NUEVOS METODOS Y CASOS PARA LA PREDICCION PROBABILISTICA DE FENOMENOS METEOROLOGICOS MEDIANTE EL ANALISIS DISCRIMINANTE

por Anselmo Peinado Serna y Carlos Almarza Mata
Sección de Meteorología Hidrológica del Instituto Nacional de Meteorología

RESUMEN

En la "Revista de Meteorología" de la Asociación Meteorológica Española, n.º 1, junio de 1983, 33-43, se expuso por los mismos autores una generalización del análisis discriminante de Fisher para N parámetros y r tipos o especies, con vistas a su aplicación a la predicción probabilística de fenómenos meteorológicos. Ahora se detalla un método, la transformación en Funciones Empíricas Ortogonales, para conseguir que los parámetros que determinan la función discriminante sean incorrelacionados y en menor número, con la mínima pérdida en la varianza de la información. Se adaptan las fórmulas del análisis discriminante a las nuevas coordenadas y se ve que resultan más sencillas para el cálculo. Se aplica el nuevo método a la predicción de precipitación en el área de Madrid con 14 parámetros predictores y se propone una metodología para el caso de 2.475 predictores. Por otra parte, se ofrecen métodos para el cálculo de la mayor raíz E/R, previo a la obtención de los coeficientes de la función discriminante.

1. Introducción y bases teóricas

Como ya se expuso en nuestro anterior artículo [1], para caracterizar un determinado tipo de tiempo en una zona se pueden utilizar varios parámetros físicos, como predictores, que razonablemente sean responsables de dicho tipo de tiempo o fenómeno meteorológico en un estado de su posible gama.

1.1. Ahora bien, dichos parámetros predictores son muy diversos, por lo cual conviene compensar las distintas unidades utilizadas para su medida, ya que pueden introducir un peso artificial sobre la importancia de la respectiva varianza, y esto se consigue si se tipifican todos ellos, tomando como unidad la respectiva desviación típica, cosa por otra parte conveniente para el análisis de la varianza, en el cual se basa el análisis discriminante de Fisher utilizado.

1.2. Por razones de cálculo es conveniente que el número de parámetros predictores sea lo más pequeño posible, pero en un fenómeno meteorológico influyen, sin embargo, un gran número de parámetros, no sólo magnitudes físicas distintas sino también su localización en distintos puntos geográficos y niveles atmosféricos. Entonces se nos plantea el problema de reducir el número de éstos, con la condición de conservar el máximo de información que la varianza de cada uno aporta. La razón de que sean tantos los parámetros influyentes es que no son independientes entre sí. La teoría de Funciones Empíricas Ortogonales [3] y [4], nos brinda la solución, ya que nos permite transformar los parámetros iniciales en otros, que al menos están incorrelacionados entre sí, y con la misma varianza total pero repartida de forma que unos cuantos cogen la mayor parte de la varianza y el resto una parte mínima, que puede despreciarse, ya que posiblemente no es otra cosa que "ruido", sin información útil.

El método aplicado es el expuesto en [5] y que, básicamente, es el siguiente:

Los N parámetros (x_1, x_2, ... x_N) iniciales pueden considerarse coordenadas en un sistema cartesiano rectangular, con una muestra empírica de n individuos o puntos $(x_1^{(i)}, x_2^{(i)}, \ldots x_N^{(i)})$, $1 \leqslant i \leqslant n$. Para analizar la varianza y para compensar las distintas unidades en cada parámetro, conviene centrar y tipificar, cambiando a

$$Y_p = \frac{X_p - \overline{X}_p}{S_p} \qquad [15]$$

donde

$$\overline{X}_p = \frac{1}{n} \sum_{i=1}^{n} X_p^{(i)}$$

$$S_p^2 = \frac{1}{n} \sum_{i=1}^{n} (X_p^{(i)} - \overline{X}_p)^2 \qquad [16]$$

con lo cual las nuevas y_p tienen todas media 0 y varianza 1, por lo que la varianza total es N y los coeficientes de correlación vienen dados por

$$r_{pk} = \frac{1}{n} \sum_{i=1}^{n} (Y_p^{(i)}, Y_k^{(i)}) \qquad [17]$$

Se hace una transformación lineal de las coordenadas Y_p por cambio a otro sistema de referencia también cartesiano rectangular con igual origen en $(\overline{x}_1, \overline{x}_2, \ldots \overline{x}_n)$, en el que las nuevas coordenadas serán $(\delta_1, \delta_2, \ldots \delta_n)$, con la condición de maximinimizar el reparto de la varianza,

$$V_h = \frac{1}{n} \sum_{i=1}^{n} \delta_h^{(i)2} \; (1 \leqslant h \leqslant N) \qquad [18]$$

sin alterar la varianza total N, que sólo depende del cuadrado de la distancia al origen, invariantes en esta transformación. Las nuevas coordenadas son las distancias $\delta_h^{(i)}$ a los N hiperplanos del nuevo sistema de referencia, cuyas ecuaciones son

$$a_{h1}y_1 + a_{h2}y_2 + \ldots + a_{hN}y_N = 0$$
$$(h = 1, 2, \ldots, N) \qquad [19]$$

y por tanto

$$\delta_h^{(i)} = \frac{a_{hi}y_1^{(i)} + a_{h2}y_2^{(i)} + \ldots + a_{kN}y_N^{(i)}}{\sqrt{a_{h1}^2 + a_{h2}^2 + \ldots + a_{hN}^2}} \qquad [20]$$

Las a_{hp} $(1 \leqslant p \leqslant N)$ se determinan al extremar V_h y por tanto, anular las derivadas, respecto de a_{hp}; de (18) sustituyendo (20), o sea

$$V_h = \frac{1}{n} \sum_{i=1}^{n} \frac{(a_{h1}y_1^{(i)} + a_{h2}y_2^{(i)} + \ldots + a_{hN}y_N^{(i)})^2}{a_{h1}^2 + a_{h2}^2 + \ldots a_{hN}^2} = \frac{A}{B} \qquad [21]$$

$$\frac{\delta V_h}{\delta a_{hp}} = \frac{\frac{2B}{n} \sum_{i=1}^{n} y_p^{(i)}(a_{h1}y_i^{(i)} + \ldots + a_{hN}y_N^{(i)}) - 2Aa_{hp}}{B^2} = 0 \qquad [22]$$

que, multiplicando por $\frac{B}{2} \neq 0$ y según (21) es

$$\frac{1}{n} \sum_{i=1}^{n} y_p^{(i)}(a_{hi}y_1^{(i)} + \ldots + a_{hN}y_N^{(i)}) = V_h \times a_{hp} \qquad [23]$$

o sea, por (17)

$$\begin{cases} a_{h1}r_{p1} + a_{h2}r_{p2} + \ldots + a_{hN}r_{pN} = V_h a_{hp} \\ p = 1, 2, \ldots, N \end{cases} \qquad [24]$$

sistema lineal homogéneo de N ecuaciones, cuya condición de compatibilidad lleva a la ecuación de grado N en V_h:

$$\begin{vmatrix} 1 - V_h & r_{12}\ldots & r_{1N} \\ r_{21} & 1 - V_h\ldots & r_{2N} \\ \cdots & \cdots & \cdots \\ r_{N1} & r_{N2}\ldots & 1 - V_h \end{vmatrix} = 0 \qquad [25]$$

es decir, la ecuación secular de la matriz simétrica de correlación $\{r_{pk}\}$ y cuyas raíces V_h son los valores propios, cada uno de los cuales sustituido en (24) da los coeficientes de los nuevos hiperplanos de referencia, componentes de los vectores propios según sus respectivos ejes perpendiculares, y de ahí, las nuevas coordenadas (20), que podemos reducir a las correspondientes a los Π mayores V_h, que absorban la mayor parte de la varianza, $\frac{1}{N} \sum_{i=1} V_h \times 100$ en porcentaje de la varianza total, a las que se les suele llamar coordenadas o componentes principales.

La ortogonalidad de las nuevas coordenadas se demuestra viendo que $\sum_{i=1}^{n} \delta_h^{(i)} \cdot \delta_k^{(i)} = 0$ si $h \neq k$, y en efecto su numerador es 0 pues, según (20) y (23) es

$$\sum_{i=1}^{n} (a_{ah1} y_1^{(i)} + a_{hN} y_N^{(i)}) \cdot (a_{k1} y_1^{(i)} + \dots + a_{kN} y_n^{(i)}) =$$

$$= a_{h1} \cdot nV_k a_{k1} + \dots + a_{nh} \cdot nV_k a_{kN} =$$

$$= a_{k1} \cdot nV_h a_{h1} + \dots + a_{kN} \cdot nV_h a_{hN} \qquad [26]$$

según que se aplique la propiedad distributiva al primero o segundo factor en (26); luego

$$(a_{h1} a_{k1} + \dots + a_{hN} a_{kN}) (V_k - V_n) \cdot n = 0 \qquad [27]$$

y si $V_k \neq V_h$ entonces será cero el primer factor de (27), por tanto (26). Si $V_k = V_h$ es raíz de (25) con multiplicidad m, la matriz simétrica del sistema (24) es de característica $N - m$ y el sistema es m-mente indeterminado y se escogen m soluciones ortogonales dos a dos, que anulan al primer factor de (27) y, por tanto, a (26).

1.3. Con estos nuevos parámetros δ_p, las fórmulas de [1] quedan simplificadas así:

$$Z_t^{(i)} = \sum_{p=1}^{\pi} \lambda_p \delta_p^{(i)} \qquad [28]$$

$$S^*pk = \sum_{t=1}^{r} n_t \bar{\delta}_{pt} \bar{\delta}_{kt} =$$

$$S_{pk} = \sum_{t=1}^{r} \left[\left(\sum_{i=1}^{n_t} \delta_{pt}^{(i)} \delta_{kt}^{(i)} \right) - n_t \bar{\delta}_{pt} \bar{\delta}_{kt} \right] = \qquad [29]$$

$$= \begin{cases} - \sum_{t=1}^{r} n_t \bar{\delta}_{pt} \bar{\delta}_{kt} = - S^*_{pk}, \text{ si } p \neq k \\ \sum_{t=1}^{r} \left[\sum_{i=1}^{n_t} \delta_{pt}^{(i)2} \right] - \sum_{t=1}^{r} n_t \bar{\delta}_{pt}^2 = \\ \sum_{i=1}^{n} D_p^{(i)2} - S^*_{pp}, \text{ si } p = k \end{cases}$$

obsérvese que $\sum_{i=1}^{n} \delta_p^{(i)2}$ es n veces la varianza δ_p en toda la base empírica, es decir nV_p, según (18), por lo que

$$S_{pp} = nV_p - S^*_{pp}$$

2. Aplicación a un caso de predicción cuantitativa de la precipitación con catorce parámetros predictores

2.1. Debe tomarse lo siguiente únicamente como ejemplo de aplicación ya que la base empírica utilizada es muy escasa, sólo 96 días, y además la elección de los parámetros predictores ha venido condicionada por las posibilidades de aceso a los bancos de datos y por los medios de cálculo disponibles.

En [1] se justifica y se expone la obtención de dos parámetros básicos en la génesis de la precipitación: "HQ" representativo del agua precipitable en cantidad y concentración y "VV" representativo de la velocidad vertical del aire. Allí puede verse la aplicación de estos dos predictores a la predicción de precipitación diaria en Galicia, de acuerdo con lo expuesto en 1.1, para tres tipos: A_1 = sin precipitación $\geqslant$ 1 mm = A; A_2 = precipitación entre 1 y 10 mm = B; A_3 = sin precipitación $\geqslant$ 10 mm = C, con una base empírica de 96 días de invierno.

Veamos ahora un caso de aplicacion de lo expuesto en 1.2 y 1.3 de este trabajo. Como zona hemos elegido Madrid y sus alrededores, con tres tipos: A_1 = sin precipitación, diaria de 6 a 6Z $\geqslant$ 0,1 mm en las estaciones Madrid (Retiro), Madrid (CU), Madrid (Barajas), Base Aérea de Cuatro Vientos y Base Aérea de Getafe; A_2 = con precipitación, diaria de 6 a 6Z, de 0,1 a 6,0 mm en alguna de dichas estaciones y A_3 = con precipitación, diaria de 6 a 6Z, > 6,0 mm, en alguna de las mismas estaciones. La base empírica está formada por 96 días entre 19 de abril y 11 de junio de 1981 a 1983, con los siguientes 14 parámetros predictores sinópticos a 12 Z:

x_1 = Q = Agua precipitable calculada según [1].

x_2 = H = Porcentaje de agua precipitable respecto de la saturación calculada según [1].

x_3 = VV = Velocidad vertical según [1].

x_4 = ΔT = Diferencia entre la temperatura que alcanzaría una partícula de aire llevada pseudoadiabáticamente de 850 mb a 500 mb y la temperatura del aire a 500 mb, sobre Madrid, a partir de los mapas del Boletín Meteorológico Diario del INM, utilizando el diagrama de Stüve.

x_5 a x_9 = $GP_{1000}{}^{(i)}$ = Geopotencial de 1000 mb en los cinco puntos 40°05'W, 45°00', 45°10'W, 35°10'W y 35°00', por este orden.

x_{10} a x_{14} = $GP_{500}{}^{(i)}$ = Geopotencial de 500 mb en los mismos cinco puntos.

ΔT es un índice de inestabilidad térmica y los GP se introducen para tener en cuenta la circulación atmosférica en el entorno de la zona en sus efectos de vorticidad, convergencia o divergencia, confluencia o difluencia y situación sinóptica en general.

Las medias y desviaciones típicas de estos parámetros pueden verse en el cuadro I.

La matriz de correlación figura en el cuadro II y los cuatro autovalores mayores son:

$V_1 = 6{,}13147 = 43{,}80$ % de N = 14
$V_2 = 2{,}66005 = 19{,}00$ % de N = 14
$V_3 = 1{,}76582 = 12{,}61$ % de N = 14
$V_4 = 1{,}11420 = 7{,}96$ % de N = 14

los cuales absorben el 83,37 % de toda la varianza y cada uno de los restantes no llega al 5 %, por lo que se desprecian.

Cuadro 1.—Media, desviación típica y CV de los parámetros en la base empírica inicial de tamaño n = 96.

Parám. inic.	p	$\overline{xp}$	s_p	$s_p/\overline{xp}$
Q (mm)	1	22,20833	7,57864	34,1
H (%)	2	68,27083	16,28680	23,9
VV (mb/h)	3	−3,06250	6,79432	221,9
ΔT (° C)	4	1,29167	4,76077	368,6
$GP^{(0)}_{1000}$(damgp)	5	10,93750	3,10179	28,4
$GP^{(1)}_{1000}$ (damgp)	6	10,20833	4,23752	41,5
$GP^{(2)}_{1000}$ (damgp)	7	9,66667	6,38629	66,1
$GP^{(3)}_{1000}$(damgp)	8	13,5625	2,68411	19,8
$GP^{(4)}_{1000}$ (damgp)	9	11,35417	2,56165	22,6
$GP^{(0)}_{500}$ (damgp)	10	565,36458	11,33773	2,0
$GP^{(1)}_{500}$ (damgp)	11	560,21875	11,56792	2,1
$GP^{(2)}_{500}$ (damgp)	12	555,76042	12,16873	2,2
$GP^{(3)}_{500}$ (damgp)	13	571,83333	10,05057	1,8
$GP^{(4)}_{500}$ (damgp)	14	576,37500	9,19154	1,6

CUADRO II.—Matriz de correlación $\{r_{pk}\}$

p \ k	Q	H	VV	ΔT	$GP^{(0)}_{1000}$	$GP^{(1)}_{1000}$	$GP^{(2)}_{1000}$	$GP^{(3)}_{10000}$	$GP^{(4)}_{1000}$	$GP^{(0)}_{500}$	$GP^{(1)}_{500}$	$GP^{(2)}_{500}$	$GP^{(3)}_{500}$	$GP^{(4)}_{500}$
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1,000													
2	0,198	1,000												
3	−0,324	−0,380	1,000											
4	0,682	−0,156	−0,392	1,000					SIMETRICA					
5	0,068	−0,322	0,220	0,014	1,000									
6	0,194	−0,343	−0,068	0,201	0,700	1,000								
7	0,053	−0,133	−0,070	−0,055	0,528	0,723	1,000							
8	−0,094	−0,239	0,315	−0,160	0,637	0,209	0,201	1,000						
9	0,021	−0,329	0,321	0,046	0,698	0,322	0,080	0,606	1,000					
10	0,556	−0,516	0161	0,430	0,572	0,581	0,313	0,329	0,461	1,000				
11	0,612	−0,365	−0,055	0,546	0,331	0,601	0,220	−0,018	0,227	0,829	1,000			
12	0,364	−0,357	0,044	0,222	0,518	0,771	0,751	0,160	0,188	0,732	0,669	1,000		
13	0,355	−0,532	0,243	0,251	0,603	0,438	0,232	0,561	0,553	0,865	0,552	0,537	1,000	
14	0,577	−0,443	0,030	0,531	0,464	0,437	0,084	0,261	0,424	0,873	0,787	0,471	0,795	1,000

Los correspondientes vectores propios unitarios o coeficientes de la transformación (20) figuran en el cuadro III.

La ecuación (14) de [1] es, haciendo $x = E/R$:

$$x^4 - 0{,}9286278211x^3 + 0{,}1249253578x^2 -$$
$$- 1{,}7 \times 10^{-9}x + 9{,}5649 \times 10^{-23} = 0$$

y su mayor raíz es $(E/R)_{MAX} = 0{,}7654153$ y para este valor, los λ_k solución de (13) de [1] son

$$\begin{aligned} \lambda_1 &= -881{,}294484 \\ \lambda_2 &= -500{,}56511 \\ \lambda_3 &= 53{,}17745 \\ \lambda_4 &= 209{,}44653 \end{aligned} \qquad [30]$$

Las Z obtenidas para cada tipo de la base empírica pueden verse en el cuadro IV, ordenadas cuantitativamente, con sus respectivas medias, medianas, desviaciones típicas, coeficientes de asimetría (γ) y curtosis (β_2), que muestran no apartarse demasiado de la distribución normal, por lo que en el cálculo de las probabilidades las consideramos normales, $\Phi(Z)$.

Sus varianzas son parecidas, y la F de Snedecor para (2,93) g de l vale 35,5918 cuya Q(F) es 3,32 $\times 10^{-12}$ que indica una discriminación altamente significativa.

2.2. Veamos el resultado en algunos días exteriores a la base empírica, pero de fechas próximas a las utilizadas:

2.2.1. *Día 23 de junio de 1983.* Los parámetros predictores x_p tomaron los valores:

$Q = 37 \qquad H = 78$

$VV = -10 \qquad AT = +9$

$GP^{(i)}_{1000} = 9; 13; 16; 14; 11.$

$GP^{(i)}_{500} = 574; 573; 570; 580; 581.$

Con (15) y (20) se obtienen

$\delta_1 = 2{,}13 \qquad \delta_3 = 0{,}62$

$\delta_2 = -2{,}53 \qquad \delta_4 = 0{,}73$

Con [28] y [30] se obtiene

$Z = -425$, para la cual $P(A_1)$, $P(A_2)$ y $P(A^3)$ son proporcionales, respectivamente, a

$$1 - \Phi\left(\frac{-425 + 1106}{1591}\right), \ \Phi\left(\frac{-425 - 822}{2288}\right),$$
$$\Phi\left(\frac{-425 - 3174}{1355}\right)$$

es decir,

$$\frac{P(A_1)}{0{,}334314} = \frac{P(A_2)}{0{,}292871} = \frac{P(A_3)}{0{,}003953}$$

de donde

$p(A_1) = 53\ \%$
$P(A_2) = 46\ \%$
$P(A_3) = 1\ \%$

La precipitación registrada fue:

2 mm en Madrid CU
0 mm en Barajas y Retiro
0,6 mm en Cuatro Vientos
0,8 mm en Getafe

es decir del tipo A_2, rondando el A_1.

2.2.2. Día 17 de abril de 1983. Predictores, en el orden del caso anterior: 28; 91; −10; 0; 2; 2; 3; 5; 4; 556; 556; 546; 558 y 572, y se obtiene Z = 5495, para la que

$P(A_1) = 0\ \%$
$P(A^2) = 2\ \%$
$P(A_3) = 98\ \%$

La precipitación registrada fue 11,0 mm en Barajas; 7,1 en Retiro; 5,0 en Madrid CU; 8,3 en Cuatro Vientos; 8,4 en Getafe, es decir tipo A_3.

2.2.3. Día 13 de abril de 1983. Análogamente los parámetros predictores son: 12; 46; −9; −6; 17; 23; 20; 14; 17; 576; 566; 568, y 580, a los que corresponde Z = −4002, con las siguientes probabilidades

$P(A_1) = 98\ \%$
$P(A_2) = 2\ \%$
$P(A^3) = 0\ \%$

No hubo precipitación en toda la zona, es decir tipo A_1.

CUADRO III.—Componentes de los vectores propios unitarios o matriz de la transformación. $a_{hj}/\sqrt{\sum_j a_{hj}^2}$

j \ h	1	2	3	4
1	0,18969	−0,44185	−0,10148	0,27712
2	−0,21700	−0,18172	0,15870	0,57907
3	0,05220	0,40220	−0,23037	−0,37375
4	0,16681	−0,43108	−0,17444	0,06572
5	0,29979	0,27408	0,13004	0,29313
6	0,30404	0,01944	0,40257	−0,02296
7	0,19400	0,08895	0,60787	0,01889
8	0,17699	0,38598	−0,12665	0,41146
9	0,22840	0,29194	−0,22887	0,32453
10	0,38224	−0,07111	−0,11905	−0,09365
11	0,32007	−0,26310	−0,03430	−0,19694
12	0,31789	−0,04920	0,35734	−0,18439
13	0,34473	0,09153	−0,21826	0,02881
14	0,33834	−0,14645	−0,27096	−0,00697

CUADRO IV.—Valores de la función discriminante.

Tipo A_1 Sin prec. > 0,1				Tipo A_2 Prec. de 0,1 a 0,6		Tipo A_3 Prec. > 6,0	
i	$Z_1^{(i)}$	i	$Z_1^{(i)}$	i	$Z_2^{(i)}$	i	$Z_3^{(i)}$
1	−5016	30	−1015	1	−3493	1	1238
2	−4650	31	− 882	2	−2715	2·	1767
3	−4607	32	− 728	3	−2416	3	1814
4	−3896	33	− 669	4	−2270	4	2107
5	−3803	34	− 555	5	−1198	5	2431
6	−3617	35	− 550	6	− 975	6	2478
7	−2756	36	− 493	7	− 902	7	2571
8	−2718	37	− 457	8	− 626	8	2885
9	−2556	38	− 416	9	− 480	9	3467
10	−2527	39	− 308	10	− 12	10	3680
11	−2344	40	− 243	11	460	11	4408
12	−2235	41	− 237	12	607	12	4566
13	−2233	42	− 212	13	1436	13	5071
14	−2125	43	− 140	14	1745	14	6008
15	−2050	44	47	15	2029		
16	−1997	45	98	16	2063		
17	−1965	46	268	17	2123		
18	−1873	47	269	18	2218		
19	−1843	48	288	19	2818		
20	−1725	49	500	20	3183		
21	−1638	50	551	21	3437		
22	−1512	51	568	22	3779		
23	−1492	52	679	23	4328		
24	−1472	53	832	24	4589		
25	−1429	54	897				
26	−1378	55	1237				
27	−1252	56	1277				
28	−1233	57	2009				
29	−1056	58	2249				
$\bar{Z}_1 = -1105{,}845$				$\bar{Z}_2 = 822{,}000$		$\bar{Z}_3 = 3174{,}071$	
$S_1 = 1590{,}695$				$S_2 = 2287{,}750$		$S_3 = 1354{,}884$	
$\gamma_1 = -0{,}359$				$\gamma_2 = -0{,}147$		$\gamma_3 = 0{,}5596$	
$\beta_{21} = 2{,}939$				$\beta_{22} = 1{,}951$		$\beta_{23} = 2{,}268$	
$md_1 = -1036$				$md_2 = 1022$		$md_3 = 2701$	

2.3. Los medios de cálculo utilizados han sido, para la obtención de la matriz de correlación, valores propios y vectores propios, un programa para un IBM 360 facilitado por el doctor Román Alba, R., a quien quedamos sumamente reconocidos, y al resto de cálculos se han hecho con una TI Programable 59 con impresora PC – 100 C y una HP 33 E.

3. Metodología propuesta para un caso de 2.475 parámetros predictores

Se trata de los valores en 99 puntos de una retícula entre 26°N y 49°N, 26°W y 13°E, de las componentes zonal y meridiana del viento, velocidad vertical, temperatura y razón de mezcla, en total 5 magnitudes, a los cinco niveles de 1.000, 850, 700 y 500 mb y 300 mb, es decir 2.475 parámetros predictores.

La metodología propuesta para reducir el número de predictores, a los cuales se aplicaría el análisis discriminante es la siguiente:

1. Aplicar la transformación en funciones empíricas ortogonales, por separado, a cada magnitud en un nivel dado, es decir, a cada uno de los 25 grupos de 99 parámetros puntuales cada uno. Supongamos, para fijar ideas, que por término medio baste con 5 funciones (nuevas coordenadas) en cada grupo para absorber un alto porcentaje de varianza, y tendríamos así 125 nuevas coordenadas δ'.

2. Aplicamos otra transformación en funciones empíricas ortogonales a estas 100 coordenadas δ', para obtener otras nuevas coordenas δ'', de las que conservamos aquellas que en total absorban el porcentaje de varianza que se estime conveniente. Supongamos que con 25 sea suficiente; pues bien, estos 25 son los parámetros predictores que utilizaremos para aplicar el análisis discriminante, obteniendo los coeficientes previa clasificación de la base empírica en los distintos tipos A_t en que se quiera descomponer el fenómeno a predecir.

Este es el plan de trabajo que estamos desarrollando, tomando como base empírica a cada una de las cuatro estaciones meteoroastronómicas los datos analizados y archivados en el Banco de Datos del Centro Europeo de Predicción a Plazo Medio, de cuatro años completos, 1979 a 1983, o sea unos 365 individuos para cada muestra.

4. Métodos para el cálculo de la mayor raíz E/R

En nuestro anterior artículo [1], vimos que el análisis discriminante exigía la resolución del sistema lineal homogéneo

$$\sum_{k=1}^{N} \lambda_k \left(S^*_{pk} - \frac{E}{R} S_{pk}\right) = 0 \qquad 1 \leqslant p \leqslant N \qquad [31]$$

de N ecuaciones con las N incógnitas λ_k, coeficientes de los respectivos parámetros predictores, en la función

$$Z^{(i)}_t = \sum_{k=1}^{N} \lambda_k X^{(i)}_{kt} \qquad [32]$$

donde S^*_{pk} y S_{pk} se calculan a partir de una muestra empírica de n_t individuos para cada tipo t, con un total de r tipos y $n = \sum_{t=1}^{r} n_t$ individuos muestrales; cada individuo viene dado por sus parámetros $X^{(i)}_{pt}$; $(1 \leqslant p \leqslant N)$; $(1 \leqslant t \leqslant r)$; $(1 \leqslant i \leqslant n_t)$; recordemos sus fórmulas:

$$S^*_{pk} = \sum_{t=1}^{r} n_t(\overline{X}_{pt} - \overline{X}_p) - (\overline{X}_{kt} - \overline{X}_k) \qquad [33]$$

o las simplificadas [29] si los parámetros son funciones empíricas ortogonales.

La condición de compatibilidad del sistema [31] lleva a la ecuación de grado N e incógnita E/R, en forma de determinante de una matriz simétrica de orden N X N:

$$\left| S^*_{pk} - \frac{E}{R} S_{pk} \right| = 0 \qquad \begin{matrix}(1 \leqslant p \leqslant N)\\(1 \leqslant K \leqslant N)\end{matrix} \qquad [34]$$

de la que sólo interesa la mayor raíz.

Para resolver la ecuación [34] podemos usar tres métodos:

1.º) Por ser E y R positivos, todas sus raíces serán positivas y por tanto la mayor será a lo sumo igual a la suma de todas, que es

$$\frac{1}{S_{pk}} \sum_{h=1}^{N} \left| S_{pk} \, (S^*_{ph}) \right| \qquad [35]$$

donde S_{pk} es el determinante de la matriz N × N $\{S_{pk}\}$ y $|S_{pk} \, (S^*_{ph})|$ es el determinante de la matriz obtenida sustituyendo en la anterior $\{S_{pk}\}$ la columna h por sus respectivas S^*_{ph}. Dicho valor [35] será, por tanto, el tope máximo al buscar por tanteos la mayor raíz de [34].

2.º) Hacer el desarrollo polinómico de (34), calculando los 2^N determinantes necesarios, ya que el coeficiente del término de grado g es la suma de los $\binom{N}{N-g}$ determinantes, que se obtiene sustituyendo sucesivamente por combinaciones en el determinante de la matriz $\{S_{pk}\}$ N − g columnas por sus respectivas S^*_{pk}, con signo $(-1)^g$.

3.º) Este problema puede reducirse, ver [2] p. 347, al cálculo del mayor valor propio de la matriz siguiente

[36] $(UD)^T A(UD)$

donde A es la matriz $\{S^*_{pk}\}$, U es la matriz cuyas columnas son los vectores propios de la matriz $\{S_{pk}\}$, ordenados por orden decreciente de sus respectivos valores propios, y D es la matriz diagonal con $d_{ij} = o$ si $i \neq j$ y $d_{ij} = \mu^{-1/2}$ siendo μ los valores propios de $\{S_{pk}\}$ colocados en orden decreciente. (T quiere decir «transpuesta».)

REFERENCIAS

(1) ANSELMO PEINADO SERNA y CARLOS ALMARZA MATA: "Predicción probabilística de fenómenos meteorológicos cuantificados, con una base empírica". REVISTA DE METEOROLOGIA de la Asociación Meteorológica Española, 33-43. Junio, 1983.

(2) FRANZ E. HOHN: "Elementary Matrix Algebra", second edition, The MacMillan Company, New York, 395 pp., 1964.

(3) LORENZ, E.N.: "Empirical Orthogonal Functions and statistical Weather Prediction". Sci Report n.º 1, Statistical Forecasting Proj, Dept. of Meteor, MIT, Cambridge, Massachuset, 44 pp., 1965.

(4) EUROPEAN CENTRE FOR MEDIUM RANGE WEATHER FORECASTING: "Workshop on the use of Empirical Orthogonal Functions in Meteorology", Bracknell, nov. 1977.

(5) PEINADO SERNA, A.: "Comparación de medidas y estimaciones de evaporación". VIII Reunión de la Ponencia de Bioclimatología del CSIC, Zaragoza, mayo 1983.