APLICACIÓN DEL ANÁLISIS *CLUSTER* PARA EL ESTUDIO DE LA RELACIÓN NAO-PRECIPITACIONES DE INVIERNO EN EL SUR DE LA PENÍNSULA IBÉRICA

D. MUÑOZ-DÍAZ y F.S. RODRIGO

Dpto. Física Aplicada, Universidad de Almería.

RESUMEN

En este trabajo se aplica el análisis cluster para analizar la relación entre la Oscilación del Atlántico Norte (NAO) y las precipitaciones invernales en 10 estaciones del Sur de la Península Ibérica. El uso del algoritmo no jerárquico de las K-medias permite obtener 3 clusters, asociados a cada una de las fases de la NAO, que explican el predominio de condiciones secas o húmedas en la región de estudio. El mismo método se aplica a las componentes principales que explican el mayor porcentaje de varianza de los datos originales. En este caso, los resultados ofrecen una regionalización similar, con una clara diferencia entre las estaciones occidentales y las orientales, pero con una mejor definición de las tres fases de la NAO.

Palabras clave: Análisis cluster, NAO, precipitaciones de invierno, Sur de la Península Ibérica.

ABSTRACT

In this work cluster analysis is applied to analyse the relationship between the North Atlantic Oscillation (NAO) and winter rainfall of 10 localities in southern Iberian Peninsula. Using the non-hierarchical K-means algorithm, 3 clusters have been obtained, corresponding to each NAO phase, that explain the predominance of wet and dry conditions in the study region. The method is also applied to the principal components explaining the greater percentage of variance of the original data. In this case, results show a similar regionalization, with a clear difference between western and eastern stations, but with better definition of the three NAO phases.

Key words: *Cluster analysis, NAO, winter rainfall, southern Iberian Peninsula.*

1. INTRODUCCIÓN

En la investigación geofísica, se necesita con frecuencia clasificar las variables en grupos homogéneos e identificar sus características con el fin de profundizar en la comprensión de los fenómenos y predecir potenciales cambios futuros. El análisis de conglomerados o análisis "cluster" (AC) es una de las muchas técnicas multivariantes que pueden usarse para obtener tales clasificaciones. Dentro del campo de la investigación climatológica, ha sido reconocido como un método eficaz para tratar problemas como el agrupamiento de estaciones en regiones climatológicamente homogéneas (GONG y RICHMAN, 1995), o el agrupamiento de periodos de tiempo (días, años, etc) en clusters que reflejan la ocurrencia de determinados sucesos o patrones meteorológicos (RAMOS, 2001).

El principal objetivo de este trabajo es analizar la relación entre la Oscilación del Atlántico Norte (NAO) y el régimen de precipitaciones de invierno en el sur de la Península Ibérica, utilizando

técnicas de AC. En los últimos años ha recibido un creciente interés el estudio sobre la NAO (por ejemplo, MARSHALL *et al.*, 2001; WANNER *et al.*, 2001), dada su inequívoca influencia en el clima del continente europeo. El modo NAO se caracteriza por un patrón dipolar de anomalías de presión de signo opuesto, con centros situados en un eje Norte-Sur sobre el Atlántico Norte, aunque exhibe una considerable oscilación estacional. Debido al mayor vigor de la circulación atmosférica en invierno, en coincidencia con el mayor gradiente de temperaturas del Ecuador a los Polos, el patrón NAO es mucho más pronunciado durante el invierno, tanto en su intensidad como en el área cubierta. La NAO presenta una acusada variabilidad temporal en escalas estacional, anual, e interanual. Dicha variabilidad se manifiesta por la existencia de dos fases opuestas: una fase positiva, caracterizada por presiones y alturas geopotenciales por encima de las normales en el centro sur y por debajo de las normales en el centro norte; y una fase negativa, caracterizada por presiones y alturas geopotenciales por debajo de las normales en el centro sur y por encima de las normales en el centro norte. Ambas fases de la NAO están asociadas con situaciones climáticas distintas en su zona de influencia, fundamentalmente la región del Atlántico Norte. En un trabajo anterior (MUÑOZ-DÍAZ y RODRIGO, 2001) se analizó la relación entre la NAO y las precipitaciones mensuales en San Fernando (Cádiz), estableciendo un simple modelo lineal. Para los meses de invierno, el coeficiente de correlación lineal obtenido, estadísticamente significativo, era negativo y del orden de -0.60. Sin embargo, los coeficientes de correlación pueden ser sensibles a hipótesis subyacentes acerca de la distribución de probabilidad de los datos, y pueden resultar inapropiados para definir la intensidad de la señal NAO en áreas con distribuciones de probabilidad sesgadas o no normales (MASON y GODDARD, 2001). Este es el caso de las series de precipitaciones totales mensuales en la Península Ibérica, que quedan mejor representadas por funciones de distribución gamma (LANA y BURGUEÑO, 2000). El AC no es una técnica de inferencia estadística, sino un método objetivo de cuantificar las características estructurales de un conjunto de observaciones y, como tal, tiene propiedades matemáticas, pero no estadísticas. Por tanto, los requisitos de normalidad y homocedasticidad, importantes en otras técnicas multivariantes, no se aplican en el AC (MARTÍNEZ ARIAS, 1999). Ello lo convierte en especialmente interesante para el estudio de las series de precipitaciones, y, en particular, de su relación con la Oscilación del Atlántico Norte.

2. DATOS

Puede obtenerse de forma simple un índice significativo de la NAO a través de la diferencia entre los valores de presión al nivel de mar de dos estaciones representativas del centro sur y el centro norte del dipolo. Los datos de presión disponibles y con los cuales se puede construir un índice NAO corresponden, en la parte sur del dipolo, a Gibraltar (36.1ºN,5.4ºW), mientras que en la parte Norte del dipolo se dispone de datos de Reykjavik y Stykkisholmur (65.0ºN, 22.8ºW) en Islandia (JONES *et al.*, 1997; POZO-VÁZQUEZ *et al.*, 2000). Para la elaboración del índice NAO mensual se utilizaron datos de presión mensual en Islandia y Gibraltar desde 1909 hasta 1997. Los datos están disponibles en varias direcciones de Internet, como por ejemplo en http://www.cru.uea.ac.uk/cru/data/nao.htm. Debido a las diferentes características estadísticas de los datos de presión en las estaciones sur y norte (sobre todo con respecto a la desviación típica), es necesario un proceso de normalización para cada serie temporal de presión antes de construir el

Tabla 1: Percentiles (mm) de las series de precipitación, 1961-1990.

Localidad	P10	P25	P75	P90
Huelva	69.8	136.3	204.0	304.1
Badajoz	64.7	115.3	164.3	236.7
San Fernando	111.6	179.3	230.9	322.9
Sevilla	73.0	160.4	237.9	366.7
Córdoba	92.7	131.0	205.4	335.8
Málaga	50.1	150.8	215.8	330.8
Jaén	92.2	128.1	199.1	285.8
Granada	53.7	99.3	129.3	184.6
Almería	25.5	42.6	67.1	90.7
Murcia	25.2	32.6	57.1	84.6

índice. El proceso consiste en calcular, para cada estación meteorológica, la diferencia entre el valor mensual de la presión y el valor medio de un periodo de referencia, y dividir el valor resultante por la desviación típica del periodo de referencia. Como periodo de referencia se eligió el periodo 1961-1990. El índice NAO mensual se obtiene como la diferencia entre los valores normalizados de la presión en Gibraltar e Islandia. Finalmente, el índice estacional se establece como el promedio de los tres índices mensuales correspondientes. Para el periodo 1909-1997, el valor medio del índice NAO obtenido fue de 0.23 y la desviación típica 1.10.

Los datos de precipitaciones utilizados en este estudio consisten en totales de precipitación estacional de 10 localidades del sur de la Península Ibérica (Huelva, Badajoz, San Fernando, Sevilla, Córdoba, Málaga, Jaén Granada, Almería y Murcia), todas ellas situadas en torno a los 37ºN de latitud, con la más occidental (Huelva) situada a 6º56'W y la más oriental (Murcia) situada a 1º13'W. Las altitudes respecto al nivel del mar oscilan entre los 7 m de Málaga y los 680 m de Granada. El periodo de medida, común a todas ellas, es 1909-1997. ESTEBAN-PARRA *et al.* (1998) analizaron la homogeneidad de estas series mediante la aplicación de tests estadísticos de homogeneidad absoluta y relativa, obteniendo como resultado que las series en estudio reúnen suficientes criterios de calidad como para ser utilizadas en análisis de variabilidad climática. El criterio para caracterizar los inviernos como secos o húmedos es considerar los percentiles 10, 25, 75 y 90 como valores umbral. Cuando las precipitaciones sean inferiores a los percentiles 10 y 25 se hablará de invierno muy seco y seco, respectivamente, y cuando sean superiores a los percentiles 75 y 90 de invierno húmedo o muy húmedo. La tabla 1 muestra los valores de los percentiles (en mm) para cada una de las series utilizadas durante el periodo de referencia 1961-1990.

3. ANÁLISIS CLUSTER DE LOS DATOS ORIGINALES

El propósito del AC es situar objetos en grupos sugeridos por los datos, y no definidos a priori, de modo que los objetos pertenecientes a un conglomerado o cluster tienden a ser similares, mientras que objetos en clusters distintos tienden a ser diferentes. Hay dos técnicas principales: divisivas

y jerárquicas. En la técnica jerárquica los objetos son progresivamente agregados hasta que se juntan en un solo cluster. El objetivo de la técnica divisiva es separar un conjunto de objetos en grupos consistentes, de modo que cada objeto se sitúa en uno y sólo en un cluster. La asignación preliminar de los objetos a un cluster puede hacerse usando una partición aleatoria, y luego los objetos se transfieren de un cluster a otro hasta alcanzar la posición en la cual la similitud es mayor.

En el AC, los datos iniciales consisten de una matriz X de pxn observaciones, donde X puede interpretarse como un conjunto de n puntos en un espacio p-dimensional. A los vectores columna se les denomina variables y a las filas observaciones. En nuestro caso tenemos n=11 variables (valores del índice NAO y de las 10 series de precipitaciones invernales) y p=89 observaciones, correspondientes a cada uno de los inviernos del periodo 1909-1997. Así, x_{ij} representa el valor para la variable j-ésima y la observación i-ésima. El primer paso es establecer una medida numérica de la similaridad o disimilaridad entre las variables para caracterizar las relaciones entre los datos. La medida de disimilaridad más utilizada en el AC es la distancia euclídea (GONG y RICHMAN, 1995). Dada la matriz X, la distancia euclídea entre las variables i y j viene dada por

$$d_{ij} = [(X_i - X_j)^T (X_i - X_j)]^{1/2}$$

donde el símbolo ^{T} indica matriz traspuesta. Trasponiendo la matriz de datos original el mismo método puede aplicarse al conjunto de observaciones. Las observaciones que tienen diferentes escalas de medida o valores medios (como es nuestro caso) pueden contribuir desigualmente a la distancia calculada. En general, las variables con mayor variabilidad tienen más impacto en la medida de similaridad. Por ello, las variables suelen estandarizarse (es decir, se transforma la matriz de datos X hallando para cada dato la anomalía respecto al valor medio de la variable, normalizada por su desviación típica) antes de calcular la distancia euclídea, para eliminar así este efecto de escala.

La idea central en la mayoría de los métodos no jerárquicos es elegir una partición inicial de los datos y luego alterar las pertenencias a los clusters para así obtener una nueva partición que revele una estructura determinada en los datos de entrada. Los métodos no-jerárquicos se diseñan para agrupar variables de datos en una sola clasificación de k clusters, donde k se especifica *a priori*. En estos métodos se utiliza un conjunto de puntos "semilla" como los centroides (o vectores de valores medios) de los clusters, y se construye el conjunto inicial de clusters asignando cada variable u observación al cluster con el centroide más próximo. Posteriormente se recalculan los centroides y las distancias de las observaciones, reasignando las observaciones a los nuevos clusters. Estos procesos se aplican de forma iterada hasta que convergen a una configuración estable. En general, las técnicas no jerárquicas ofrecen mejores resultados que las jerárquicas, y el método no jerárquico más comúnmente empleado es el denominado de K-medias (GONG y RICHMAN, 1995), que utiliza la distancia euclídea como medida de similaridad. No existen procedimientos óptimos para determinar a priori el número k de clusters (MARTÍNEZ ARIAS, 1999). En nuestro caso se eligió k=3, basándose en el hecho de que tres son las fases principales que caracterizan el comportamiento del índice NAO (negativa, normal y positiva). La elección de los puntos semilla se realizó de modo secuencial, usando las primeras k variables de datos de la matriz X. La tabla 2 resume los centroides encontrados en el análisis de los datos originales.

Tabla 2: Centroides del AC (método de K-medias y distancia euclídea).

Variable	Cluster1	Cluster2	Cluster3
NAO	0.05	0.92	-1.09
Huelva	213.1	125.7	338.0
Badajoz	181.3	111.5	296.3
San Fernando	262.0	133.4	415.8
Sevilla	247.5	119.4	411.5
Córdoba	253.5	147.3	428.1
Málaga	207.1	131.8	427.9
Jaén	221.6	144.3	372.0
Granada	142.2	97.6	229.8
Almería	69.7	60.8	113.1
Murcia	65.7	71.2	68.6

El Cluster 1 contiene el 34.83 % de las observaciones, el cluster 2 el 46.07 % y el cluster 3 el 19.10 %. La interpretación física de los clusters obtenidos es inmediata. El cluster 1 corresponde a una situación de índice NAO normal (~ 0), el cluster 2 corresponde a la fase positiva de la oscilación (+0.92), mientras que el cluster 3 corresponde a la fase negativa de la NAO (-1.09). La aplicación del test F y del test de diferencias significativas de Fisher lleva a la conclusión de que los tres valores medios son significativamente diferentes a un nivel de confianza del 95 %. Esto se traduce en un valor para las precipitaciones en el cluster 2 inferior al primer cuartil para los casos de Huelva, Badajoz, San Fernando, Sevilla y Málaga, y superior al percentil 90 en el caso del cluster 3 en todas las estaciones, excepto en Murcia (véase la tabla 1).

La figura 1 muestra la distribución de las observaciones en los tres clusters para los casos de San Fernando (a), Granada (b) y Murcia (c). Excepto en el caso de San Fernando, se encuentra que las fronteras de los clusters no pueden definirse de forma separada. Esto suele ocurrir cuando se trabaja con casos reales. En nuestro caso, el mayor solapamiento entre los clusters correspondientes a la fase positiva y la fase normal de la NAO en Granada sugiere que, a efectos de precipitaciones en la región de estudio, la distinción entre ambas situaciones no es tan clara en comparación con el comportamiento de la fase negativa. Esta incertidumbre sugiere o bien la necesidad de incorporar otros factores causales al análisis (otros mecanismos circulatorios distintos a la NAO, influencia del Mediterráneo, etc.), o bien la conveniencia de introducir algoritmos que la tengan en cuenta, como aquellos que usan la lógica difusa en su procedimiento (HALKIDI *et al.*, 2001). En general, se aprecia una clara regionalización, con la estación de Murcia claramente separada del resto, indicando que la NAO apenas influye en las precipitaciones de esta estación. La figura 2 muestra los diagramas de Box-and-Whisker correspondientes a cada cluster para las tres estaciones seleccionadas en la figura 1. La comparación de los valores medios resulta en diferencias estadísticamente significativas a un nivel de confianza del 95 % para las estaciones de San Fernando y Granada, mientras que este resultado no se obtiene en el caso de Murcia.

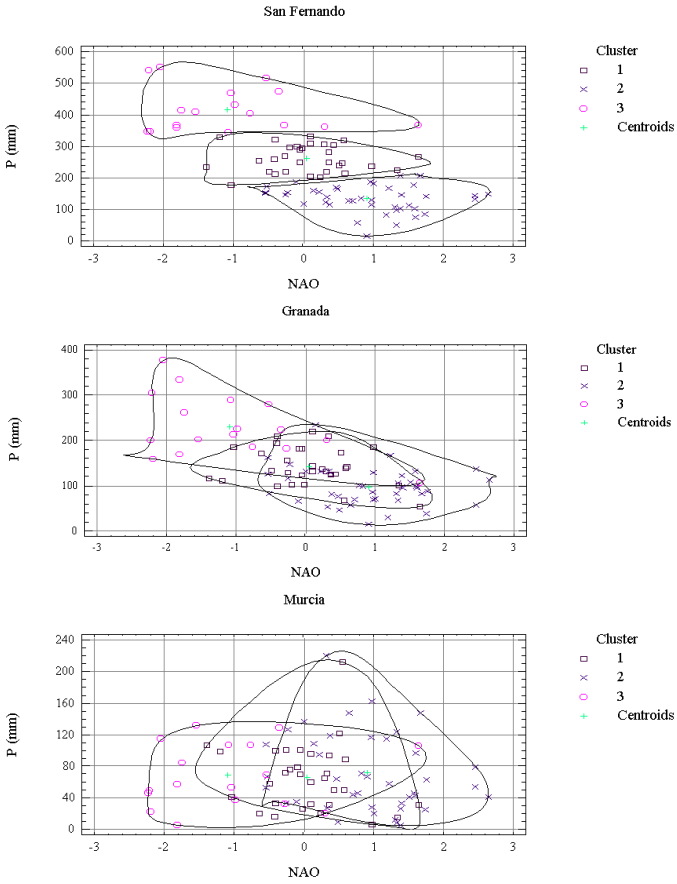

Figura 1: Diagramas de dispersión (método de K-medias, distancia euclídea), para los casos de San Fernando (a), Granada (b) y Murcia (c).

4. COMPONENTES PRINCIPALES

Un problema importante en el AC es el de la multicolinealidad, es decir, la presencia de variables muy correlacionadas puede tener el efecto de atribuir un peso excesivo a algún aspecto de las variables. Así, si usamos tres variables muy correlacionadas, proporcionan la misma información que una sola, aunque tienen un peso triple en el cálculo de la similaridad. (MARTÍNEZ ARIAS, 1999). Surge entonces la necesidad de utilizar alguna técnica de reducción de datos, como el análisis de componentes principales (ACP). El método de las componentes principales es una alternativa a los métodos tradicionales de AC para conseguir grupos homogéneos. En este método, las distintas variables se asignan a grupos según los valores de sus factores de carga. Mientras que el AC tradicional incluye la varianza completa de datos originales, el ACP permite separar la señal buscada del ruido de fondo. Para ello se procede mediante la selección de las componentes

Tabla 3: Factores de carga de las series de precipitaciones del Sur de la Península Ibérica, 1909-1997.

Serie	Hu	Ba	SF	Se	Co	Ma	Ja	Gr	Al	Mu
PC1	0.84	0.86	0.90	0.94	0.93	0.82	0.87	0.88	0.50	-0.16
PC2	0.14	-0.06	0.18	0.04	-0.06	0.27	-0.12	0.11	0.57	0.88

Tabla 4: Centroides del AC (método de K-medias y distancia euclídea aplicado a las dos primeras componentes principales).

Variable	Cluster1	Cluster2	Cluster3
NAO	-0.01	1.38	-1.45
PC1	-0.4	-4.5	10.5
PC2	0.1	-0.9	1.5

principales que representan el mayor porcentaje de varianza explicada, mediante un procedimiento de rotación, como el método Varimax (VON STORCH y ZWIERS, 1999). La tabla 3 muestra los valores de los factores de carga correspondientes a los dos primeras componentes principales obtenidos para cada serie de precipitación mediante la aplicación del ACP y la rotación Varimax. La primera componente explica el 65.5 % y la segunda el 12.1 % de la varianza total de los datos, siendo la varianza total acumulada explicada por ambas componentes el 77.6 %. Los valores de los factores de carga indican una clara regionalización, con Murcia como principal estación de la segunda componente, y Badajoz así como la mayoría de las estaciones andaluzas incluidas en la primera. A diferencia de los métodos AC, las soluciones aquí pueden ser solapantes, es decir, algunas variables pueden estar incluidas en más de un cluster (GONG y RICHMAN, 1995). Este es el caso de la estación de Almería, con factores de carga 0.50 y 0.57 para, respectivamente, la primera y la segunda componente. Podemos aplicar ahora nuevamente el método de las K-medias para analizar la dependencia NAO-precipitaciones, usando las series de componentes principales PC1 y PC2 como datos de entrada. Utilizando nuevamente k=3, los valores de los centroides vienen recogidos en la tabla 4.

Igual que en el caso anterior, el cluster 1 (48.31 % de las observaciones) corresponde a la fase NAO normal, el cluster 2 (34.83 % de las observaciones) corresponde a la fase positiva y el cluster 3 (16.85 % de las observaciones) a la fase negativa de la NAO. Los valores de las componentes principales indican anomalías negativas para la primera componente y el cluster 2 y anomalías fuertemente positivas para esta componente en el caso del cluster 3. Las diferencias entre clusters no son significativas en el caso de la segunda componente, correspondiente a Murcia y, en menor medida, Almería. La figura 3 muestra los diagramas de dispersión correspondientes a ambas componentes. El ACP permite obtener clusters con fronteras más definidas, y, en consecuencia, reducir la incertidumbre obtenida en el análisis anterior. Mientras que para la primera componente

se obtiene una clara relación negativa (coeficiente de correlación -0.69), en el caso de la segunda componente los clusters se obtienen en función de los valores del índice NAO, sin una clara influencia de los valores de las precipitaciones, indicando así la escasa dependencia entre ambas variables.

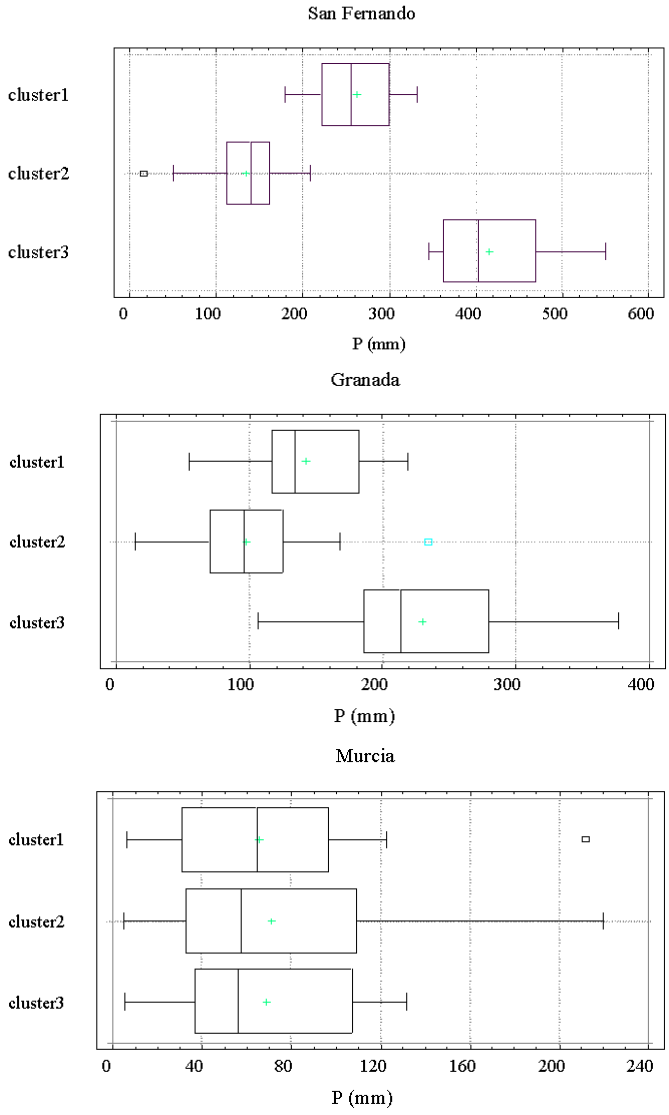

Figura 2: Diagramas de Box-Whisker de las distribuciones resultantes en los tres clusters para los casos de San Fernando (a), Granada (b) y Murcia (c).

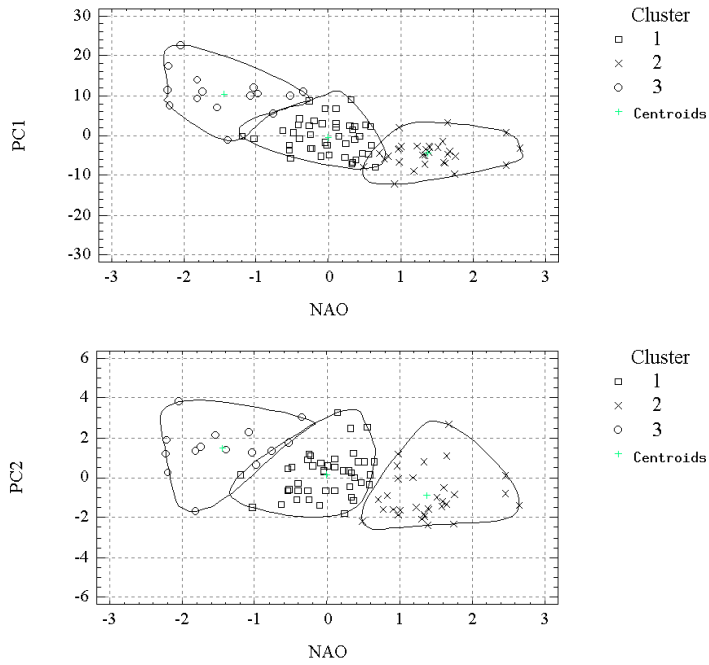

Figura 3: Diagramas de dispersión (método de K-medias, distancia euclídea) para las dos primeras componentes principales rotadas PC1 (a) y PC2 (b).

5. CONCLUSIONES

El análisis cluster ha demostrado ser útil para el estudio de las relaciones entre el comportamiento del índice NAO y las series de precipitaciones del Sur de la Península Ibérica, con la obtención de 3 clusters correspondientes a cada una de las fases NAO descritas en la bibliografía. Sin embargo, la aplicación del método de las K-medias a las series de datos originales ofrece un cierto grado de incertidumbre, con fronteras mal definidas entre los clusters correspondientes a la fase normal y la fase positiva. Ello puede indicar una menor diferencia entre ambas fases en cuanto a su influencia en el régimen de precipitaciones. Para evitar la posible influencia de las correlaciones entre los datos originales se ha aplicado el análisis a las primeras componentes principales, que explican un 77.6 % de la varianza total de los datos. Los resultados determinan una regionalización similar a la anterior, pero permiten una mejor definición de los clusters.

6. AGRADECIMIENTOS

Este trabajo se desarrolló con financiación del Ministerio de Ciencia y Tecnología, Proyecto REN2001-3923-C02-02/CLI.

7. REFERENCIAS

ESTEBAN-PARRA, M.J., RODRIGO F.S. y CASTRO-DÍEZ, Y. (1998): Spatial and temporal patterns of precipitation in Spain for the period 1880-1992.*International Journal of Climatology*, 18, pp. 1557-1574.

GONG, X. y RICHMAN, M.B. (1995): On the application of cluster analysis to growing season precipitation in North America East of the Rockies, *Journal of Climate*, 8, pp. 897-931.

HALKIDI, M., BATISKAKIS, Y. y VAZIRGIANNIS, M. (2001): On clustering validation techniques, *Journal of Intelligent Informations Systems*, 17, pp. 107-145.

JONES, P.D., JONSSON, T. and WHEELER, D. (1997): Extension to the North Atlantic Oscillation index using early instrumental pressure observations from Gibraltar and south-west Iceland, *International Journal of Climatology*, 17, pp. 1-18.

LANA, X. y BURGUEÑO, A. (2000): Some statistical characteristics of monthly and annual pluviometric irregularity for the Spanish Mediterranean Coast. *Theoretical and Applied Climatology*, 65, pp. 79-97.

MARSHALL, J., KUSHNIR, Y., BATTISTI, D., CHANG, P., CZAJA, A., DICKSON, R., HURRELL, J., McCARTNEY, M., SARAVANAN, R. y VISBECK, M. (2001): North Atlantic climate variability: phenomena, impacts and mechanisms, *International Journal of Climatology*, 21, pp. 1863-1898.

MARTÍNEZ ARIAS, R. (1999): *El análisis multivariante en la investigación científica*. Editorial La Muralla, S.A., Madrid, 143 pp.

MASON S.J. y GODDARD, L. (2001): Probabilistic precipitation anomalies associated with ENSO. *Bulletin of the American Meteorological Society*, 82, pp. 619-638.

MUÑOZ-DÍAZ, D. y RODRIGO, F.S. (2001): Fases extremas del índice NAO mensual y precipitaciones en el SW de la Península Ibérica. En PÉREZ-CUEVA, A.J., LÓPEZ BAEZA, E. y TAMAYO CARMONA, J. (Eds): *El tiempo del clima*. Publicaciones de la Asociación Española de Climatología (AEC, serie A, nº2), pp. 177-186.

POZO-VÁZQUEZ, D., ESTEBAN-PARRA, M.J., RODRIGO, F.S. y CASTRO-DÍEZ, Y. (2000): An analysis of the variability of the North Atlantic Oscillation in the time and frequency domains. *International Journal of Climatology*, 20, pp. 1675-1692.

RAMOS, M.C. (2001): Divisive and hierarchical clustering techniques to analyse variability of rainfall distribution patterns in a Mediterranean region, *Atmospheric Research*, 57, pp. 123-138.

VON STORCH, H. y ZWIERS, F.W. (1999): *Statistical analysis in climate research*. Cambridge University Press, Cambridge, 484 pp.

WANNER, H., BRÖNNIMANN, S., CASTY. C., GYALISTRAS, D., LUTERBACHER, J., SCHMUTZ, C., STEPHENSON, D.B. y XOPLAKI, E. (2001): North Atlantic Oscillation-Concepts and Studies. *Surveys in Geophysics*, 22, pp. 321-382.