MUNIBE (Antropología y Arqueología)	Suplemento N.º 6	25-33	SAN SEBASTIAN	1988	ISSN 0027 - 3414

Métodos estadísticos aplicables a la reconstrucción prehistórica.

Statistical methods applicable to prehistoric reconstruction.

C. M. CUADRAS *

PALABRAS CLAVE: Comparación de modelos de regresión, Regresión sobre variables cualitativas, Análisis de coordenadas principales, Datación relativa, Estadística en arqueología.

RESUMEN

En este trabajo se exponen y discuten algunos problemas relacionados con la predicción estadística. Se aborda en especial la comparación de modelos de regresión, la regresión múltiple sobre variables cualitativas y la asignación de un individuo a un grupo cuando hay otras causas de variación. Las técnicas utilizadas están relacionadas con el análisis de proximidades, el análisis discriminante y la taxonomía numérica. Se comentan además algunos ejemplos sobre datos arqueológicos.

SUMMARY

In this paper some problems related to statistical prediction are exposed and discussed. We especially deal with the comparison of regression models, multiple qualitative variables regression, cronogical order of the data and the allocation of an individual to one of different groups when other causes of variation are present. The techniques used are related to multidimensional scaling, discriminant analysis and numerical taxonomy. In addition, some archeological examples are included.

1. INTRODUCCION

Desde que a finales del siglo pasado F. GALTON introdujera los conceptos de correlación y regresión, para ser aplicados en Antropología, han aparecido nuevos métodos de regresión y análisis multivariantes, que han tenido una enorme influencia en las ciencias biológicas, especialmente Antropología, Genética, Ecología, etc., y en las ciencias históricas, como la Arqueología y la Prehistoria. Tales métodos han permitido resolver problemas de predicción, clasificación, ordenación, datación y filogenia, propiciando una forma moderna y objetiva de tratar los datos procedentes de la observación experimental, en contraste con los métodos más tradicionales.

El presente trabajo es una contribución al tema de la predicción, en el que se abordan los siguientes aspectos: regresión múltiple entre variables cuantitativas, regresión múltiple cuando las variables independientes son cualitativas, comparación entre modelos de regresión, datación relativa de objetos, identificación de un individuo cuando hay mas de una causa de variabilidad, etc., discutiendo algunas aplicaciones. Las técnicas propuestas están relacionadas con el Análisis de Coordenadas Principales, el Análisis Discriminante y la Taxonomía Numérica.

* Universidad de Barcelona, Departamento de Estadística Diagonal, 645 08028 BARCELONA. SPAIN

2. REGRESION LINEAL MULTIPLE

La regresión lineal múltiple de una variable dependiente y sobre m variable independientes $x_1, \ldots, x_m$ es un problema bien conocido en Estadística. Dadas n observaciones independientes de las variables, que indicamos en la forma.

$$Y = (y_1 \ \ldots \ y_n)'$$

$$X = \begin{pmatrix} x_{11} & \ldots & x_{1m} \\ x_{21} & \ldots & x_{2m} \\ \ldots & \ldots & \ldots \\ x_{n1} & \ldots & x_{nm} \end{pmatrix}$$

el modelo básico es (SEBER, 1977; TROCÓNIZ, 1987).

$$y_i = \mu + x_{i1}\,\beta_1 + \ldots + x_{im}\,\beta_m\, e_i \quad i = 1, \ldots, n \qquad (1)$$

Introduciendo los vectores

$$\beta = (\beta_1 \ \ldots \ \beta_m)' \quad e = (e_1 \ \ldots \ e_n) \quad 1 = (1 \ \ldots \ 1)'$$

la formulación matricial del modelo es

$$Y = \mu\, 1 + X\beta + e \qquad (2)$$

Como es bien conocido, indicando $\tilde{X} = (1,X)$, la estimación por mínimos cuadrados de los parámetros μ, β es solución de las llamadas ecuaciones normales.

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta} \end{pmatrix} = (\tilde{X}'\tilde{X})^{-1}\,\tilde{X}'Y \qquad (3)$$

suponiendo rang X = m. Si rang X<m, entonces utilizaremos una g-inversa en (3).

Sean $\bar{x}_1, \ldots, \bar{x}_m$ las medias de las variables x. Entonces el modelo (1) puede expresarse como

$$y_i = \mu' + (x_{i1} - \bar{x}_1)\,\beta_1 + \ldots + (x_{im} - \bar{x}_m)\,\beta_m\,e_i$$

siendo $\mu' = \mu + \bar{x}_1\,\beta + \ldots + \bar{x}_m\,\beta_m$. Podemos entonces suponer que las variables x en (1) son centradas. Las soluciones de (3) son ahora

$$\hat{\mu} = \bar{y} = \frac{1}{n}\,\Sigma\, y_i \qquad \hat{\beta} = (X'X)^{-1}\,X'Y \qquad (4)$$

También es posible transformar linealmente las variables x. Consideramos la transformación.

$$W = XA \qquad (5)$$

donde A es matriz mxm, ran A = m . El modelo (2) se convierte en

$$Y = \mu\,1 + W\gamma + e \qquad (6)$$

siendo

$$A\,\gamma = \beta \qquad (7)$$

Algunos problemas importantes que interesa resolver en regresión lineal múltiple son los siguientes:

1. Predicción

El vector de valores predichos es

$$\hat{Y} = \hat{\mu}\,1 + X\,\hat{\beta} = \hat{\mu}\,1 + X\,(X'X)^{-1}\,X'Y \qquad (8)$$

que no depende del modelo. En efecto, si el modelo es (6) tendremos

$$\hat{Y}_W = \hat{\mu}\,1 + W\,(W'W)^{-1}\,W'Y$$

pero como

$$W\,(W'W)^{-1}\,W' = XA\,(A'\,X'\,XA)^{-1}\,A'X = X\,(X'X)^{-1}\,X'$$

vemos que $\hat{Y} = \hat{Y}_W$.

Dada una nueva observación de las variables independientes

$$x = (x_1, \ldots, x_m)'$$

la predicción de la variable dependiente es

$$y\,(\chi) = \mu + \chi'\,\hat{\beta} \qquad (9)$$

2. Significación de los coeficientes de regresión

Suponiendo que los términos e_i son independientes con distribución normal $N\,(0, \sigma^2)$, la hipótesis nula

$$H_0 : \beta_1 = \ldots = \beta_m = 0$$

puede decidirse utilizando el estadístico

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} \qquad (10)$$

cuya distribución es F de Snedecor con m y n—m—1 grados de libertad. R^2 es el coeficiente de determinación, que puede calcularse utilizando la fórmula.

$$R^2 = \frac{\Sigma\,(\hat{y}_i - \bar{y})^2}{\Sigma\,(y_i - \bar{y})^2} \qquad (11)$$

Obsérvese que, dada la relación (7), $\beta = 0$ implica $\gamma = 0$.

3. Comparación de dos modelos de regresión

Supongamos que tenemos otro modelo de regresión, correspondiente a las mismas variables, pero con datos sobre otra población:

$$z_i = \delta + x'_{i1}\,\gamma_1 + \ldots + x'_{im}\,\gamma_m + e'_i \quad i = 1, \ldots, n' \qquad (12)$$

¿Son los modelos (1) y (12) el mismo? Es decir, ¿es la regresión múltiple idéntica en ambas poblaciones? En otras palabras, se trata de ver si puede aceptarse la hipótesis nula.

$$H_0 : \mu = \delta,\ \beta_1 = \gamma_1 = \ldots = \beta_m = \gamma_m \qquad (13)$$

Sea Q_0^2 (1) la suma de cuadrados residual para el modelo (1)

$$Q_0^2\,(1) = \Sigma\,(y_i - \hat{y}_i)^2 = \Sigma\,(y_i - \bar{y})^2\,(1 - R^2)$$

Sea análogamente Q_0^2 (2) para el modelo (12), Q_1^2 para el modelo que resulta de juntar ambos conjuntos de datos y suponer cierta (13). Suponiendo normalidad e igualdad de varianzas, la decisión so-

bre (13) puede tomarse considerando el estadístico (CUADRAS, 1979).

$$F = \frac{[Q_1^2 - Q_0^2(1) - Q_0^2(2)]}{[Q_0^2(1) + Q_0^2(2)]} \cdot \frac{n + n' - 2(m+1)}{m+1} \qquad (14)$$

con distribución F de Snedecor con m + 1 y n + n' - 2 (m + 1) grados de libertad.

4. Ejemplo

La capacidad craneal (C) puede ser predicha conociendo la longitud del cráneo (L), la anchura parietal máxima (A), la altura basio bregma (h), utilizando la fórmula (RAO, 1952):

$$C = \alpha L^{\beta_1} A^{\beta_2} H^{\beta_3}$$

Se llega a un modelo de regresión lineal tomando logaritmos

$$\log C = \mu + \beta_1 \log L + \beta_2 \log A + \beta_3 \log H$$

siendo $\mu = \log \alpha$. Obsérvese que a es un parámetro relacionado con el tamaño del cráneo, y que las diferencias entre ß1, ß2 y ß3 influyen en la forma. Partiendodo de una muestra, se puede contrastar la hipótesis Ho : ß1 = ß2 = ß3 = 0 utilizando (10).

Supongamos ahora que tenemos dos muestras de cráneos, una procedente de una primera población (ejemplo: *Homo sapiens fossilis)* y la otra de una segunda población (ejemplo: *Homo sapiens neanderthalensis).* Entonces, utilizando (14), podemos contratar la hipótesis de que ambos grupos de cráneos tienen la misma forma y tamaño.

3. PREDICCION SOBRE VARIABLES CENTRADAS ORTOGONALES

Las consideraciones expuestas anteriormente nos permiten suponer que en el modelo (2) se cumplen las condiciones siguientes:

1) Las columnas de X son centradas,
2) Las columnas de X son ortogonales.

En otras palabras, si $X_1, \ldots, X_m$ son las columnas, entonces formulamos (3) como sigue

$$Y = \mu\, 1 + ß_1 X_1 + \ldots + ß_m X_m + e \qquad (15)$$

con las restricciones

$$\begin{aligned} X'_i \cdot 1 &= 0 \\ X'_i \cdot X_j &= 0 \ (i \neq j) \\ X'_i \cdot X_i &= \alpha_i > 0 \end{aligned} \qquad (16)$$

Para llegar a este modelo ortogonal, basta realizar una adecuada transformación lineal sobre X, lo que no altera el objetivo de predecir y en función de χ. Cuando además X_i son las componentes principales sobre la matriz original X, se puede abordar el problema de la multicolinealidad (JOLLIFE, 1986), tomando solamente las primeras componentes.

4. PREDICCIONES SOBRE VARIABLESCUALITATIVAS

Supongamos ahora que las variables x son cualitativas. Nuestro deseo es predecir y dado x

$\hat{y} = f(c_1, \ldots, c_m)$ $\quad C_1, \ldots, C_m$ son cualitativas.

Un caso particular importante lo constituyen las variables dicotómicas, es decir:

$$\begin{array}{cc} & C_1\ C_2 \ldots C_m \\ y_1 & c_{11}\ c_{12} \ldots c_{1m} \\ y_2 & c_{21}\ c_{22} \ldots c_{2m} \\ & \ldots\ldots \\ y_n & c_{n1}\ c_{n2} \ldots c_{nm} \end{array} \qquad (17)$$

siendo:

c_{ij} = 1 si la característica c_j se presenta en la observación i,
= 0 en caso contrario.

La regresión de una variable cuantitativa y sobre variables cualitativas, no puede plantearse (en general) utilizando directamente el modelo (1) por razones obvias. En este trabajo se propone un procedimiento basado en el Análisis de Coordenadas Principales sobre una matriz de similaridades o de distancias (CUADRAS, 1981a), que puede ser útil para cualquier tipo de variables cualitativas.

Indiquemos las observaciones por 1, 2, ..., n, y supongamos definida una matriz de distancias.

$$D = (d_{ij}) \qquad d_{ij} = d(i,j)$$

donde dij (dij = 0, dij = dji) es la distancia entre las observaciones i,j. Existen numerosos procedimientos para construir una distancia. Partiendo de un índice de similaridad s_{ij}, podemos definir.

$$d_{ij} = 1 - s_{ij} \qquad (18)$$

Por ejemplo, dada la información (17), podemos tomar el índice

$$s_{ij} = \frac{a}{m} \qquad \text{siendo} \qquad a = \sum_{h=1}^{m} c_{ih}\, c_{jh}$$

o bien el índice de SOKAL y MICHENER

$$s_{ij} = \frac{a + d}{m} \text{ siendo } d = \sum_{h=1}^{m} (1 - c_{ih})(1 - c_{jh})$$

Otro índice (quizás más recomendable) es el de Jaccard

$$s_{ij} = \frac{a}{a + b + c}$$

donde a es el número de dobles presencias, b es el número de presencias/ausencias, d es el número de dobles ausencias, etc.

Vamos a imponer la condición de que la distancia d, es Euclídea, es decir, que existen n puntos en RP de coordenadas

$$P_i = (X_{i1} \ldots X_{ip})$$

tales que

$$d_{ij} = d(P_i, P_j) = \sum_{h=1}^{p} (x_{ih} - x_{jh})^2 \qquad (19)$$

La condición necesaria y suficiente para que una distancia sea Euclídea es la siguiente (MARDIA et al., 1979; SEBER, 1984):

Sea $A = (a_{ij})$ la matriz n x n cuyos elementos son

$$a_{ij} = -\frac{1}{2} d_{ij}^2$$

y H la matriz

$$H = I_n - \frac{1}{n} J$$

donde I_n es la identidad, $J = 1\ 1'$ es una matriz con todos sus elementos iguales a 1. Obsérvese que $HH = H$, $H1 = 0$, $H J = 0$. Considerando entonces la matriz $B = HAH$, la distancia d_{ij} es Euclídea si B es semidefinida positiva. La dimensión euclídea es $p = \text{ran } B$ y las coordenadas euclídeas son las filas de la matriz X (nxp) tal que

$$B = X X' \qquad X'X = \Lambda \qquad (20)$$

donde $A = \text{diag}(\lambda_1 \ldots \lambda_p)$ es la matriz con los valores propios (positivos) de B. Las columnas de X contiene los vectores propios de B. La relación de y con las variables cualitativas toma ahora una expresión cuantitativa.

$$\begin{array}{llll} y_1 & x_{11} & x_{12} & \ldots\ x_{1p} \\ y_2 & x_{21} & x_{22} & \ldots\ x_{2p} \\ & \ldots\ldots & & \\ y_n & x_{n1} & x_{n2} & \ldots\ x_{np} \end{array}$$

con la propiedad de que las coordenadas euclídeas obtenidas (llamadas coordenadas principales) son compatibles con d_{ij} en el sentido de que reflejan las diferencias cualitativas entre i, j, medidas a través de d_{ij}.

Parece entonces razonable admitir el siguiente modelo de regresión lineal múltiple.

$$y_i = \mu + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_i \qquad (21)$$

donde $X = (x_{ij})$ se obtiene de (20). Además, por las propiedades que poseen las coordenadas principales, estamos ante un modelo centrado ortogonal, es decir, un modelo (15) con las restricciones (16).

GOWER y LEGENDRE (1986) realizan un amplio estudio sobre las propiedades métricas y euclídeas de las distancias, basadas en similaridades por aplicación de las fórmulas $d_{ij}^2 = 1 - s_{ij}$, $d_{ij} = 1 - s_{ij}$. El cuadro 1 nos puede orientar sobre la elección del coeficiente de similaridad.

Es obvio que la utilización de $\sqrt{1 - s_{ij}}$ es preferible sobre $1 - s_{ij}$.

CUADRO 1

Similaridad	$\sqrt{1 - s_{ij}}$ es Euclídea?	$(1 - s_{ij})$ es Euclídea?
$\frac{a}{a+b+c+d}$	SI	NO
$\frac{a}{a+b+c}$	SI	NO
$\frac{a+d}{a+b+c+d}$	SI	NO
$\frac{a}{a+2(b+c)}$	SI	NO
$\frac{a+d}{a+2(b+c)+d}$	SI	NO
$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	NO	NO
$\frac{a}{\sqrt{(a+b)(a+c)}}$	SI	NO
$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	SI	NO
$\frac{ad-bc}{ad+bc}$	NO	NO

Aplicando (4), como $X'X = \Lambda$, la estimación de los parámetros en el modelo (21) es

$$\hat{\mu} = \bar{y} \qquad \hat{\beta} = \Lambda^{-1} X'Y \qquad (22)$$

El vector de predicción es, por lo tanto

$$\hat{Y} = \hat{\mu}\, 1 + X\hat{\beta} = \hat{\mu}\, 1 + X \Lambda^{-1} X'Y$$

Aplicando (11) el coeficiente de determinación es

$$R^2 = \| \hat{Y} - \bar{y}\ 1 \|^2 / n\ s_y^2$$

Indiquemos $B^* = X\ \Lambda^{-1}\ X'$. Entonces

$$B^*\ B^* = X\ \Lambda^{-1}\ X'\ X\ \Lambda^{-1}\ X' = X\ \Lambda^{-1}\ \Lambda\ \Lambda^{-1}\ X' = B^*$$

$$\| \hat{Y} - \bar{y}\ 1 \|^2 = Y'\ B^*\ B^*\ Y = Y'\ B^*\ Y \tag{23}$$

luego obtenemos la fórmula que da el coeficiente de determinación

$$R^2 = Y'\ X\ \Lambda^{-1}\ X'\ Y / n\ s_y^2 \tag{24}$$

Se verifica $0 \leqq R^2 \leqq 1$. Cuanto más próximo es R^2 a 1 mayor es el grado de relación lineal de Y con las variables cualitativas.

5. PREDICCION DE UNA NUEVA OBSERVACION

Nos planteamos el problema de predecir el valor y_{n+1} de la variable dependiente, conocidas las características cualitativas de una nueva observación que indicaremos por n + 1. Mediante el índice de distancia que se haya elegido previamente será posible calcular las distancias.

$$d_{1,\,n+1},\ d_{2,\,n+1}\ ,\ldots,\ d_{n,\,n+1} \tag{25}$$

entre la nueva observación y cada una de las observaciones previas. Para obtener y_{n+1} es necesario hallar el conjunto de coordenadas $X = (x_1\ \ldots\ x_p)'$ cuyas distancias euclídeas a las observaciones previas coincidan con (25). Utilizando el modelo (21) y (22) la predicción será.

$$y_{n+1} = \hat{\mu} + \chi'\ \hat{\beta} \tag{26}$$

Suponiendo que χ pertenece también al espacio Euclídeo R^p, utilizaremos un resultado debido a GOWER (1968), que permite obtener las coordenadas de un punto añadido a un conjunto de puntos P_1, ..., P_n, cuyas coordenadas X son conocidas y verifican (20). La solución es

$$\chi = \frac{1}{2}\ \Lambda^{-1}\ x'\ f$$

siendo

$$f = (f_1\ \ldots\ f_n)' \qquad f_i = b_{ii} - d^2_{i,n+1}$$

donde b_{ii} es elemento diagonal de B. Sustituyendo en (26)

$$y_{n+1} = \mu + \frac{1}{2}\ f'\ X\ \Lambda^{-1}\ X'\ Y$$

Indicando $\bar{B} = X\ \Lambda^{-2}\ X'$ resulta

$$B\ \bar{B}\ B = X\ X'\ X\ \Lambda^{-2}\ X'X\ X' = X\ \Lambda\ \Lambda^{-2}\ \Lambda\ X' = B$$

luego $\bar{B}$ es una g-inversa de B. Indicando

$$b = (b_{11}\ \ldots\ b_{nn})' \qquad d = (d^2_{1,n+1}\ \ldots\ d^2_{n,n+1})'$$

obtenemos finalmente que la predicción $y_{n+1} = y\ (\chi)$ es

$$y\ (\chi) = \bar{y} + \frac{1}{2}\ (b - d)'\ \bar{B}\ Y \tag{27}$$

6. ESTUDIO DE UN CASO PARTICULAR

Supongamos que las variables x son binarias. Sea X la matriz de datos binarios: $x_{ij} = 0,1$ (ver (17). El modelo de regresión es

$$Y = \mu\ 1 + X\ \beta + e$$

Tomando la matriz de datos centrados H X = x podemos escribir

$$Y = \mu' 1 + \bar{X}\ \beta + e \tag{28}$$

Dada la información binaria x = (x1 ... x,.,,)' la predicción es

$$y\ (\chi) = \bar{y} + \chi'\ \hat{\beta} = \bar{y} + \chi'(\bar{X}'\bar{X})^{-1}\ \bar{X}'Y \tag{29}$$

Supongamos ahora que elegimos el coeficiente de similaridad de SOKAL y MICHENER s_{ij} = (a+d) / (a+b+c+d) = (a+d) /m, la matriz de similaridades es entonces

$$S = (XX' + (K - X)(K - X)' / m$$

donde K es una matriz n x m con todos los elementos iguales a 1. Es fácil ver que.

$$S = (2\ XX' + KK' - KX' - XK') / m$$

Tomando la distancia $d_{ij} = \sqrt{1 - s_{ij}}$ tenemos que

$$D^{(2)} = (d^2_{ij}) = J - S$$

$$A = -\frac{1}{2}\ D^{(2)} = -\frac{1}{2}\ (J - S)$$

Multiplicando por H, como HJ, HK y K'H se anulan, resulta

$$B = HAH = HXX'H / m$$

Es decir, prescindiendo del factor 1/m,

$$B = \bar{X}\,\bar{X}'$$

Sea ahora B = V V' donde V contiene las coordenadas principales. Tomando el modelo (21) basado en estas coordenadas, tendremos

$$Y = \mu' 1 + V \gamma + e \qquad (30)$$

donde $\bar{X}\,T = V$ para alguna matriz T de orden mxp, siendo p = ran B. Relacionando (29) con (30) tenemos que $\beta = T\gamma$. Sea $x=(x_1...x_m)'$ la información binaria para una nueva observación. Entonces la información transformada para el modelo (30) es $v = \chi' T$. Luego

$$y(v) = \bar{y} + v'\hat{\gamma} = \bar{y} + \chi' T \hat{\gamma} = \bar{y} + \chi'\hat{\beta}$$

es decir

$$y(v) = y(\chi)$$

En otras palabras, para la distancia definida a través del índice de SOKAL y MICHENER, la predicción con variables binarias utilizando el método propuesto Y la predicción utilizando el modelo clásico de regresión, dan el mismo resultado. Sin embargo, la predicción es diferente si consideramos otros índices de similaridad.

7. UN EJEMPLO DE PREDICCION CUALITATIVA

MARDIA et al. (1979) exponen una aplicación en la cual describen unos datos acerca de la presencia (1) o ausencia (0) de unos objetos sobre 6 tumbas a,b ,..., f.

		Objetos				
		1	2	3	4	5
Tumbas	a	0	0	1	1	0
	b	1	1	0	0	1
	c	0	1	1	1	1
	d	0	0	1	1	0
	e	1	0	0	0	1
	f	1	0	1	1	1

Utilizando el coeficiente de JACCARD a/(a+b+c) y la distancia $\sqrt{1 - s_{ij}}$ se obtienen las matrices

$$S = \begin{pmatrix} 1 & 0 & .5 & 1 & 0 & .5 \\ 0 & 1 & .4 & 0 & .66 & .4 \\ .5 & .4 & 1 & .5 & .2 & .6 \\ 1 & 0 & .5 & 1 & 0 & .5 \\ 0 & .66 & .2 & 0 & 1 & .5 \\ .5 & .4 & .6 & .5 & .5 & 1 \end{pmatrix}$$

$$D = \begin{pmatrix} 0 & 1 & .71 & 0 & 1 & .71 \\ 1 & 0 & .77 & 1 & .57 & .77 \\ .71 & .77 & 0 & .71 & .89 & .63 \\ 0 & 1 & .71 & 0 & 1 & .71 \\ 1 & .57 & .89 & 1 & 0 & .71 \\ .71 & .77 & .63 & .71 & .71 & 0 \end{pmatrix}$$

Los valores propios de la matriz B (ver (20)) son

$$\lambda_1 = 0.922 \quad \lambda_2 = 0.290 \quad \lambda_3 = 0.221 \quad \lambda_4 = 0.105$$
$$\lambda_5 = \lambda_6 = 0$$

es decir, los 6 objetos se localizan en un espacio Euclídeo de dimensión 4.

Supongamos que es conocida una cierta característica numérica asociada a las tumbas (por ejemplo, el orden cronológico)

$$Y' = (\underset{1}{a}\ \underset{5}{b}\ \underset{3}{c}\ \underset{2}{d}\ \underset{6}{e}\ \underset{4}{f})$$

La información binaria para una nueva tumba es

g=(01001)

Las distancias entre g y las demás tumbas son

a	b	c	d	e	f
(1	0.57	0.71	1	0.82	0.89)

Aplicando (27) se obtiene

$$y(g) = 4.153$$

La tumba g podría estar cronológicamente comprendida entre f y b.

8. DATACION RELATIVA

Supongamos que tenemos un conjunto de objetos $a_1, a_2, ..., a_n$ y que estamos interesados en ordenarlos cronológicamente. El problema se puede plantear haciendo corresponder a cada objeto unas coordenadas euclídeas en dimensión p

$$a_i: [x_1(t_i), x_2(t_i), ..., x_p(t_i)]$$

donde t_i significa el tiempo. Los objetos presentarán una sucesión cronológica

$$a_{i_1} < a_{i_2} < ... < a_{i_n}$$

si se verifica

$$t_{i_1} < t_{i_2} < ... < t_{i_n}$$

donde (i_1, i_2, ..., i_n) es una permutación de (1, 2, ..., n).

El problema tiene en sí una considerable dificultad conceptual y matemática, existiendo diversos criterios algebraicos, geométricos y estadísticos, dando lugar a una abundante literatura al respecto (HODSON et al., 1971). Una vía de solución es la siguiente. Supongamos que podemos encontrar una matriz de similaridades entre los objetos S = (s_{ij}). Apliquemos entonces un Análisis de Coordenadas Principales a la matriz de distancias D = (d_{ij}), donde d, es una distancia Euclídea obtenida por una transformación de sij. Si la distancia no es euclídea, entonces aplicaremos una transformación monótona a d_{ij}, sea a, = f (d_{ij}), y a continuación ajustaremos una distancia Euclídea a $\hat{d}_{ij}$, utilizando la técnica del Análisis de Proximidades («Multidimensional Scaling»), véase CUADRAS, 1981a). En ambos casos, el resultado será una configuración formada por n puntos P_1, P_2, ..., P, en un espacio Euclídeo R^p.

Si tomamos los dos primeros ejes principales, o las dos primeras dimensiones relevantes, obtendremos una disposición de los objetos que quedarán situados aproximadamente a lo largo de una curva (Figura 1).

Para conseguir una representación de este tipo es necesario que se verifique el llamado efecto ahorseshoe» (KENDALL, 1971), es decir, que la similaridad entre los objetos a_i, a_j sea alta si están cronológicamente próximos, y sea baja en caso contrario.

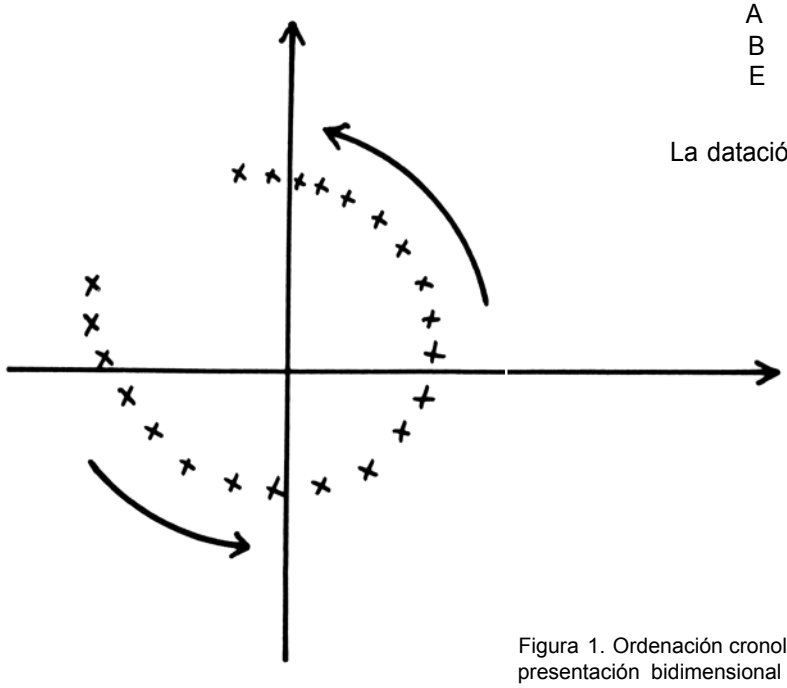

Figura 1. Ordenación cronológica de n objetos mediante una representación bidimensional por «Multidimensional Scaling».

En otras palabras, permutando convenientemente los objetos, se debe conseguir una matriz de similaridades S tal que las similaridades sean altas cuando estén próximas a la diagonal de S, y bajas para los elementos alejados de la diagonal.

Un ejemplo debido a SPAULDING (1971) (véase también SEBER, 1984), puede aclararnos esta estructura. Supongamos que 5 herramientas cortantes A, B, C, D, E han sido hechas utilizando piedra, bronce o hierro, de acuerdo con la siguiente matriz de incidencia:

	Piedra	Bronce	Hierro
A	0	1	0
B	1	1	0
C	0	1	1
D	0	0	1
E	1	0	0

Utilizando el coeficiente de Jaccard, la matriz de similaridades es

	A	B	C	D	E
A	1/2	1/2	0	0	1
B	1/2	1/3	0	1/2	1
C	1/2	1/3	1/2	0	1
D	0	0	1/2	1	0
E	0	1/2	0	0	1

A continuación, reordenando la matriz, obtenemos

	D	C	A	B	E
D	1	1/2	0	0	0
C	1/2	1	1/2	1/3	0
A	0	1/2	1	1/2	0
B	0	1/3	1/2	1	1/2
E	0	0	0	1/2	1

La datación relativa de los objetos sería entonces

E<B<A<C<D

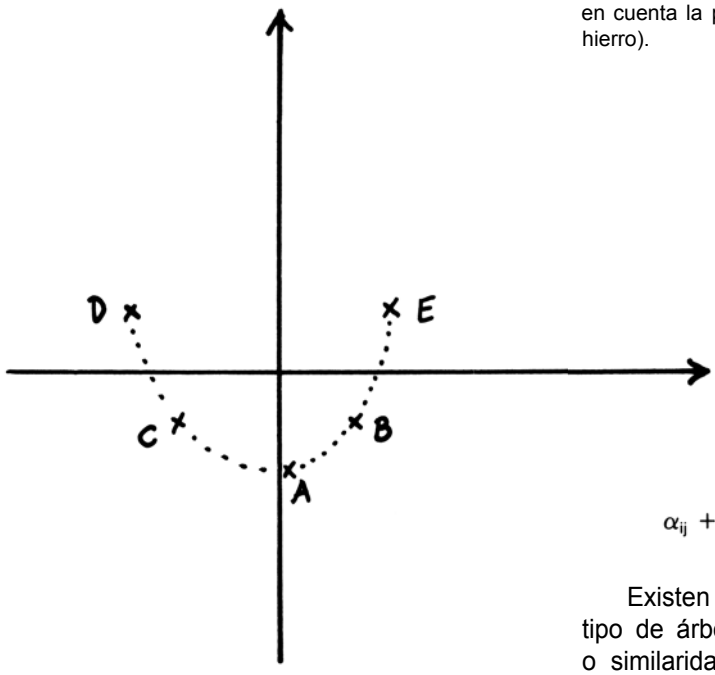

Figura 2. Ordenación cronológica de 5 herramientas, teniendo en cuenta la presencia de diversos materiales (piedra, bronce, hierro).

en correspondencia con el orden cronológico natural: piedra, piedra-bronce, bronce, bronce-hierro, hierro. Pera esta datación puede obtenerse directamente aplicando análisis de coordenadas principales, resultando las siguientes coordenadas A (0, —0.48), B (0.34, —0.13), C (—0.34, —0.13), D (—0.4,0.37), E (0.47, 0.37) (ver Figura 2).

9. ARBOLES EVOLUTIVOS ADITIVOS

Una forma muy recurrida de clasificación consiste en representar los objetos utilizando un dendograma, lo que significa aproximar la matriz de distancias entre los objetos D = (d_{ij}) a otra matriz U = (u_{ij}) donde u_{ij} es una distancia que cumple la llamada propiedad ultramétrica para toda terna i, j, k (CUADRAS, 1981a).

$$u_{ij} \leqq \max (u_{ik}, u_{ik})$$

Una aplicación a la clasificación de asentamientos puede verse en COMAMALA y ARCAS (1984).

Sin embargo, una estructura más realista o más aplicable al tipo de datos arqueológicos, es el llamado árbol evolutivo aditivo (WATERMAN et al., 1977). Se trata de una representación de objetos mediante los vértices de un árbol conexo sin ciclos, utilizando como métrica la distancia entre dos vértices igual a la suma de longitudes de los segmentos que los unen. Matemáticamente equivale a aproximar D a una matriz de distancias $\alpha = (\alpha_{ij})$ verificando, para toda cuaterna i,j,k,m, la llamada desigualdad aditiva.

$$\alpha_{ij} + \alpha_{km} \leqq \max (\alpha_{jm} + \alpha_{ki}; \alpha_{im} + \alpha_{jk})$$

Existen diversos algoritmos para construir este tipo de árbol partiendo de una matriz de distancias o similaridades (ARCAS y CUADRAS, 1987). Una vez obtenida una representación en árbol aditivo, en la que se supone la existencia de un vértice raíz, la datación de objetos arqueológicos es una consecuencia de la misma representación (ver Figura 3).

La forma de representación mediante árbol aditivo puede aplicarse a la clasificación de objetos en general, permitiendo descubrir las relaciones entre los mismos, como una alternativa al conocido dendograma, pues la desigualdad aditiva es menos restrictiva que la ultramétrica. Sobre las ventajas de una representación sobre la otra, véase CUADRAS et al. (1985).

10. UNA GENERALIZACION DEL ANALISIS DISCRIMINANTE

Se puede generalizar la asignación de un individuo a uno de entre do o más grupos, utilizando una determinada información multivariante, en la línea del Análisis Canónico Generalizado (CUADRAS, 1974, 1980, 1981a, 1981b). Por ejemplo, supongamos que queremos decidir si ciertas mandíbulas pertenecen a hombres o a mujeres, pero las muestras son escasas o proceden de diferentes yacimientos y no es posible un análisis para cada yacimiento por separado (VAN VARK, 1985). Entonces se puede discriminar entre sexos, eliminando previamente la variabilidad entre yacimientos, interpretando la combinación sexo x yacimiento, como un diseño de dos factores. Se puede abordar incluso el caso de observaciones faltantes o datos no balanceados (CUADRAS, 1983). La función discriminante obtenida podría clasificar el sexo independientemente del yacimiento.

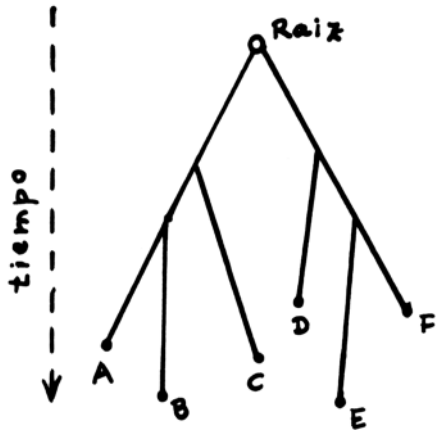

Figura 3. Representación de objetos mediante un árbol aditivo evolutivo.

BIBLIOGRAFIA

ARCAS, A. y CUADRAS, C.M.

1987. Métodos geométricos de representación mediante modelos en árbol. *Pub. Bioest. Biomat.* 20, Univ. Barcelona, 67 pp. Barcelona.

COMAMALA, A. y ARCAS, A.

1984. Clasificación por métodos estadísticos de unos asentamientos epipaleolíticos entre el Júcar y el Segura. *Arqueología Espacial 1,* 135-147, Teruel.

CUADRAS, C.M.

1974. Análisis discriminante de funciones paramétricas estimables. *Trab. Estad. Inv. Oper. 25* (3). 3-31.

1979. Sobre la comparació estadistica de corbes experimentals. *Qüestió 3* (1), 1-10.

1980. Análisis discriminante. Posibilidades de aplicación a la Prehistoria. Comun. II y III Col. *Int. Prehistoria, Morella,* 1975, 1976 (Anna Mir, ed.). Morella, 108-117.

1981.a *Métodos de Análisis Multivariante.* Euníbar, Barcelona.

1981.b Análisis y representación multidimensional de la variabilidad. *Intern. Symp. Concept. Meth. Paleo.* 287-297. Barcelona.

1983. Diseños no balanceados y con observaciones faltantes en MANOVA. *Actas XIIIJorn. de Estad., I. Oper. Inform.,* Dep. Estadística, Univ. de Valladolid, Vol. 2, secc. 2, 168-173.

CUADRAS, C.M.; OLLER, J.M.; RIOS, M. y ARCAS, A.

1985. Métodos geométricos de la Estadística. *Qüestió 9* (4). 219-250.

GOWER, J.C.

1986. Adding a point to vector diagrams in multivariate analysis. *Biometrika 55,* 582-585.

GOWER, J.C. y LEGENDRE, P.

1986. Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification 3,* 5-48.

HODSON, F.R.; KENDALL, D.G. y TAUTU, P. (eds).

1971. *Mathematics in the Archaelogical and Historical Sciences.* Edinburgh University Press, Edinburgh.

JOLLIFE, I.T.

1986. *Principal Components Analysis.* Springer-Verlag, New York.

KENDALL, D.G.

1971. Seriation from abundance matrices. En: *Mathematics in the Archaelogical and Historical Sciences.* (Hodson, F.R.; Kendall, D.G. y Tautu, P., eds.). Edinburgh University Press Edinburgh, 215-251.

MARDIA, K.V.; KENT, J.T. y BIBBY, J.M.

1979. *Multivariate Analysis.* Academic Press, London.

RAO, C.R.

1952. *Advanced Statistical Methods in Biometric Research.* J. Wiley, New York.

SEBER, G.A.F.

1977. *Linear Regression Analysis.* J. Wiley, New York.

1984. *Multivariate Observations.* J. Wiley, New York.

SPAULDING, A.C.

1971. Some elements of quantitative archaelogy. En: *Mathematics in the Archaelogical and Historical Sciences.* (Hodson F.R.; Kendall, D.G. y Tautu, P., eds.). Edinburgh University Press, Edinburgh, 3-13.

TROCONIZ, A.F. DE

1987. *Modelos Lineales.* Serv. Edit. Univ. País Vasco.

VAN VARK, G.N.

1985. Multivariate Analysis in Physical Anthropology. En: *Multivariate Analysis* VI (P.R. Krishnaiah, ed.), Elsevier Science Publishers, B.V., North-Holland, Amsterdam, 599-611.

WATERMAN, M.S.; SMITH, T.F.; SINGH, M. y BEYER, W.A.

1977. Additive Evolutionary Trees. *J. Theor. Biology 64,* 199-213.