GEOACTA,vol.11,n.1 (octubre de 1981) pag 179 a 190

ESTIMACION DEL RUIDO EN ARCHIVOS DE DATOS AEROLOGICOS UTILIZANDO FUNCIONES ORTOGONALES EMPIRICAS

María Luz D. de Lloret y Gustavo V. Necco*
Departamento de Meteorología, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires.
Buenos Aires, República Argentina

RESUMEN

Uno de los principales problemas que se enfrentan al utilizar una base de datos es la estimación del ruido. Una metodología posible consiste en el estudio del comportamiento de las Funciones Ortogonales Empíricas (FOE) de la muestra, particularmente a través del valor de diagnóstico que se evidencia en las variaciones del logaritmo de los autovalores asociados en función del número de los mismos. Este método, originado por Craddock (1965), fue aplicado al campo de temperaturas del archivo de datos aerológicos de la República Argentina que posee el Departamento de Meteorología de la Facultad de Ciencias Exactas y Naturales (FCEYN), facilitado por el Servicio Meteorológico Nacional (SMN).

Se determinó que dicho ruido, en el estado actual del archivo, afecta en algunos casos hasta el 30% de la varianza de la información (referida al estado medio), aunque dicho porcentaje varía según la estación aerológica y el mes del año.

ABSTRACT

One of the major problems faced for in the use of data bases es noise estimation.

A possible methodology involves the analisys of the sample Empirical Orthogonal Functions behaviour, in particular through the diagnostic value evidenced by the variationsof the logarithm of eigenvalues as functions of its caractheristics numbers. This methods, proposed by Craddock (1965), was applied to temperature series taken from Argentine aerological files supplied to the Departament of Meteorology of the Natural and Exact Sciences Faculty (FCEyN) by the National Weather Service (S.M.N.).

Is was found that in the present state of the files such noice affects, in some cases, up to 30% of the information variance (referred to the mean state) although this value varies with site and season.

*Miembro de la Carrera de Investigador Científico del CONICET.
Jefe del Instituto de Investigaciones Sinópticas del Servicio Meteorológico Nacional.

1. INTRODUCCION

Si se desean utilizar funciones ortogonales empíricas con el objeto de compactar la información contenida en grandes volúmenes de datos meteorológicos para su posterior aplicación en pronósticos objetivos, surge el interrogante de la determinación de la cantidad de autovectores que es posible desechar sin perder información del campo meteorológico real.

Muy pocos autores se han dedicado a analizar este problema en el campo de las aplicaciones meteorológicas: Juahni Rinne y Simo Järvenoja (1979) han realizado una extensa discusión sobre el tema, utilizando distintos métodos de truncado que aplicaron a funciones obtenidas de los análisis de 500mb del Hemisferio Norte. De la literatura consultada, el único método originado de consideraciones meteorológicas es el planteado por Craddock y Flood (1969) y analizado posteriormente por Farmer (1971).

Este último método, basado en el comportamiento de los autovalores de campos aleatorios y que se explica en el siguiente apartado, es aplicado en este trabajo al campo térmico de los sondeos aerológicos de estaciones argentinas.

El Departamento de Meteorología de la Facultad de Ciencias Exáctas y Naturales de la Universidad de Buenos Aires posee archivados en cintas magnéticas la información aerológica de los niveles estándar, correspondiente al período 1958-1971, de trece estaciones argentinas, facilitada por el Servicio Meteorológico Nacional.

Uno de los objetivos del proyecto de investigación "Tratamiento estadístico de datos aerológicos de la República Argentina", de dicho Departamento, consiste en lograr una descripción compacta de la información, a través de distintas herramientas estadísticas y a su vez obtener un mejor conocimiento del comportamiento de las distintas variables atmosféricas a ser utilizadas en pronósticos objetivos.

Ello se ha encarado, por una parte, a través de la determinación de funciones ortogonales empíricas de dichas variables, las cuales también pueden ser utilizadas, en algunos casos, como interpoladores.

En primer término se decidió analizar el comportamiento del campo térmico, determinándose los autovalores y autovectores de los desvíos de las temperaturas de los sondeos diarios respecto del sondeo medio muestral de la estación considerada, ya que en un estudio piloto realizado previamente (Lloret y Necco, 1979) se vió la necesidad de eliminar de la información inicial la estratificación normal atmosférica pues en la misma está contenida aproximadamente el 90% de la varianza del campo de temperaturas.

Se determinaron conjuntos de autovalores y autovectores para cada uno de los doce meses del año en forma independiente con muestras correspondientes a las estaciones aerológicas Ezeiza, Espora y Resistencia separadamente debido a que la memoria total de la computadora utilizada es insuficiente para trabajar con matrices de mayor dimensión. Cada muestra consistió en las temperaturas de todos los sondeos de la estación y mes considerados que tienen información en todos los niveles estándar comprendidos entre 1000 y 100 mb incluidos ambos. Una vez determinados los conjuntos de autovalores y autovectores surgió la necesidad de establecer cuántos de estos últimos se debían utilizar para conservar un porcentaje óptimo de la varianza del campo real de desvíos de la temperatura, desechando la varianza del campo inicial debida a procesos aleatorios (ruido). Para ello se utilizó un criterio debido a Farmer (1971).

2. METODOLOGIA DE LA ESTIMACION

Sea una matriz P tal que cada elemento p_{ij} corresponda al desvío de la temperatura en el nivel i el día j. La misma se puede expresar en forma de un producto matricial:

$$P = MV \tag{1}$$

donde V es una matriz tal que satisface las siguientes condiciones:

$$VV' = I \tag{2}$$

$$VAV' = D \tag{3}$$

aquí A es la matriz de covarianzas definida por:

$$A = P^{*\prime} P^{*}$$

((') indica matriz traspuesta y (*) son desvíos respecto de la media muestral), I es la matriz identidad y D es una matriz cuyos elementos no diagonales son nulos.

De acuerdo con (2) y (3), es posible determinar M tal que cumpla con (1) y tal que:

$$M = PV' \tag{4}$$

$$M^{*\prime}M^{*} = (VP^{*\prime})\,(P^{*}V') = D \tag{5}$$

Por lo anterior la matriz V contiene los autovectores de la matriz de covarianzas A y los elementos de la diagonal de D son sus autovalores, denominados generalmente λ. Dado que A es una matriz simétrica y los autovectores $\vec{V}_i$ normales, los autovalores λ seran reales. De tal forma cualquier sondeo se puede expresar como:

$$\widetilde{P_j} = \sum_{i=1}^{N} mij\, V$$

donde los N autovectores constituyen una base ortogonal del espacio muestral y por lo tanto tienen la propiedad de no estar correlacionados entre sí. Por otra parte el porcentaje de varianza explicado por el k-ésimo autovector es:

$$S_k = \lambda_k \Big/ \sum_{i=1}^{N} \lambda_i$$

En este caso el número de orden k se encuentra relacionado a la escala característica del fenómeno representado por el término correspondiente.

Farmer (1971) ha mostrado que si se tiene un conjunto de k autovalores (λ_k) ordenados en forma decreciente y se grafica el logaritmo de los autovalores en función de su número de orden (llamado en la literatura inglesa, diagrama LEV) se encuentra que dicho gráfico tiene un comportamiento característico tal como se indica en la figura 1.

La curva tiene para los autovalores de orden menor una forma de tipo exponencial, mientras que los de mayor orden se acercan a una recta. La parte lineal del diagrama LEV corresponde a los autovalores asociados a autovectores

ligados a la parte azarosa de la información original, mientras que la porción relacionada a los autovalores de menor orden corresponde a los autovectores asociados a patrones de escala mayor, que representan un mayor porcentaje de la varianza original y pueden estar ligados a procesos físicos.

De esta manera es posible cuantificar el porcentaje de varianza correspondiente a los errores aleatorios en la muestra.

3. RESULTADOS

En primer término, mediante un programa computacional fueron generados números al azar con una distribución uniforme, que fueron dispuestos en matrices de dimensión 600x13, con el objeto de simular las temperaturas en los trece niveles estándar entre 1000 y 100 mb, correspondientes a 600 radiosondeos (ésta era la dimensión de las diversas muestras utilizadas con información de las distintas estaciones aerológicas). Se hallaron los autovalores y autovectores de dicha matriz y se graficó el diagrama LEV de los primeros (Figura 2, con puntos). El primer autovector explica el 75.9% de la varianza total del campo generado al azar y los doce restantes explican entre el 2.7% y el 1.4%.

Estos últimos se encuentran sobre una recta mientras que el primero se aleja considerablemente de la misma indicando que si bien los doce últimos autovectores muestran que los datos generados corresponden a información al azar, el primer autovector no. En la Figura 3 se ha graficado el primer autovector, que toma un valor constante en la vertical, lo cual coincide con el valor medio del campo al azar, trazado en la misma figura.

Se hallaron entonces los autovalores y autovectores de los desvíos de los datos generados al azar respecto del valor medio correspondiente a cada nivel. El diagrama LEV de los autovalores obtenidos en este caso se ve en la Figura 2 (con cruces). El efecto dominante del primer autovector ha desaparecido y los nuevos trece autovectores explican un porcentaje de varianza que va desde el 10.7% para el primero hasta el 5.2% para el último. En este caso todos los puntos se acercan mucho a la recta de regresión obtenida anteriormente indicando la total azarosidad de la información analizada.

Se obtuvieron resultados totalmente coincidentes con lo antes expuesto con aproximadamente diez muestras diferentes obtenidas al azar, verificandose, de esta manera, lo establecido por Farmer.

De estos resultados es evidente que si se toma una muestra de información meteorológica tranzando el diagrama LEV de los autovalores determinados a partir de la misma se podrán considerar como debidos a procesos aleatorios los autovectores asociados a los autovalores de la parte de la curva que se ajusten a una recta.

En la figura 4, se puede ver el diagrama LEV de los autovalores de los desvíos de las temperaturas diarias de Ezeiza, donde se han ploteado los valores obtenidos para cada uno de los doce meses del año en un mismo gráfico.

Cuando los mismos se observan en forma individual, en algunos meses el punto de truncado no está bien definido y parecen no estar de acuerdo con los resultados obtenidos para el resto del año.

Al analizar los valores que toman en dichos casos los coeficientes multiplicadores mij de la ecuación (4) se nota un salto en el valor absoluto de uno o más de los de orden intermedio con respecto a su rango normal de variación. Recurriendo a la información de entrada y analizando el sondeo correspondiente se encuentra que en él hay capas con gradientes superadiabáticos o superinversiones lo cual aumenta el porcentaje de varianza explicado por los autova-

lores antes mencionados. Cabe destacar que la información básica utilizada solo había sido consistida parcialmente por rangos (Velasco y Necco, 1980). Estos resultados nos indican que las FOE son también una herramienta muy útil para detectar información dudosa en una base de datos, identificando los posibles errores.

En las figuras 5 y 6 se han trazado los diagramas LEV correspondientes a Resistencia y Espora respectivamente. Tanto en Ezeiza (Figura 4) como en Espora y Resistencia, a partir del cuarto autovector, según el criterio de Farmer, la varianza explicada correspondería al ruido (comportamiento aleatorio) existente en la base de datos.

En la tabla I se presentan los porcentajes de varianza acumulada explicada por los diez últimos autovectores para cada una de las treinta y seis muestras analizadas. Estas varianzas son indicativas del ruido blanco existente en la base de datos analizada que varía según sea la estación y el mes considerados. Los valores obtenidos alcanzan al 31% para la muestra correspondiente al mes de enero de la estación aerológica Resistencia, que de las tres analizadas es la que más ruido presenta. La estación Espora presenta el ruido más parejo en todas las muestras analizadas.

4. CONCLUSIONES

Mediante una metodología basada en el comportamiento de los autovalores de la matriz de covarianzas se ha mostrado la utilidad de las funciones ortogonales empíricas como identificadores de la parte aleatoria en un muestreo de datos meteorológicos, así como de los errores presentes.

La aplicación del método a algunas estaciones de la base de datos aerológicos de la República Argentina que posee el Departamento de Meteorología indica que aproximadamente un 90% de la varianza de la información está dada por la estratificación atmosférica, y que del 10% restante alrededor de un 2 a 3% corresponde a procesos aleatorios.

Agradecimientos: Los autores agradecen a las autoridades del Servicio Meteorológico Nacional por la información facilitada; al personal del Instituto de Cálculo de la Facultad de Ciencias Exactas y Naturales (UBA) por el desarrollo de programas computacionales y por posibilitar el procesamiento de los mismos y a la Srta. Gilda Mercado por su colaboración técnica y el mecanografiado del trabajo.

Este trabajo contó con el apoyo económico de la Secretaría de Estado de Ciencia y Tecnología a través de los subsidios 42°875/77 y 15466/79 y del Consejo de Investigaciones Científicas y Técnicas a través del subsidio 8773/79.

BIBLIOGRAFIA

Craddock, J. M., 1965: A meteorological application of principal component analysis. The Statistician, Vol. 15, No. 2.

Craddock y Flood, 1969: Eigenvectors for representing the 500 mb geopotential surface over the Northern Hemisphere. Quarterly Journal of the Royal Meteorological Society, Vol. 95.

Farmer, S. A., 1971: An investigation into the results of Principal Component Analysis of data derived from random numbers. The Statistician, Vol. 20, No. 4

Juhani Rinne y Sino Jarvenoja, 1979: Truncation of the EOF series representing 500 mb leights. Quarterly Journal of the Royal Met. Soc., Vol. 105.

Lloret y Necco, 1979: Resultados preliminares de la aplicación de funciones ortogonales empíricas a radiosondeos de la República Argentina. Meteorológica, Vol. X, No. 2.

Velasco y Necco, 1980: Valores medios extremos y desviaciones estándar de datos aerológicos de la República Argentina. Publicación del Departamento de Meteorología. FCEyN. UBA. Buenos Aires.

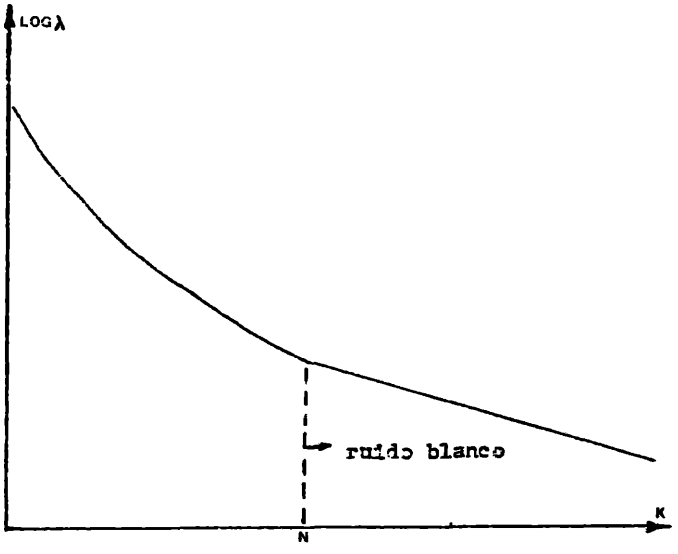

Fig.1 :Diagrama LEV característico ($\log \lambda_K = f(k)$)

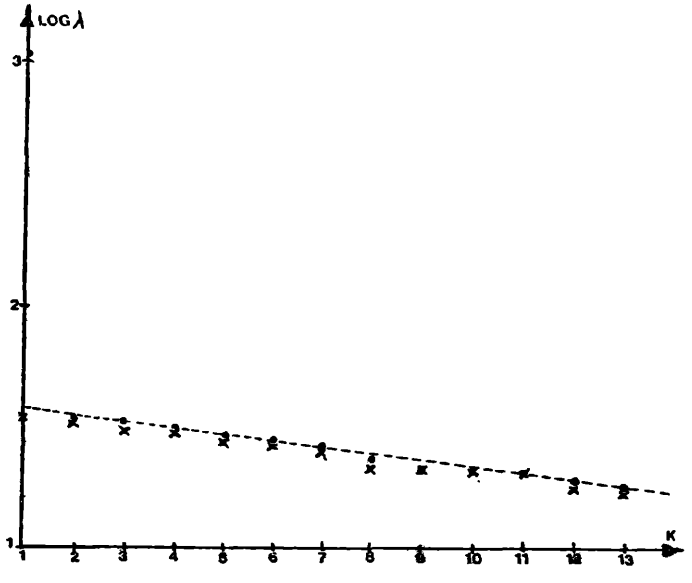

Fig.2 :Diagrama LEV de autovalores de números al azar.

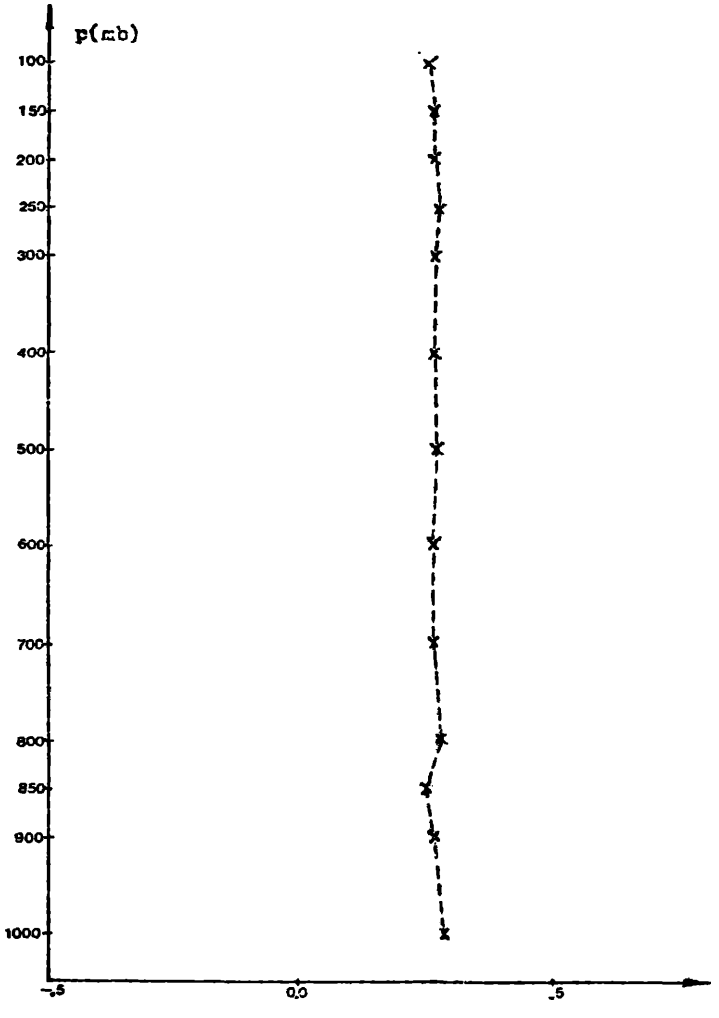

Fig.3 : Primer autovector de un campo de números obtenidos al azar.

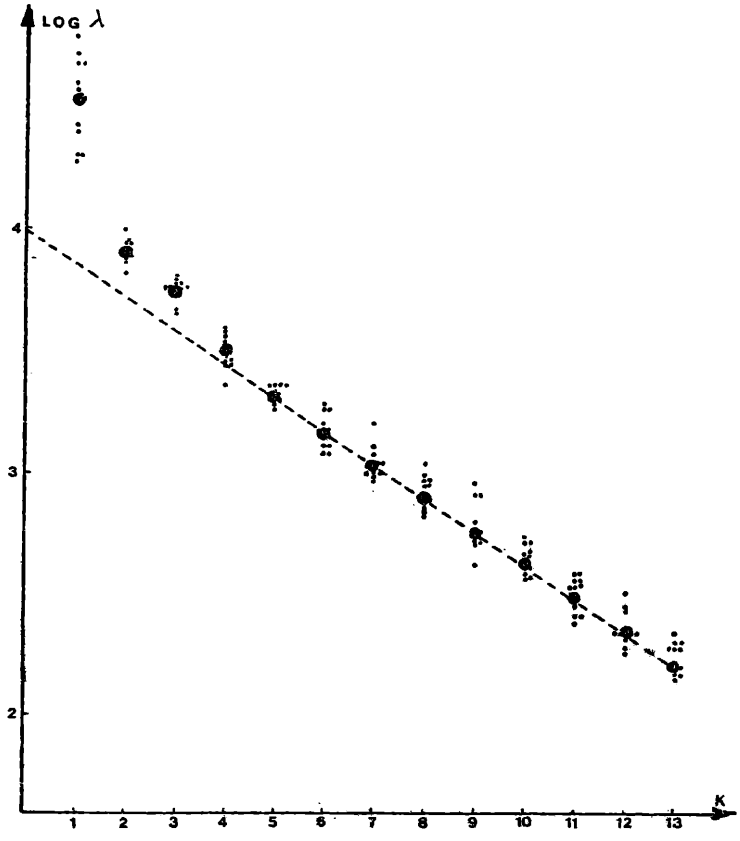

Fig.4 :Diagrama LEV de los sondeos térmicos de la estación aerológica Ezeiza.

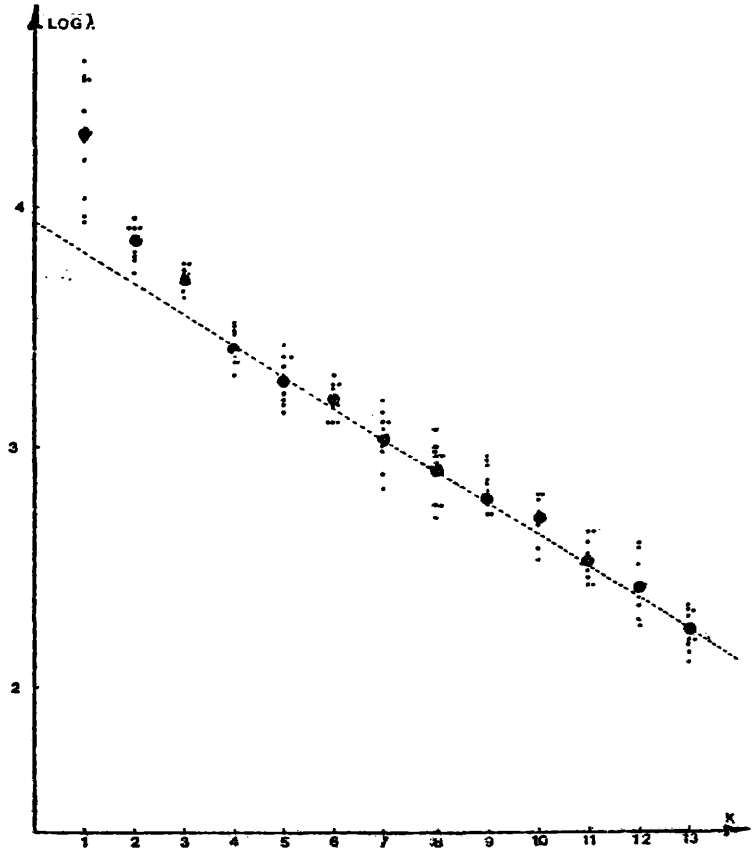

Fig.5 :Diagrama LEV de los sondeos térmicos de la estación aerológica Resistencia.

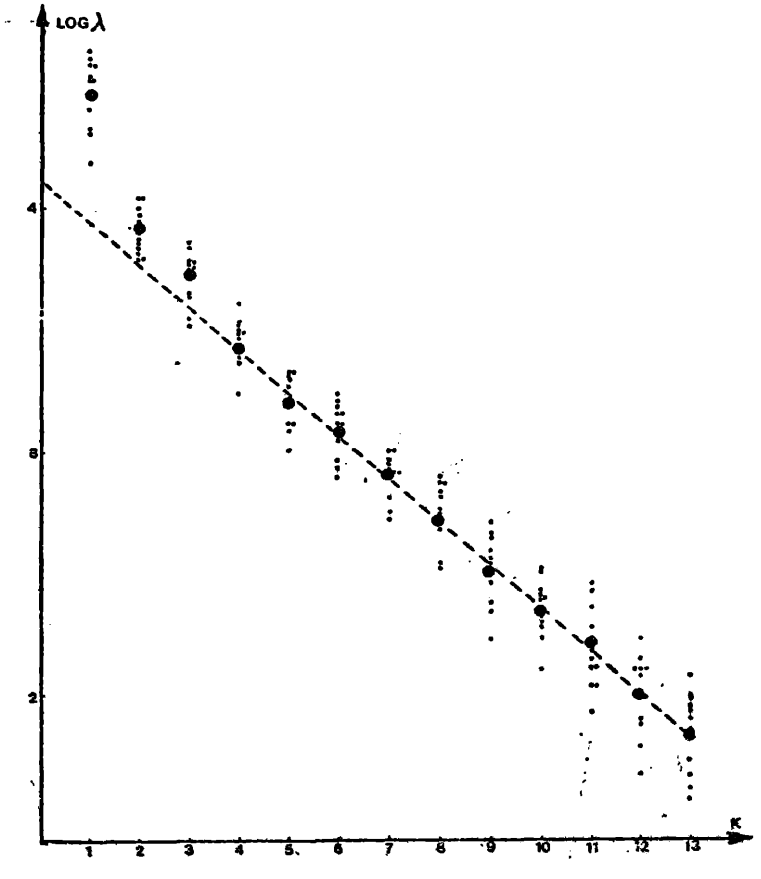

Fig.6 :Diagrama LEV de los sondeos térmicos de la estación aerológica Esoora.

Tabla I: Porcentaje de varianza explicada por los diez últimos autovectores.

ESTACION	EZEIZA	ESPORA	RESISTENCIA
Enero	26	18	31
Febrero	24	16	26
Marzo	20	15	25
Abril	20	18	20
Mayo	14	14	21
Junio	13	16	21
Julio	14	14	18
Agosto	15	16	21
Septiembre	16	15	20
Octubre	18	16	24
Noviembre	16	15	25
Diciembre	24	16	25