GEOACTA,Vol.12,n.2 (1985) pag 147 a 156

COMPACTADO DE ARCHIVOS METEOROLOGICOS MEDIANTE EL USO DE FUNCIONES ORTOGONALES EMPIRICAS

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

RESUMEN

Se ensaya el uso de funciones ortogonales empíricas para construír un archivo compactado de la información meteorológica a ser utilizada en minicomputadoras.

Determinando los errores cuadráticos medios por sondeo y por nivel para las distintas reconstrucciones posibles se realiza un análisis de la varianza para confirmar que los autovectores desechados no agregan información sobre el campo meteorológico real.

Se establece que con seis autovectores se logran los objetivos propuestos reduciendo en más del 50% la cantidad de dígitos a almacenar.

1. INTRODUCCION

Edward N. Lorenz (1956), introdujo el uso de funciones ortogonales empíricas (e.o.fs. :"Empirical orthogonal functions") en meteorología. Varios investigadores han aplicado las mismas a grandes volúmenes de información con distintos objetivos.

En noviembre de 1977 se llevó a cabo en el European Centre for Medium Range Weather Forecast (ECMWF) un seminario sobre el uso de funciones ortogonales empíricas en meteorología en el cual se mostró la utilidad de dichas funciones para comprimir grandes volúmenes de datos, transformándolos en conjuntos más facilmente manejables, de los cuales se elimina la información redundante, manteniendo un alto porcentaje de la varianza del conjunto original.

En la publicación interna del ECMWF se muestran los resultados obtenidos al explorar la eficiencia del uso de e.o.fs. para archivar los campos de alturas de 500 mb., encontrando que para la misma cantidad de información a archivar, dichas funciones son aproximadamente el doble de efectivas con respecto al uso de funciones armónicas esféricas, cuando se quiere realizar un archivo de alta densidad de información.

Por otra parte, se ha visto que el análisis de grandes volúmenes de información mediante el uso de computadoras está sujeto a errores computacionales debidos al truncado. Especialmente cuando se llevan a cabo largos cálculos , la acumulación de dichos errores puede llevar a resultados bastante alejados de la solución real . En éstos casos las e.o.fs.

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

tienen la ventaja frente a los polinomios y algunas otras funciones de ser practicamente inmunes respecto a los errores por truncado, además de ser más adecuadas para obtener la representación de campos con discontinuidades tales como los perfiles verticales de temperaturas.

2. APLICACION

En la actualidad, debido a la poca disponibilidad de computación, se ha intensificado en nuestro país el uso de minicomputadoras en el ámbito meteorológico. Uno de los principales inconvenientes de su uso es la limitada capacidad de almacenaje.

El Departamento de Meteorología de la Facultad de Ciencias Exactas y Naturales posee una minicomputadora Apple. La información necesaria para procesar los diferentes programas puede ser almacenada en diskettes cuya configuración es de 35 pistas de 16 sectores con una capacidad de 256 bytes cada uno, dando una capacidad total de almacenaje de 143.000 bytes por diskette.

En el proyecto "Tratamiento estadístico de los datos aerológicos de la República Argentina", de dicho Departamento, se utiliza como base de datos la información de las estaciones aerológicas de la República Argentina correspondiente a 20 años, actualmente almacenada en cintas magnéticas de 2800 pies, incompatibles con el sistema del minicomputador. Para disponer la información en forma conveniente para ser utilizada con la misma sería necesario grabar 500 diskettes. Ello significaría que durante un período de tiempo extremadamente prolongado debería utilizarse la computadora unicamente para ingresar la información, no contándose, para los distintos trabajos que se llevan a cabo, con el agravante de que posteriormente el acceso sería bastante dificultoso.

En un trabajo anterior , Lloret y Necco (1979) , se mostró la posibilidad de lograr una descripción compacta de la información de radiosondeos mediante la determinación de las e.o.f.s.

Sea P la matríz de datos correspondientes a uno de los elementos medidos en los radiosondeos: alturas geopotenciales, temperaturas, humedad o velocidad del viento, donde cada elemento Pij de la misma corresponde al desvío en el nivel i el día j con respecto al valor medio muestral en dicho nivel. Cada vector $\vec{P}j$ puede ser expresado como:

$$\vec{P_j} = \sum_{k=1}^{n} m_{kj} \vec{V_k} \qquad (1)$$

donde $\vec{V_k}$ son los autovectores de la matriz de covarianza de P y n es igual a la cantidad de niveles considerados.

En Lloret y Necco (1981) se realizó un análisis de la cantidad de autovectores a considerar, de acuerdo con el criterio de Farmer (1971), para conservar el mayor pocentaje de la varianza del campo original, encontrándose que si se considera el campo de temperaturas en los niveles tipo entre 1000 y 100 mb. y se ordenan los autovectores en orden decreciente respecto a la varianza explicada por los mismos, la representada por todos los autovectores desde el quinto en adelante corresponde al ruido (blanco) de la base de datos utilizada. Por lo tanto serán suficientes cuatro autovectores para conservar el mayor porcentaje de la varianza original.

No. AV \ MS	MS_D V_D+V_E	MS_N V_N+V_E	MS_T V_T+V_E	MS_{DN} $V_{DN}+V_E$	MS_{DT} $V_{DT}+V_E$	MS_{NT} $V_{NT}+V_E$	V_E MS_E
1 AV.	7.066	15.080.472	16	.904	.967	5.526	.365
2 AV.	6.910	15.054.879	188	1.205	964	328	282
3 AV.	6.792	15.089.612	16	1.488	943	243	211
4 AV.	6.677	15.070.419	126	1.599	871	215	188
5 AV.	7.009	15.077.743	1	1.616	821	136	182
6 AV.	7.043	15.080.310	1	1.653	833	80	171
7 AV.	7.071	15.073.706	4	1.682	835	21	161
8 AV.	7.065	15.073.002	4	1.700	835	12	160
9 AV.	7.062	15.073.693	3	1.713	834	11	157
10 AV.	7.059	15.073.987	5	1.721	835	9	155
11 AV.	7.056	15.074.197	6	1.727	833	6	155
12 AV.	7.055	15.074.307	6	1.729	834	5	155

TABLA I-Valores de cuadrados medios (MS) correspondientes a las distintas fuentes de variación, obtenidos para las doce muestras consideradas según la expresión (4).

En la tabla I se tienen los valores de MS calculados, correspondientes a todos los términos de la ecuación (5) , indicándose las varianzas estimadas por los mismos. Analizando las varianzas debidas a las interacciones vemos que en todas las muestras consideradas son muy significativas las correspondientes a días y niveles y días y tipo de sondeo y cuando se consideran las reconstrucciones con un autovector también hay interacción entre tipo de sondeo y nivel. Las dos primeras interacciones son obvias ya que dependen de la estructura térmica en cada caso. La última se puede explicar facilmente observando la figura 1 y la tabla II, allí se ve que tanto el primero como el segundo autovector contribuyen significativamente en los distintos niveles del sondeo mientras que el resto tienen asociados coeficientes mucho menores.

Debido al distinto comportamiento de los diferentes niveles es necesario analizar en cada nivel las varianzas dentro de las series y entre las series. Dichos resultados se presentan en la Tabla III, los valores grisados son aquellos que al serle aplicada la prueba de la razón de las varianzas de Snedecor (Moroney, 1965), para un nivel de confianza del 1%, dan un valor no significativo, indicando que las diferencias entre muestras son del mismo orden de magnitud

	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
Máx.	34	11	11	7	7	5	5	3	3	2	3	2
Min.	-18	-17	-9	-6	-9	-5	-5	-4	-4	-4	-2	-1

TABLA II. Rango de variación de los multiplicadores de enero en Santa Rosa.

Posteriormente se realizó el mismo análisis con las alturas geopotenciales y velocidades del viento también entre 1000 y 100 mb. y con las humedades relativas entre 1000 y 300 mb., encontrándose que en dichos casos es suficiente conservar tres, cuatro y cuatro autovectores respectivamente.

De acuerdo con éstos resultados la ecuación (1) se puede expresar como:

$$\vec{P}_j = \sum_{k=1}^{M-1} m_{kj} \vec{V}_k + \vec{E}_j \qquad (2)$$

con

$$\vec{E}_j = \sum_{k=M}^{n} m_{kj} \vec{V}_k \qquad (3)$$

donde M es el número de autovectores a partir del cual la varianza explicada corresponde al ruido y tal como se vió anteriormente depende de la variable que se considere.

3. RESULTADOS

Se reconstruyeron los sondeos a partir de los autovectores $(\vec{V}_k)$ y los coeficientes asociados (m_{ij}) obtenidos de la matríz de datos P :

$$\vec{P}^*_j(L) = \sum_{k=1}^{L} m_{kj} \vec{V}_k \qquad (4)$$

y se obtuvieron doce muestras haciendo variar L desde 1 hasta 12 , o sea que $\vec{P}^*_j(1)$ es el sondeo del día j obtenido a partir de la reconstrucción con un autovector, $\vec{P}^*_j(2)$ es el obtenido con dos autovectores y así sucesivamente.

Para cada una de las muestras así obtenidas se estableció la hipótesis nula de que provenía de la misma población que los sondeos originales $\vec{P}_j$ y se comprobó dicha hipótesis a través del análisis de la varianza.

Para ello se supone que la varianza total (V) puede ser reducida a distintas componentes, a saber:

$$V = V_D + V_N + V_T + V_{DN} + V_{DT} + V_{NT} + V_E \qquad (5)$$

V_D son las variaciones entre días, V_N entre niveles, V_T entre tipos de sondeo (sondeo real y reconstruído), V_{DN}, V_{NT} y V_{DT} son las interacciones entre las distintas fuentes de variación antes mencionadas y V_E es la varanaza residual debida a los errores.

Los cuadrados medios (MS)

$$MS_x = \frac{SS_x}{df_x} \qquad (6)$$

(donde SS_x es la suma de los cuadrados correspondientes al efecto x y df_x los grados de libertad asociados a dicha suma) son estimadores de las correspondientes varianzas.

que las correspondientes dentro de las muestras. A partir de la reconstrucción con seis autovectores las diferencias entre sondeos de uno y otro tipo, en todos los niveles, son despreciables frente a las diferencias entre sondeos, o sea que los seis últimos autovectores contienen unicamente en ruido de la base de datos utilizada.

Para verificar los resultados obtenidos se calculó el error cuadrático medio de la reconstrucción en cada nivel, mediante:

$$\vec{E_i}\ (L) = \sum_{j=1}^{N} \frac{(P_{ij}^{*}\ (L) - P_{ij})^2}{N}$$

donde $\vec{E_i}\ (L)$ es el error cuadrático medio de la reconstrucción con L autovectores y se puso como cota el error observacional (1°C). En la figura 2 se han graficado dichos errores en función de L para cada nivel, $E_i\ (L)$ disminuye rápidamente hasta L=4 donde, sin embargo, en varios niveles es superior al °C, luego el decrecimiento es más lento hasta alcanzar un valor de aproximadamente 0.3°C en L=12 que es el error debido al truncado en los diversos cálculos. A partir de la reconstrucción con seis autovectores (L=6) en todos los niveles el error es menor que el error observacional.

También se calculó el error cuadrático medio de la reconstrucción de cada sondeo mediante:

$$\vec{E_j}\ (L) = \sum_{i=1}^{n} \frac{(P_{ij}^{*}\ (L) - P_{ij})^2}{N}$$

y luego se calcularon las frecuencias porcentuales de dichos errores las cuales se encuentran graficadas en la figura 3 .

Allí se ve que para una reconstrucción con cuatro vectores en el 40% de los casos el error es superior a 1° C, mientras que con seis autovectores dicha frecuencia se reduce al 15%.

Se han graficado algunos sondeos reales y sus correspondientes reconstrucciones obtenidas a partir de seis autovectores y sus coeficientes asociados, notándose que el ajuste es totalmente satisfactorio pese a ser éstos los casos que presentan el mayor $E_j(6)$.

4. CONCLUSIONES

Los resultados indican que las temperaturas que se obtienen a partir de seis autovectores y sus coeficientes asociados reflejan totalmente las características de los sondeos originales.

Para almacenar las temperaturas de trece niveles de los sondeos correspondientes a diez años de una estación cualquiera se requieren 62 bytes por sondeo lo cual hace un total de 446.400 bytes mientras que para almacenar los correspondientes autovectores y coeficientes asociados serán necesarios 190.000 bytes lo cual reduce en más del 50% la cantidad de información a ser ingresada y consecuentemente el tiempo necesario para llevar a cabo su archivo. Resultados similares se han encontrado para otras estaciones y periodos.

De este trabajo surge así la gran utilidad de la aplicación de funciones ortogonales empíricas en la comprensión de archivos que, por su volúmen, resultan de uso engorroso o inconveniente en sistemas de cómputo con limitada capacidad.

Es importante destacar que la base de datos utilizada ha sido consistida sólo por rangos (Velasco y Necco (9)), no así hidrostáticamente, por lo tanto los niveles de error que se han tomado como cota son suficientemente débiles como para permitir que se introduzcan errores en la base de datos generada pero aseguran que la información contenida en los diskettes sea equivalente a la original.

Agradecimientos: Los autores agradecen la colaboración técnica de la Srta. Mónica Yacobone.

Este trabajo contó con el apoyo económico de la Secretaría de Estado de Ciencia y Técnología a través del subsidio 15466/79 y del Consejo Nacional de Investigaciones Científicas y Técnicas a través del subsidio 8773/79.

BIBLIOGRAFIA

European Centre for Medium Range Weather Forecast, 1978 : Verification and Storing with empirical orthogonal functions, Internal Report, No. 18.

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.

Lorenz, Edward N., 1956 : Empirical orthogonal functions and statistical weather prediction, Scientific Report No. 1, Statistical Forecasting Project.

Lloret, M. L. D. de y Necco, Gustavo V., 1979 : Resultados preliminares de la aplicación de funciones ortogonales empíricas a radiosondeos de la República Argentina, Meteorológica, Vol. 10, No. 2.

Lloret, M. L. D. de y Necco, Gustavo V., 1981 : Estimación del ruído en archivos de datos aerológicos utilizando funciones ortogonales empíricas, Geoacta, Vol. 11, No. 1.

Moroney, M. J., 1965 : Hechos y Estadísticas, Eudeba.

Velasco, I. y Necco, Gustavo V., 1982 : Aplicación de métodos objetivos al control de datos de radiosondeos en estaciones argentinas, Geoacta, Vol. 11, No. 2.

TABLA III- Valores de cuadrados medios MS (D dentro de las muestras y E entre muestras correspondientes a las doce muestras consideradas nivel por nivel. (En grisado diferencias significativas al 1 %, ver texto)

Niv. No. AV.	Niv. 1	Niv. 2	Niv. 3	Niv. 4	Niv. 5	Niv. 6	Niv. 7	Niv. 8	Niv. 9	Niv. 10	Niv. 11	Niv. 12
1 - D	2162	1791	1891	851	737	712	618	488	509	788	4222	688
E	12709	3115	292	1995	4494	3010	3258	6225	4537	330	8570	12567
2 - D	2022	1626	1229	883	852	875	818	697	644	801	695	969
E	714	5	209	788	268	64	245	1	1	1102	264	112
3 - D	2031	1634	1243	897	890	907	827	684	757	1544	828	971
E	660	3	243	870	372	29	214	15	141	59	69	13
4 - D	2040	1659	1273	907	888	891	823	699	799	1545	836	1238
E	677	5	189	851	388	14	102	1	58	56	81	70
5 - D	1981	1635	1270	978	1032	989	849	684	799	1538	799	1238
E	278	1	214	348	13	28	196	18	6	28	95	77
6 - D	1930	1649	1331	1050	1027	992	841	708	835	1547	814	1239
E	81	1	53	83	38	44	90	0	23	156	30	70
7 - D	1913	1647	1336	1106	1058	989	836	702	839	1695	824	1235
E	40	0	41	0	1	29	8	46	36	20	12	2
8 - D	1898	1660	1363	1109	1080	991	831	698	876	1608	825	1238
E	11	3	12	2	4	23	35	33	3	7	6	1
9 - D	1909	1658	1361	1115	1091	1006	831	700	883	1610	825	1238
E	16	1	18	4	2	13	38	23	1	6	6	1
10 - D	1913	1661	1365	1148	1079	1014	832	698	883	1609	825	1239
E	4	11	5	6	9	2	35	20	1	5	5	0
11 - D	1920	1657	1365	1150	1078	1009	859	685	898	1609	824	1242
E	3	17	6	8	6	21	1	8	0	2	3	0
12 - D	1918	1676	1360	1151	1077	1007	862	684	897	1006	823	1241
E	6	4	0	10	5	26	1	8	0	2	3	0

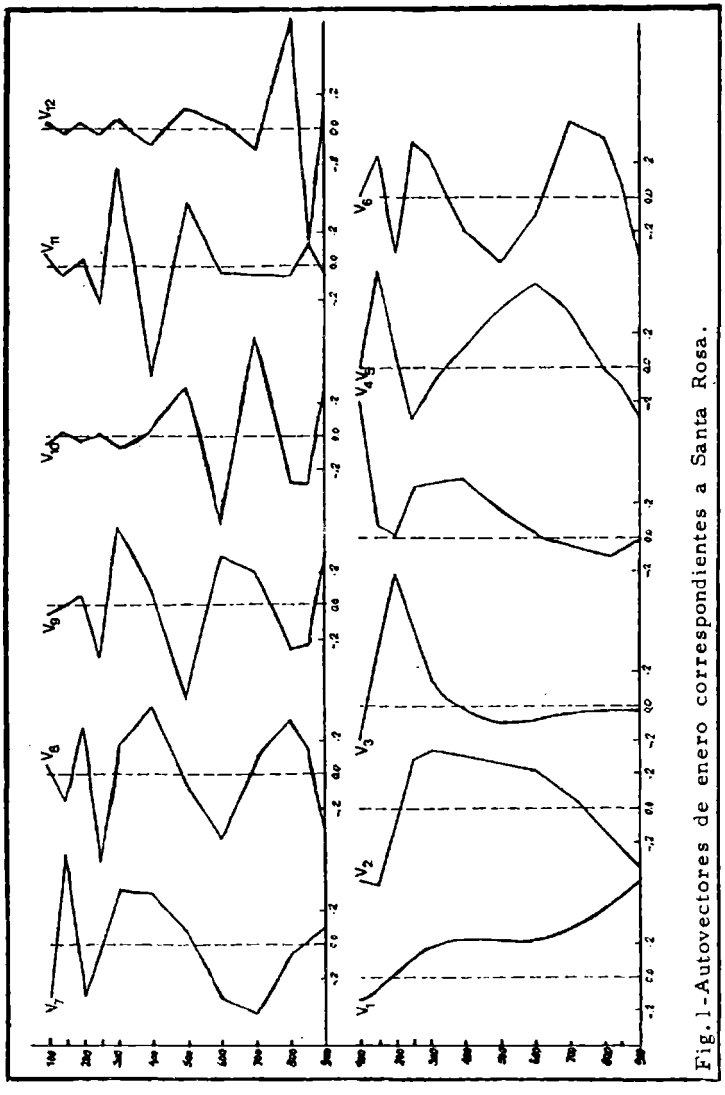

Fig.1-Autovectores de enero correspondientes a Santa Rosa.

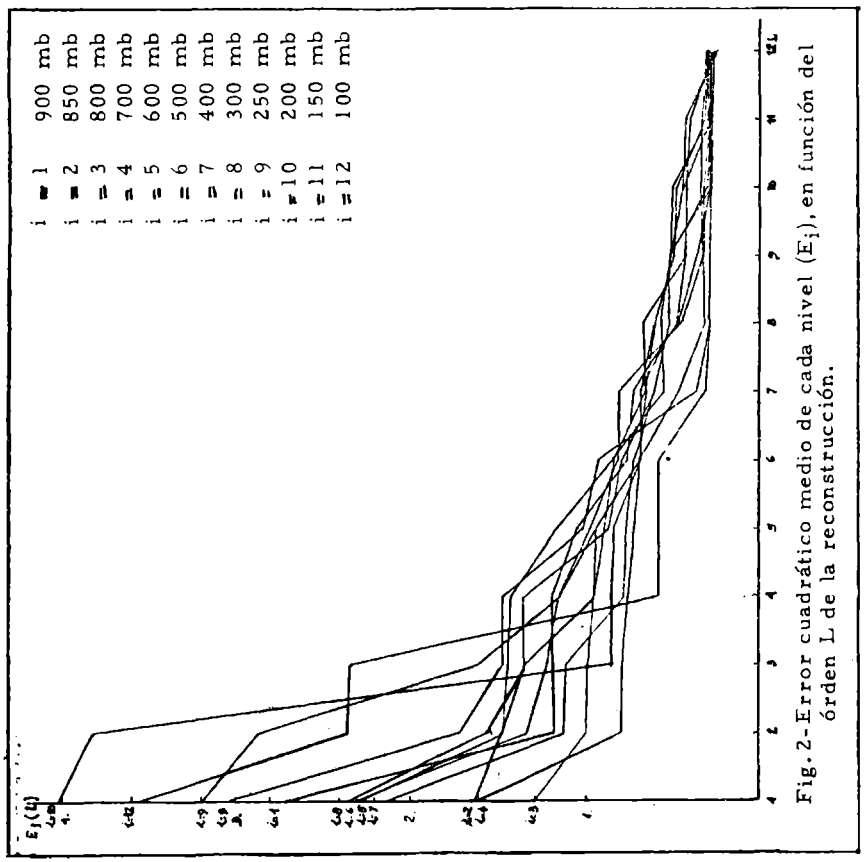

Fig. 2- Error cuadrático medio de cada nivel (E_i), en función del órden L de la reconstrucción.

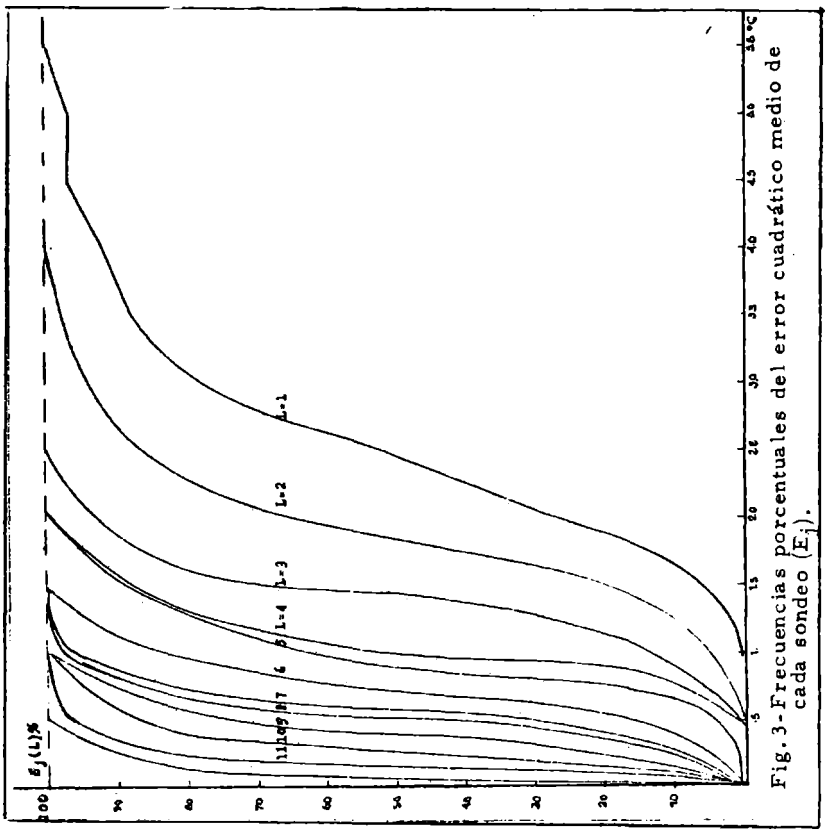

Fig.3-Frecuencias porcentuales del error cuadrático medio de cada sondeo (E_j).