matemáticas de carácter general, obtener un sistema de información suficientemente consistente para su aplicación a los problemas que se presentan en Hidrología cuando se trata de seleccionar los parámetros más representativos del fenómeno, así como el obtener la información más precisa acerca de la importancia relativa de los mismos.

Es de destacar en el análisis de la estructura de los datos, que parámetros que tradicionalmente se han considerado como independientes en sentido estadístico, no lo son, sino que por el contrario, se observan fuertes dependencias de unos con los otros. Por ello, un parámetro no puede aislarse repentinamente del resto sin analizarlo convenientemente, pues la información que aporta puede pesar lo suficiente como para desvirtuar el fenómeno real.

La técnica de obtención de componentes principales puede utilizarse con buenos resultados en la hipótesis de dependencia lineal entre los distintos parámetros representativos del fenómeno, para deducir unas transformaciones que conviertan los parámetros de origen en otros cuyas jerarquías sean evaluables, lo que permitirá reducir el número de parámetros de que inicialmente se dispone, obteniéndose la consiguiente economía en la modelación del fenómeno.

a) Fundamento del método.

En orden a lo indicado en los párrafos iniciales, pasamos a describir el fundamento del método. En esencia, éste consiste en, sobre la hipótesis de dependencia lineal entre los distintos parámetros representativos del fenómeno, deducir unas transformaciones que conviertan los parámetros de origen en otros cuyas jerarquías sean evaluables, lo que permitirá reducir el número de parámetros de que inicialmente se dispone, obteniéndose la consiguiente economía en la modelación del fenómeno.

Si a causa de la no linealidad de los datos originales no resulta practicable el efectuar transformaciones del tipo mencionado, siempre podrá soslayarse el problema efectuando una transformación previa de tipo logarítmico, para evitar los inconvenientes que en el método aparecían como consecuencia de la susodicha falta de linealidad en los datos de partida.

El método, consistente en efectuar unas transforma-

ciones lineales sobre los parámetros primitivos que conviertan éstos en otro conjunto de nuevos parámetros, se conoce como análisis de los componentes principales representativos del fenómeno.

Ahora bien, de las prácticamente infinitas transformaciones lineales que nos es dado escribir, elegiremos aquellas que se ajustan a los dos criterios siguientes:

El primer componente principal se elige de forma que tenga máxima varianza, eligiéndose el segundo de forma que siendo incorrelado con el primero, tenga la mayor varianza posible; este proceso se repite en la extracción de los siguientes componentes principales.

En términos matemáticos, se trata de efectuar la transformación a ejes principales de una forma cuadrática relativa al espacio multidimensional que constituye la masa muestral de los parámetros observados. Es decir, a partir de un sistema complicado de relaciones obtener una forma canónica formada por un conjunto ortogonal de ejes. De esta forma los componentes principales pueden

obtenerse como las líneas de las sucesivas tendencias que adoptan los individuos de la población, en el espacio muestral.

Es obvio que la forma más sencilla de representar los individuos de la muestra es referirlos a una base ortonormal del espacio muestral multidimensional.

El procedimiento seguido en la obtención de los componentes principales es un método puramente matemático que proporciona unas relaciones entre los datos de partida; corresponde al análisis estadístico evaluar qué dimensión es más significativa, real o importante, así como evaluar aquellas que pueden desprejarse. No obstante, la ventaja que proporciona el que se utilicen componentes incorrelados, es el poder analizar éstos con más comodidad que los parámetros originales.

b) Rectas de regresión mínimo cuadrática.

Consideremos que el conjunto de parámetros está constituido por p variables aleatorias x_i de las que disponemos muestras iguales de extensión n , es decir:

$$x_i (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{in}) \quad \forall i \leftarrow (1, 2, \dots, p)$$

Consideremos asimismo que dichas variables están tipificadas, es decir:

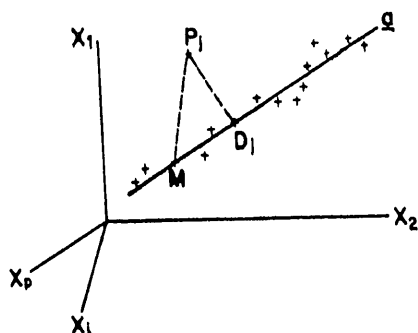
$$\bar{x}_i = \frac{1}{n} \sum_j x_{ij} = 0 \quad \text{Var}(x_i) = \frac{1}{n} \sum_j x_{ij}^2 = 1$$

para $\forall i \leftarrow (1, 2, \dots, p)$

Vamos a obtener la recta de regresión múltiple mínimo cuadrática relativa al espacio muestral p -dimensional así definido.

Sea a , cuyo versor es $\vec{u}(l_1, l_2, \dots, l_p)$ la recta que se ajusta al sistema discreto de puntos que constituyen la muestra. Si M es un punto de la misma, la ecuación de la recta será:

$$\frac{x_1 - m_1}{l_1} = \frac{x_2 - m_2}{l_2} = \dots = \frac{x_p - m_p}{l_p} \quad (1)$$



con la condición de que:

$$\sum_i l_i^2 = 1 \quad (2)$$

Consideremos la expresión que da la varianza residual:

$$\emptyset = \sum_j P_j D_j^2$$

podemos obtener la recta a condición de que $\emptyset$ sea mínima, es decir, efectuando un ajuste por mínimos cuadrados:

$$\emptyset = \sum_j (M P_j^2 - M D_j^2) = \sum_j \sum_i (x_{ij} - m_i)^2 - \sum_j [\sum_i l_i (x_{ij} - m_i)]^2$$

Obtengamos las:

$$\frac{\delta \emptyset}{\delta m_i} = 0 \quad \forall i \leftarrow (1, 2, \dots, p)$$

$$\begin{aligned} \frac{\delta \emptyset}{\delta m_i} &= -2 \sum_j (x_{ij} - m_i) + 2 \sum_j l_i \sum_i (x_{ij} - m_i) = \\ &= -2 (n m_i - n l_i \sum_i m_i) = 0 \end{aligned}$$

De las anteriores resulta:

$$\frac{m_i}{l_i} = \sum_i l_i m_i \quad \forall i \leftarrow (1, 2, \dots, p) \quad (3)$$

las expresiones (3) implican que:

$$\frac{m_1}{l_1} = \frac{m_2}{l_2} = \dots = \frac{m_p}{l_p} \quad (4)$$

Sumando un miembro cualquiera de esta expresión en todos los miembros de las (1) se tiene la recta:

$$\frac{x_1}{l_1} = \frac{x_2}{l_2} = \dots = \frac{x_p}{l_p} \quad (5)$$

esto equivale a que $m_i = 0$, o lo que es lo mismo, que la recta a pasa por el origen de coordenadas.

Teniendo en cuenta la condición (2) y añadiendo un multiplicador de Lagrange, la varianza residual queda de la forma:

$$\emptyset = \sum_j \sum_i x_{ij}^2 - \sum_j [\sum_i l_i x_{ij}]^2 + c (\sum_i l_i^2 - 1)$$

continuando el proceso de minimización de $\emptyset$, calculemos las derivadas que restan, es decir:

$$\frac{\delta \emptyset}{\delta l_i} = 0 \quad \forall i \leftarrow (1, 2, \dots, p)$$

$$\frac{\delta \emptyset}{\delta l_i} = -2 \sum_j x_{ij} \sum_i l_i x_{ij} + 2 c l_i = 0 \quad (6)$$

Si el coeficiente de correlación entre dos variables v_i y v_j se define como:

$$r_{ij} = \frac{\frac{1}{n} \sum_k (v_{ik} - \bar{v}_i)(v_{jk} - \bar{v}_j)}{\left[\frac{1}{n} \sum_k (v_{ik} - \bar{v}_i)^2 - \frac{1}{n} \sum_k (v_{jk} - \bar{v}_j)^2 \right]^{1/2}}$$

al tratarse de variables v_i y v_j tipificadas se convierte en

$$r_{ij} = \frac{1}{n} \sum_k v_{ik} v_{jk}$$

y al sustituir en la expresión (6) se tiene:

$$\frac{\delta \phi}{\delta l_i} = 2 c l_i - 2 n \sum_j l_j r_{ij} = 0$$

si se tiene en cuenta, además de la anterior, la expresión (2), resulta el sistema:

$$\sum_j l_j r_{ij} = \frac{c}{n} l_i \quad (7)$$

con $\sum_j l_j^2 = 1$

las expresiones (7) dan lugar a un sistema lineal homogéneo en l_j :

$$\left\{ \begin{array}{l} \left(r_{11} - \frac{c}{n} \right) l_1 + r_{12} l_2 + \dots + r_{1p} l_p = 0 \\ r_{21} l_1 + \left(r_{22} - \frac{c}{n} \right) l_2 + \dots + r_{2p} l_p = 0 \\ \dots \dots \dots \\ r_{p1} l_1 + r_{p2} l_2 + \dots + \left(r_{pp} - \frac{c}{n} \right) l_p = 0 \end{array} \right. \quad (8)$$

en el que para que existan soluciones distintas de la trivial se debe verificar la nulidad del determinante de los coeficientes, es decir:

$$\begin{vmatrix} \frac{c}{n} & r_{12} & \dots & r_{1p} \\ r_{12} & \frac{c}{n} & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ r_{1p} & r_{2p} & \dots & \frac{c}{n} \end{vmatrix} = 0$$

polinomio de grado p en c , que dada la simetría existente nos proporciona raíces $c_i \forall i \leftarrow (1, 2, \dots, p)$ reales no necesariamente distintas.

Para cada raíz c_i del polinomio característico (o autovalores de la ecuación secular de Laplace), de las expresiones (8) se obtiene el correspondiente vector propio $l^i_j \forall j \leftarrow (1, 2, \dots, p)$.

En el párrafo siguiente mostraremos que se ha obtenido una base ortonormal del hiperespacio vectorial de dimensión p , la cual optimiza las sumas de cuadrados extendidas a los valores de una muestra de extensión n relativa a la población considerada.

Así, pues, existen en el espacio muestral p direcciones (los p vectores propios obtenidos) preponderantes para la distribución de la masa muestral, que se corresponden con los mínimos de la varianza residual.

c) Obtención de los componentes principales.

Efectuemos ahora un cambio lineal que transforme las variables x_i en otras Y_i denominadas componentes principales, de la siguiente forma:

Si los autovalores de la matriz de correlación obtenidos en el párrafo anterior son:

$$c_1 \geq c_2 \geq \dots \geq c_p$$

se elige el vector propio asociado al mayor de ellos:

$$(l^1_1, l^1_2, \dots, l^1_p)$$

y se forma el primer componente principal según la transformación lineal:

$$Y_1 = \sum_j l^1_j x_j \quad \forall j \leftarrow (1, 2, \dots, p)$$

cuya esperanza matemática es:

$$E(Y_1) = \sum_j l_j E(x_j) = 0 \quad (9)$$

y cuya varianza vale:

$$\begin{aligned} \text{Var}(Y_1) &= E(Y_1^2) = E\left(\sum_j l^1_j x_j\right) = \sum_j E(l^1_j x_j \sum_k l^1_k x_k) = \\ &= \sum_j l^1_j \sum_k l^1_k r_{jk} \text{ para } \forall k \leftarrow (1, 2, \dots, p). \end{aligned}$$

Teniendo en cuenta la expresión (7) la varianza de la transformación lineal es:

$$\text{Var}(Y_1) = \sum_j l^1_j c_1 l^1_j = c_1 \sum_j (l^1_j)^2 = c_1 \quad (10)$$

La interpretación de las (9) y (10) nos dice que se ha obtenido el primer componente principal como un estimador centrado de máxima varianza.

Se elige ahora el segundo componente Y_2 incorrela-

do con Y_1 y de varianza la mayor posible. Se toma, por tanto, para Y_2 la transformación lineal:

$$Y_2 = \sum_j b_j x_j$$

de forma que los coeficientes b_j estén normalizados, es decir, tales que $\sum_j b_j^2 = 1$ en aras de una mayor facilidad operativa.

Se calcularán los coeficientes b_j a condición de que $\text{Var}(Y_2)$ sea máxima y además se impondrá previamente la condición de que este segundo componente esté incorrelado con el primero, es decir, que:

$$\text{cov}(Y_1, Y_2) = 0$$

Si anulamos la covarianza se tiene:

$$\begin{aligned} \text{cov}(Y_1, Y_2) &= E[(\sum_j l_j x_j)(\sum_j b_j x_j)] = \\ &= \sum_j E(b_j x_j \sum_k l_k x_k) = \sum_j b_j \sum_k l_k r_{jk} = \sum_j b_j c_1 l_j = \\ &= c_1 \sum_j b_j l_j = 0 \end{aligned}$$

que nos da la condición:

$$\sum_j b_j l_j = 0 \quad (11)$$

que es la condición para que los vectores:

$$(b_1, b_2, \dots, b_p) \\ \text{y } (l_1, l_2, \dots, l_p)$$

sean ortogonales.

Maximicemos $\text{Var}(Y_2)$ mediante aplicación del método de los multiplicadores de Lagrange, por lo que definiremos la extremal:

$$\Phi = \frac{1}{n} \sum_i (\sum_j b_j x_{ij})^2 - f (\sum_j b_j^2 - 1) - 2g (\sum_j b_j l_j)$$

que habrá de cumplir:

$$\begin{aligned} \frac{\delta \Phi}{\delta b_j} &= 0 \quad \forall j \leftarrow (1, 2, \dots, p) \\ \frac{\delta \Phi}{\delta b_j} &= \frac{2}{n} \sum_i x_{ij} \sum_j b_j x_{ij} - 2f b_j - 2g l_j = 0 \end{aligned}$$

expresión que da lugar al sistema de ecuaciones lineales:

$$\sum_k r_{jk} b_k = f b_j + g l_j \quad \forall j \leftarrow (1, 2, \dots, p)$$

si multiplicamos por $l_j \forall j \leftarrow (1, 2, \dots, p)$ cada una de las ecuaciones y sumamos:

$$\sum_k b_k \sum_j l_j r_{jk} = f \sum_j b_j l_j + g \sum_j (l_j)^2 = g$$

ya que el primer sumando es nulo por la condición (11), y el segundo es la unidad por estar normalizados los l_j .

Por otra parte, teniendo en cuenta (7):

$$\begin{aligned} \sum_k b_k \sum_j l_j r_{jk} &= \sum_k b_k \sum_j l_j r_{kj} = \sum_k b_k \frac{c_1}{n} l_k = \\ &= \frac{c_1}{n} \sum_k b_k l_k = 0 \end{aligned}$$

y por consiguiente resulta: $g = 0$.

es decir, que los autovalores de la transformación son los mismos que los de la ecuación característica de partida, y por tanto:

$$b_i^2 = l_i^2 \quad i, k \leftarrow (1, 2, \dots, p)$$

Hemos visto, pues, que existe una base ortonormal de vectores formada por los vectores propios relativos a la matriz de correlación que maximiza la varianza, es decir, que de todas las direcciones del hiperespacio en el que se ha ubicado la muestra, existen p de ellas (que forman un sistema ortogonal) en las cuales la varianza toma los mayores valores posibles.

Para seleccionar los componentes principales, basta, pues, obtener la serie ordenada de autovalores:

$$c_1 \geq c_2 \geq \dots \geq c_p$$

correspondiente a la matriz de correlación y mediante un análisis de la varianza contrastada al nivel de significación que se desee, prescindir de aquellos componentes de escasa significación.

Hemos reducido el conjunto de los parámetros $x_i, i \leftarrow (1, 2, \dots, p)$ de partida al de los $Y_i = \sum_j l_{ij} x_j, i \leftarrow (1, 2, \dots, t)$ que forman un conjunto de parámetros independientes entre sí (donde t marca el nivel de significación).

Debe ahora abordarse el problema de la determinación de los autovalores de la matriz de correlación (al menos de los más significativos), y de sus vectores propios asociados.

En la práctica es conveniente suponer que la distribución de los x_i es una distribución normal conjunta (lo que en la mayoría de los casos es bastante real), con lo que los nuevos parámetros Y_i se distribuyen cada uno de ellos asimismo en forma normal:

d) *Forma lineal de parámetros independientes representativa del fenómeno.*

Efectuada la transformación a los nuevos parámetros Y_i incorrelados entre sí, el vector z que representa el fenómeno puede obtenerse de la forma lineal:

$$z = \sum_{i=1}^n d_i Y_i + e_r$$

donde los d_i son los coeficientes obtenidos en la regresión de una muestra Z (de extensión n) de la población z , con cada uno de los Y_i independientes entre sí. Y donde e_r es un residuo aleatorio que permite cuantificar la importancia de aquellos parámetros que aun influyendo en el fenómeno, no se han considerado; así como la bondad de los datos de partida.

Así, pues, los d_i pueden obtenerse con la condición:

$$\phi = \sum_{k=1}^n (Z_k - d_i Y_{ki})^2 = \text{mínima},$$

es decir:

$$\frac{\delta \phi}{\delta d_i} = \sum_{k=1}^n Y_{ki} (Z_k - d_i Y_{ki}) = 0$$

o sea:

$$d_i = \frac{\sum_{k=1}^n Y_{ki} Z_k}{\sum_{k=1}^n Y_{ki}^2}$$

e) *Elección de los componentes principales significativos.*

En la hipótesis de que los parámetros incorrelados Y_i se distribuyen $N(0, s_i^2)$, la forma lineal del párrafo anterior:

$$z = \sum_{i=1}^p d_i Y_i + e_r$$

en la que se han supuesto las z referidas a la media (es decir, $E(z_k) = 0$), tiene por desviación típica:

$$s^2 = d_1^2 s_1^2 + d_2^2 s_2^2 + \dots + d_p^2 s_p^2 = \sum_{i=1}^p d_i^2 s_i^2$$

La variabilidad de la forma lineal es:

$$\sum_{k=1}^n z_k^2 = \sum_{k=1}^n d_1^2 Y_{k1}^2 + \sum_{k=1}^n d_2^2 Y_{k2}^2 + \dots + \sum_{k=1}^n d_p^2 Y_{kp}^2 + \text{residuo}$$

donde, al tener el primer miembro $n-1$ g. l. y 1 g. l. cada uno de los p primeros sumandos del segundo miembro, el residuo tendrá (teorema de Cochran) $n-p-1$ g. l. ...

Dividiendo por s^2 :

$$\frac{\sum_{k=1}^n z_k^2}{s^2} = \frac{\sum_{k=1}^n d_1^2 Y_{k1}^2}{s^2} + \dots + \frac{\sum_{k=1}^n d_p^2 Y_{kp}^2}{s^2} + \frac{\text{Residuo}}{s^2}$$

el primer miembro resulta ser una χ^2_{n-1} , cada uno de los primeros sumandos del segundo miembro una χ^2_1 y por tanto el residuo es una χ^2_{n-p-1} , con lo que el cociente:

$$\frac{\sum_{k=1}^n d_1^2 Y_{k1}^2}{s^2} = F_{1, n-p-1}$$

es una F de Snedecor con los g. l. que se indican.

Según lo anterior, si suponemos la variabilidad total formada por dos sumandos, uno debido al componente principal de mayor variabilidad, o sea, esta:

$$\sum_{k=1}^n d_j^2 Y_{kj}^2$$

y el otro en forma de resto global, es decir:

$$\sum_{k=1}^n z_k^2 = \sum_{k=1}^n d_j^2 Y_{kj}^2 + \text{Resto}$$

las esperanzas valen:

$$E \frac{\sum_{k=1}^n z_k^2}{s^2} = n-1; \quad E \frac{\sum_{k=1}^n d_j^2 Y_{kj}^2}{s^2} = 1;$$

$$E \frac{\text{Resto}}{s^2} = n-2$$

o lo que es lo mismo:

$$E \frac{\sum_{k=1}^n z_k^2}{n-1} = s^2; \quad E \frac{\sum_{k=1}^n d_j^2 Y_{kj}^2}{1} = s^2;$$

$$E \frac{\text{Resto}}{n-2} = s^2$$

y por tanto:

$$\frac{\sum_{k=1}^n z_k^2}{n-1} = \hat{s}^2 \quad \frac{\sum_{k=1}^n d_j^2 Y_{kj}^2}{1} = \hat{s}^2 \quad \frac{\text{Resto}}{n-2} = \hat{s}^2$$

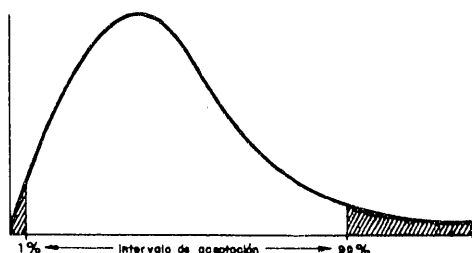
es decir, que los cocientes indicados son estimadores centrados de s^2 :

Al ser:

$$\frac{\sum_{k=1}^n d_{kj}^2 Y_{kj}^2}{\text{Resto}} = F_{1, n-2}$$

si elegimos el nivel de significación correspondiente al 1 por 100, y elegida la mayor variabilidad:

$$\sum_{k=1}^n d_{kj}^2 Y_{kj}^2$$



relativa a los componentes principales, si el cociente:

$$Q = \frac{\sum_{k=1}^n d_{kj}^2 Y_{kj}^2}{\text{Resto}}$$

supera el valor $F_{1, n-2}$ correspondiente (zona a la derecha del intervalo de aceptación), rechazamos la hipótesis de que z depende sólo del Y elegido y de nuevo necesitaremos ver si el parámetro que sigue (con variabilidad mayor) se acepta o se rechaza.

La aceptación de nuevos componentes Y_i como términos de la forma lineal continúa hasta que el cociente anterior es inferior al correspondiente valor $F_{1, n-2-h}$ (siendo h el número de parámetros que se han aceptado). En este instante aceptamos que z sólo depende ya de este parámetro y de ninguno más de los siguientes.

En el caso de que la dependencia de z del parámetro en estudio sea muy pequeña (región a la izquierda del intervalo de aceptación), es decir, cuando $1/Q$ exceda del valor $F_{n-2-h, 1}$ deberemos rechazar el que dicho parámetro se integre en la forma lineal definitiva.

El residuo e_j se puede evaluar en unidades de la variabilidad residual de los parámetros que se utilizan. Será tanto menor cuanto mayor sea la bondad de los datos originales, es decir, si los parámetros que se eligen representan fielmente al fenómeno de forma que la influencia de aquellos parámetros no considerados sea escasa, el residuo será de poca importancia.

f) Solución numérica del polinomio característico.

La obtención de las raíces de la ecuación secular se realiza mediante el método del producto escalar. Este mé-

todo particularmente útil cuando se aplica, como en este caso, a matrices simétricas, proporciona el mayor valor autovalor c_1 en el límite de las sucesivas potencias de la matriz R .

Sea $T_K = R^K T_0$ el vector K -ésimo obtenido por iteración del vector T_0 por la matriz de correlación R , se puede demostrar que:

$$\frac{(T_K \cdot T_K)}{(T_{K-1} \cdot T_K)} \rightarrow c_1$$

para un cierto valor de K .

Para asegurarnos de que la elección del vector de partida T_0 no recae sobre un vector mal condicionado respecto de los distintos hiperplanos, se definen sus componentes de forma aleatoria incorporando al sistema informático soporte del método un generador de números pseudoaleatorios.

El primer vector propio $B^{(1)}$, asociado al autovalor c_1 , se obtiene normalizando el vector T_K .

El segundo autovalor c_2 de la matriz R coincide con el primero de la matriz:

$$R_2 = R - c_1 B^{(1)} B^{(1)}$$

por lo que puede aplicarse de nuevo a esta matriz el proceso iterativo descrito con anterioridad y obtener así el autovalor c_2 y su vector propio asociado $B^{(2)}$.

En general, los sucesivos autovalores c_{j+1} se obtienen por el método descrito con anterioridad aplicado a la matriz:

$$R_{j+1} = R_j - c_j B^{(j)} B^{(j)}$$

g) Sistema informático soporte del método.

Se ha elaborado una estructura informática que permite la aplicación de lo expuesto a casos reales, así como su programación para un ordenador IBM 1130 con 16 K, lo que impide que en todo momento sea adaptable a cualquier otro tipo de ordenador.

Las etapas de que consta el modelo, en términos generales, son las siguientes:

- Almacenamiento de la matriz $n \times p$ relativa a muestras de extensión n correspondientes a p parámetros observados de los que es presumible que dependa el fenómeno.
- Obtención de la matriz r_{ij} de correlación (matriz de orden p) de los parámetros originales.
- Cálculo de los autovalores y sus vectores propios asociados correspondientes al polinomio característico de la matriz de correlación.
- Obtención de los p principales (incorrelados entre sí) como funciones lineales de los parámetros de partida.
- Representación del fenómeno mediante una función lineal de los parámetros incorrelados.

- Obtención de los coeficientes de la forma lineal por regresión directa de una muestra del fenómeno sobre los parámetros de que depende.
- Análisis de las variabilidades, contraste a nivel significativo de los componentes principales y consecuente eliminación de aquellos cuya significación sea escasa.
- Obtención de los coeficientes de la forma lineal generadora del fenómeno referida a los parámetros de origen y evaluación de la influencia de los mismos.

h) Estudio de una aplicación.

Partimos de los resultados de ensayos sobre calidad de aguas correspondientes a muestras puntuales obtenidas en el río Ebro. La finalidad de estos ensayos esta en la investigación de la calidad del agua, en primer lugar; y en la búsqueda de índices que puedan representar ésta significativamente, para su vigilancia y control sistemático.

En la hipótesis de que en el río solamente existiera una fuente de contaminación que vertería en sus aguas una colección de sustancias contaminantes $I_i \forall i \leftarrow (1, 2, \dots, i_m)$ en proporción a_i , no fijas, sino variables, tuvieran una distribución que se ajustase a una normal m -dimensional. Seguramente la realidad no sería esa, sobre todo si la fuente contaminante procede de la actividad humana, pero sabemos que aumentando el número de muestras, de los valores medios si que podríamos esperar su distribución normal; y esto es lo que suponemos en lo que sigue. Pues bien, de sucesivos muestreos conseguiremos obtener una visión de las proporciones relativas de contaminantes, con las desviaciones, etc.

Pasamos ahora a la hipótesis de existencia de i_m fuentes de contaminación. Cada una de ellas tendrá un comportamiento similar al anteriormente descrito y diremos que verterían a las aguas del río las sustancias contaminantes I_i en las proporciones $a_{ij}, \forall i, j/i \leftarrow (1, 2, \dots, i_m), j \leftarrow (1, 2, \dots, j_m)$.

En las muestras tomadas del río se encuentran mezcladas las sustancias de las distintas fuentes. Si todas ellas fueran semejantes, de modo que existiera proporcionalidad entre las distintas a_{ij} para las diferentes j , no cabe duda de que el caso sería análogo a si se tratara de una única fuente de contaminación suma de todas ellas. Algo parecido podría decirse si esto sólo se verificase para grupos parciales de ellas. Lo que puede interesar entonces es el separar las diferentes influencias distintas que se superponen en los vertidos al río.

Si suponemos que agrupamos a cada fuente de contaminación F_j todas las que pueden suponerse semejantes u homólogas, y que dos fuentes distintas son independientes, entonces ocurriría que podríamos estudiar el comportamiento de las contaminaciones separando los vectores aleatorios independientes, tal como se hace en este trabajo. Si aquellos vectores originales tenían dis-

tribución normal, también tendrán distribución normal los que obtenemos, cuya varianza dada por los valores propios calculados.

Los autovectores obtenidos son como las distintas fotografías que podrían hacerse de las variables que intervienen; serían la representación del tipo de contaminación. Ahora bien, esto es en el supuesto de que se hayan incluido en los análisis recogidos en las muestras todos y cada uno de los elementos contaminantes vertidos al agua.

Podría ocurrir el que alguno faltara o que hubiera errores sistemáticos o equivocaciones en la obtención. Podríamos entonces elegir otra medida que estuviera ligada funcionalmente de alguna forma con las cantidades de contaminantes y comparándola con ellas obtener su interdependencia.

En nuestro caso se ha elegido para este objeto la medida de la conductividad. Se ha determinado la regresión de esta variable respecto de las distintas componentes independientes, y se ha valorado la bondad del ajuste por el "tanto por diez" de la varianza que presenta la variable conductividad que es explicada por la influencia de la o las componentes que intervienen en forma significativa. En el caso en que se obtiene una buena calificación podremos confiar en que los resultados concuerden más o menos con nuestras hipótesis de partida. Por el contrario, una calificación baja será indicio de alguna anomalía que puede tener distintos orígenes:

- Errores o equivocaciones.
- Existencia de otros elementos contaminantes no incluidos en los datos recogidos.
- Discrepancia respecto de la distribución normal del vector de contaminación.

Los errores o equivocaciones pueden investigarse en su origen. Si se confirman, se ha obtenido un primer fruto de este estudio; en otro caso, investigaríamos las otras posibilidades.

En cuanto a la existencia de otros posibles elementos contaminantes, no es difícil comprobar su existencia o ausencia. Se observa en algún caso que los valores medidos de la conductividad son siempre inferiores o superiores a los calculados por medio de la regresión establecida. Pero en nuestro caso hay que señalar que por facilidad de representación de los resultados se ha limitado arbitrariamente el número de parámetros incluidos en el estudio.

Sólo resta comentar el último supuesto de no normalidad de los datos de partida. En este caso, aunque en esencia podría acometerse el estudio, igualmente hay que tener en cuenta que dejan de tener distribución normal todas las variables que pueden deducirse, y el estudio práctico se hace inabordable. Por ello si se llega a la consecuencia de que falla la normalidad lo ideal será el procurar aumentar la frecuencia de las medidas.

Esto no tiene por qué suponer un aumento de trabajo, pues se verá compensado por la disminución de las magnitudes a medir.

Del examen de los diferentes valores de los vectores propios se obtienen en los sucesivos años que se han estudiado, puede verse cómo existen pocas magnitudes que influyan sistemáticamente en la configuración de la muestra. Es, pues, preferible limitarse a las medidas de dichos factores, pero aumentando la frecuencia de las observaciones.

En el cuadro adjunto puede observarse la aparente distribución normal de la diferencia entre las muestras original y calculada. Se ha comprobado asimismo la relativa constancia de los vectores propios más significativos en el período que se ha estudiado (del que disponemos de datos) y que abarca desde 1962 a 1972, referido a años hidráulicos. Caso de mejorar el muestreo podría perfeccionarse el conocimiento de estas distribuciones y el estudio de los orígenes de las contaminaciones y efecto de los remedios que podrían aplicarse.

SISTEMA EST-1***ANÁLISIS MULTIPARAMETRICO DE LAS SUCESIVAS TENDENCIAS POBLACIONALES, MEDIANTE MUESTRAS DE EXTENSION N.

E-11 • RIO EBRO EN ZARAGOZA. AÑO 1971-72

MATRIZ DE ORDEN 12 X 10 RELATIVA A LOS PARAMETROS DE PARTIDA

X 1	X 2	X 3	X 4	X 5	X 6	X 7	X 8	X 9	X 10
91.00	233.00	0.00	216.00	215.00	128.00	32.00	1.80	9.00	120.00
25.00	193.00	6.00	215.00	305.00	130.00	23.00	1.80	5.00	145.00
14.00	143.00	3.00	144.00	385.00	88.00	50.00	2.30	2.00	126.00
17.00	199.00	6.00	135.00	170.00	140.00	21.00	2.60	4.00	111.00
37.00	182.00	6.00	69.00	89.00	83.00	12.00	2.90	3.00	39.00
34.00	182.00	6.00	77.00	103.00	90.00	19.00	1.80	2.00	45.00
34.00	164.00	6.00	107.00	132.00	70.00	26.00	1.80	2.00	61.00
85.00	158.00	6.00	108.00	103.00	75.00	18.00	2.30	3.00	60.00
73.00	172.00	6.00	107.00	153.00	91.00	19.00	1.90	2.00	76.00
351.00	185.00	6.00	145.00	190.00	104.00	23.00	2.00	4.00	117.00
18.00	179.00	23.00	222.00	298.00	139.00	38.00	2.00	5.00	140.00
168.00	196.00	6.00	163.00	200.00	115.00	31.00	2.70	3.00	117.00
M.Súsp.	CO ₃ H ⁻	CO ₃ ⁻	Cl ⁻	SO ₄ ⁻	Ca ⁺⁺	Mg ⁺⁺	M.Org.	K ⁺	Na ⁺

MATRIZ DE CORRELACION

R 1	R 2	R 3	R 4	R 5	R 6	R 7	R 8	R 9	R 10
1.0000000									
0.1840743	1.0000000								
-0.1572502	-0.1756180	1.0000000							
0.0466777	0.4684394	0.2848683	1.0000000						
-0.1281283	-0.0903599	0.1907490	0.7232966	0.9999999					
-0.0193775	0.6696864	0.3279373	0.7846574	0.4672165	1.0000000				
-0.1109435	-0.1768976	0.1540226	0.5559389	0.8461834	0.2414217	0.9999999			
-0.0077395	-0.0708057	-0.0407855	-0.3105671	-0.1717839	-0.0274473	-0.1290408	1.0000000		
0.0997894	0.8047170	-0.0357645	0.7463330	0.2545600	0.6742017	0.1566758	-0.2365301	1.0000000	
0.1452176	0.2913477	0.2248864	0.9068128	0.8555460	0.7741129	0.6490476	-0.1562823	0.5240182	1.0000000

MATRIZ DE ORDEN 12 X 10 DE LOS NUEVOS PARAMETROS INCORRELADOS

Z 1	Z 2	Z 3	Z 4	Z 5	Z 6	Z 7	Z 8	Z 9	Z 10
3.0835938	2.5155554	-0.9700017	0.7780952	1.2008607	-0.8119073	0.0603122	-0.0501614	-0.0188643	0.0003144
2.5045851	0.1379416	0.0110129	0.7669148	0.3331759	-1.0641652	-0.6462487	-0.1605030	0.1835089	0.0127513
0.7285837	-3.4886167	-1.2979650	-0.7681914	1.2785173	-0.0894509	0.0416521	0.2097615	0.0197511	0.0073517
0.5232898	0.9446685	1.3698203	-0.8869007	0.7722030	0.7024936	0.2886511	0.1887682	-0.3484304	-0.0364617
-2.9743835	0.8870752	1.2531694	-0.9960480	0.6164679	-0.4274385	-0.3797665	0.1566313	0.3364590	-0.0282312
-2.2661194	0.3183019	0.1257794	1.3002838	-0.0247668	0.2115932	0.7036177	0.1582112	0.1093894	0.0777211
-1.9966723	-0.7929861	-0.4862915	1.2386126	-0.0275349	-0.2787951	0.1417918	-0.3027750	-0.0760506	-0.0925213
-2.3091228	-0.0777293	-0.0144464	0.0442123	0.3312964	-0.3972987	-0.7250515	-0.0921552	-0.3400060	0.0669413
-1.5194342	-0.1411184	-0.2001145	0.7172498	-0.03329720	0.5395374	0.0731334	-0.0216136	0.0261419	-0.0187212
0.3384988	0.9969286	-1.7380101	-0.8961934	-2.0717638	0.1272251	-0.0600478	0.2654805	0.0153140	-0.0222219
3.2539860	-1.5978568	2.0039288	0.5639188	-1.3474154	-0.5906731	0.0528894	0.0634359	0.0205915	0.0003513
0.6331949	0.2978360	-0.0568815	-1.8619540	-0.1654753	-0.0494507	0.4490665	-0.5154033	0.0721951	0.0327310

COMPONENTES PRINCIPALES

L 1	L 2	L 3	L 4	L 5	L 6	L 7	L 8	L 9	L 10
4.441	2.049	1.117	0.990	0.872	0.284	0.158	0.049	0.034	0.001

PORCENTAJES DEL TOTAL DE LA VARIANZA

44.42	20.49	11.17	9.91	8.72	2.85	1.59	0.50	0.34	0.02
-------	-------	-------	------	------	------	------	------	------	------

MATRIZ DE LAS SUCEASIVAS TENDENCIAS MUESTRALES

B 1	B 2	B 3	B 4	B 5	B 6	B 7	B 8	B 9	B 10
0.0170212	0.2508841	-0.5228077	-0.4069550	-0.6752959	-0.0740606	0.0485025	0.1067183	0.0446811	0.1432378
0.2331933	0.5788328	0.0806258	0.0078083	0.1443751	-0.0488629	0.4634256	-0.2223683	0.5169350	-0.2221109
0.1241482	-0.2470327	0.6240634	0.1160599	-0.6198501	-0.2743694	0.0414021	0.0473247	0.1887756	-0.1468765
0.4622090	0.0111885	-0.0064625	0.0659775	-0.0566273	-0.0783593	-0.3317504	-0.6950438	-0.0327311	0.4222515
0.3660910	-0.3907211	-0.1589104	-0.0921643	0.1393946	0.2048678	-0.1839338	0.3482506	0.6617579	0.1584086
0.4022122	0.1843774	0.3350010	-0.1109894	-0.0189817	0.4098640	0.3170898	0.3308694	-0.3573024	0.4167719
0.2904951	-0.4404431	-0.2389506	-0.1018640	0.1730245	-0.5174137	0.5606228	-0.0255145	-0.2030958	0.0158296
-0.1246328	0.0088466	0.3557763	-0.8550601	0.2413396	-0.1821510	-0.1647800	-0.0802753	0.0388804	0.0152361
0.3494484	0.2840714	-0.0115557	0.1333557	0.1477907	-0.5278117	-0.4234369	0.4483940	-0.1487184	-0.0979412
0.4454758	-0.1123475	-0.0968945	-0.1907953	-0.0769380	0.3482355	-0.1264309	-0.1221980	-0.2522078	-0.7220894

PARAMETROS QUE INFLUYEN

1	2
119.581	-26.791

CALIFICACION DE LOS DATOS DE PARTIDA

9
VALORES DE LAS MUESTRAS ORIGINAL Y DE CALCULO

1259.00	1281.84
1282.00	1276.30
1124.00	1161.08
943.00	1017.76
578.00	601.05
649.00	700.98
763.00	762.97
792.00	706.45
833.00	802.58
1006.00	994.26
1414.00	1412.42
1160.00	1048.23

COEFICIENTES DE LA FORMA GENERADORA

147.52166									
-0.05066	0.56568	4.10135	1.08389	0.61362	1.79319	4.68909	-40.11880	16.32604	1.55651
1.37	3.63	6.30	16.15	15.93	12.67	13.67	4.44	9.25	16.53

POPCENTAJES DE INFLUENCIA DE LOS PARAMETROS ORIGINALES

1.37	3.63	6.30	16.15	15.93	12.67	13.67	4.44	9.25	16.53
------	------	------	-------	-------	-------	-------	------	------	-------