Revista INTEGRACIÓN
Universidad Industrial de Santander
Escuela de Matemáticas
Vol. 16, No 2, p. 67–86, julio–diciembre de 1998

Análisis de la estructura común de varias tablas de datos a través de diferentes técnicas factoriales*

Lydia Lera Marqués**
Amparo Vallejo Arboleda***
Humberto González****
Armando Gutiérrez Arias*****

(Sometido a publicación en diciembre 15, 2000)

Resumen

En este trabajo se analizan diferentes técnicas de tipo factorial, entre las que se encuentran el método de Krzanowski, el método `STATIS`, el método de Procusto Generalizado, el Análisis de Componentes Principales Triple y el Doble Análisis de Componentes Principales. El objetivo fundamental de todos los métodos es la búsqueda de una estructura común entre diferentes tablas o grupos de datos y/o la definición de un compromiso. Los métodos se comparan entre sí a través de un ejemplo.

Palabras claves: Técnicas factoriales, estructura común, compromiso.

*Este artículo fue escrito en el marco del proyecto de investigación *Estudio de Técnicas Factoriales para el Análisis de Datos Longitudinales y provenientes de una función de crecimiento*, financiado por la Universidad de Antioquia.

**Centro de Matemática y Física Teórica, ICIMAF, La Habana, CUBA. *E–mail*: `lydia@cidet.icmf.inf.cu`. Profesora Departamento de Matemáticas, Universidad de Antioquia, Medellín, COLOMBIA.

***Profesora del Departamento de Matemáticas, Universidad de Antioquia, Medellín, COLOMBIA. (*E–mail*: `avallejo@matematicas.udea.edu.co`)

****Departamento de Contaminación, Aguas de Cartagena; Cartagena, COLOMBIA. (*E–mail*: `hgonzalez@acuacar.com`)

*****Profesor del Departamento de Matemáticas, Universidad de Antioquia, Medellín, COLOMBIA. (*E–mail*: `aga@matematicas.udea.edu.co`). Centro de Matemática y Física Teórica, ICIMAF, La Habana, CUBA.

Abstract

In this work we analyze several factor techniques, among which are the Krzanowski method, the STATIS method, the Generalized Procusto method, the three–way principal component analysis and the double principal component analysis. The main objective of the methods is to search a common structure between different data tables or groups and the definition of a compromise.These methods are compared by using an example.

Key words: Factor methods, common structure, compromise.

1. Introducción

En los últimos años han surgido métodos de tipo factorial cuyo objetivo principal es la búsqueda de una estructura común para diferentes tablas de datos y a partir de esta estructura común explicar las diferencias entre las distintas tablas. Estos métodos tienen en común su relación directa con el **A**nálisis de **C**omponentes **P**rincipales (ACP). En muchas aplicaciones, cuando se tienen diferentes tablas de datos, que han sido analizadas separadamente por un ACP, se presenta el problema de compararlas, lo que no es posible generalmente de forma directa ya que los coeficientes de la transformación componente principal presentan variabilidad muestral, pudiendo ocurrir que la transformación obtenida en los diferentes grupos difiera solo por error muestral (Flury and Riedwyl [5]). De ahí la importancia de estos métodos.

Las tablas de datos pueden estar formadas por: los mismos individuos medidos en diferentes ocasiones por las mismas variables, los mismos individuos medidos en diferentes ocasiones por diferentes variables y diferentes conjuntos de individuos a los que se les miden las mismas variables. Los dos primeros casos, que se refieren a fenómenos que evolucionan en el tiempo, forman parte de los datos longitudinales.

Estos datos se organizan en matrices cúbicas (tablas de 3 entradas donde una o más de las entradas puede ser el tiempo). Sea X una matriz cúbica, cuyo elemento x_{ijk}, $(i, j, k) \in I \times J \times K$ representa al individuo i con $i = 1, \ldots, n$, para la variable j con $j = 1, \ldots, P$, en la ocasión k con $k = 1, \ldots, K$. Se estudia el primer caso en donde se tienen K tablas de datos formadas por los mismos individuos y las mismas variables medidas en el tiempo ya que es un caso común a todos los métodos analizados.

En este trabajo se presentan los siguientes métodos: el **M**étodo de **C**omparación de **G**rupos de **C**omponentes **P**rincipales (MCGCP) de Krzanowski [9], el Método STATIS (**S**tructuration des **T**ableaux **A** **T**rois **I**ndices de la **S**tatistique) (Lavit [15]), el Método Generalizado de Procusto (MGP) (Gower [6]), el Análisis

de Componentes Principales Triple (Tucker [22] y Kroonenberg [10]) y el Doble Análisis de Componentes Principales (Bouroche [2] citado por GERI [7]), así como su utilización e interpretación a través de un ejemplo de aplicación en estudios sobre contaminación marina, por metales pesados, en sedimentos, de la Bahía de Cienfuegos, Cuba (Lera y González [17]). Todos estos métodos tienen en común la búsqueda de un compromiso y/o el análisis de la estructura de las diferentes tablas de datos.

2. Descripción de los métodos

2.1. Método de Comparación de Grupos de Componentes Principales

Este método tiene como objetivo fundamental la búsqueda de una estructura común entre K tablas de observaciones, a las que se les miden las mismas variables. El método se basa en la aplicación del Análisis de Componentes Principales a cada una de las tablas y su descripción en términos de un número menor de componentes principales. A continuación se comparan los componentes principales de las K tablas de datos, calculando los ángulos (θ_i) entre los subespacios generados por los primeros componentes principales de cada tabla. Como medida de similitud se utiliza $\sum_{i=1}^{K} \cos^2 \theta_i$. Se cumple que $0 \leq \sum_{i=1}^{K} \cos^2 \theta_i \leq K$, donde K corresponde a espacios coincidentes y 0 a espacios ortogonales.

Este método tiene las desventajas de que no brinda la posibilidad de gráficos y no da una estimación de la matriz de varianzas y covarianzas.

2.2. El Método STATIS

El método STATIS es un método exploratorio del Análisis de Datos, cuyo principal objetivo es también la búsqueda de una estructura común entre K tablas, que pueden estar formadas por los mismos individuos a los que se les han hecho mediciones a través del tiempo o por individuos diferentes a los que se les miden las mismas variables. Esta estructura (intraestructura) está descrita por las distancias mutuas entre las observaciones dada por

$$\mathbf{W}_k = \mathbf{X}_k \mathbf{X}_k',$$

donde $\mathbf{X}_k$ es la tabla de datos $n \times p$ en la ocasión k (o por $\mathbf{V}_k = \mathbf{X}_k'\mathbf{X}_k$ en el caso dual), la cual representa la evolución de los individuos (o de las variables).

Contrariamente a este término se define la interestructura como las relaciones entre grupos de observaciones, con el objetivo de comparar las tablas, descritas por las distancias entre las $\mathbf{W}_k$, y se deduce del producto escalar de Hilbert–Schmidt que equivale a la suma de los cuadrados de las covarianzas entre las variables en diferentes ocasiones, y está dado por

$$(\mathbf{W}_{k_1}, \mathbf{W}_{k_2}) = tr(D\mathbf{W}_{k_1} D\mathbf{W}_{k_2}),$$

donde D es la matriz diagonal obtenida de los pesos de los individuos. Si los pesos de los individuos son iguales, entonces

$$D = \tfrac{1}{n} I_n,$$

y se construye una imagen euclidiana plana de las tablas; si $\mathbf{G}_1, \ldots, \mathbf{G}_k$ es la nube de puntos formada por los grupos, el coseno del ángulo entre los vectores $\mathbf{OG}_k$ y $\mathbf{OG}_1$ es la aproximación del producto escalar normado entre $\mathbf{W}_k$ y $\mathbf{W}_1$, y se lo llama coeficiente RV, que representa un índice de similitud:

$$RV(k_i, k_j) = \frac{S_{k_i,k_j}}{\sqrt{S_{k_i,k_i} S_{k_j,k_j}}},$$

donde

$$\begin{aligned} S_{k_i k_j} &= (W_{k_i}, W_{k_j}) = tr(DW_{k_i} DW_{k_j}), \\ S_{k_i k_i} &= (W_{k_i}, W_{k_i}) = tr(DW_{k_i} DW_{k_i}), \\ S_{k_j k_j} &= (W_{k_j}, W_{k_j}) = tr(DW_{k_j} DW_{k_j}). \end{aligned}$$

Si el coeficiente RV es aproximadamente igual a 1 se dice que las observaciones tienen la misma estructura en el interior de las tablas $\mathbf{X}_k$ y $\mathbf{X}_l$. Se define un compromiso, que representa una tabla promedio y se describe su estructura, así como también pueden calcularse las trayectorias que definirán la evolución de los individuos (o de las variables). El método es como una generalización del `ACP` y brinda la posibilidad de gráficos.

2.3. El Análisis de Componentes Principales Triple

Este método tuvo su origen en las Ciencias Sociales, pero su uso se ha extendido a otras ramas de la ciencia. Se considera que los datos se pueden arreglar

en bloques tridimensionales de *sujetos* × *variables* × *condiciones*. Luego, para la tabla de datos

$$\mathbf{X} = (x_{ijk}), \qquad \begin{aligned} i &= 1, \ldots, I, \\ j &= 1, \ldots, J, \\ k &= 1, \ldots, K, \end{aligned}$$

se tiene el modelo (conocido como Tucker 3)

$$x_{ijk} = \sum\sum\sum a_{ip} b_{jq} c_{kr} g_{pqr} + e_{ijk}, \qquad \begin{aligned} i &= 1, \ldots, I, \\ j &= 1, \ldots, J, \\ k &= 1, \ldots, K, \end{aligned}$$

donde a_{ip}, b_{jq} y c_{kr} son elementos de las matrices $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$, que son las matrices de componentes principales de cada una de las clasificaciones, respectivamente; g_{pqr} son los elementos de la matriz central $\mathbf{G}$, que representan las interacciones entre las componentes y e_{ijk} son los errores de aproximación.

El modelo puede escribirse en forma matricial como

$$\mathbf{X} = \mathbf{A}\mathbf{G}\left(\mathbf{C}^t \otimes \mathbf{D}^t\right) + \mathbf{E}.$$

En las aplicaciones prácticas, generalmente se utilizan sólo los dos primeros componentes principales de cada una de las matrices $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$. Se define la función de pérdida

$$(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{G}) = \left\| \mathbf{X} - \mathbf{A}\mathbf{G}\left(\mathbf{C}^t \otimes \mathbf{D}^t\right) + \mathbf{E} \right\|,$$

donde $\|\cdot\|$ es la norma euclidiana. Sin pérdida de generalidad, se asume que $\mathbf{A}$, $\mathbf{B}$ y $\mathbf{C}$ son ortonormales.

Una solución para la estimación mínimo cuadrática de $\mathbf{G}$ a partir de la minimización de esta función de pérdida fue obtenida por Kroonenberg y De Leeuw [11], y es la siguiente:

$$\mathbf{G} = \mathbf{A}^t \mathbf{X} \left(\mathbf{C} \otimes \mathbf{D}\right).$$

Se utiliza como medida de ajuste del modelo

$$\mathbf{SC}_{\text{Ajuste}} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}^2.$$

2.4. El Método Generalizado de Procusto

Este método (Gower [6]) trata el problema de la determinación de un compromiso $\mathbf{C}$ tal que

$$\text{mín}_{\mathbf{C}} \sum_{k=1}^{T} \|\rho_k \mathbf{X}_k \mathbf{P}_k - \mathbf{C}\|^2$$

con la restricción

$$\sum \rho_i^2 trace\left(\mathbf{X}_i^t \mathbf{X}_i\right) = \sum trace\left(\mathbf{X}_i^t \mathbf{X}_i\right).$$

Existen diversas variantes del Método Generalizado de Procusto (Quannari et al. [19]; Lacourly y Lera [13]; Lacourly, Lera y Farías [14]). En este estudio se utiliza una de las variantes propuestas por Lacourly, Lera y Farías [14].

El criterio elegido para obtener el compromiso $\mathbf{C}$ es tal que:

$$\text{mín}_{\mathbf{C},\rho_k,\mathbf{P}_k} \sum_{k=1}^{T} \alpha_k \|\mathbf{X}_k - \rho_k \mathbf{C}\mathbf{P}_k\|^2 = \text{mín}_{\mathbf{C},\rho_k,\mathbf{P}_k} \sum_{k=1}^{T} tr\alpha_k \left(\mathbf{X}_k - \rho_k \mathbf{C}\mathbf{P}_k\right)\left(\mathbf{X}_k - \rho_k \mathbf{C}\mathbf{P}_k\right)^t,$$

donde

- X_k : es la i–ésima tabla de dimensión $m \times n$ y $\|\cdot\|$ es la norma de Frobenius, ($n =$ no. de observaciones, $m =$ no. de variables);
- α_k : ponderación de las tablas;
- ρ_k : k–ésimo coeficiente de escala;
- $\mathbf{P}_k$: k–ésima matriz de rotación de dimensión $m \times m$ tal que $\mathbf{P}_k \mathbf{P}_k^t = \mathbf{I}$;
- $\mathbf{C}$: es la matriz compromiso de dimensión $m \times n$.

En resumen, se tiene que:

$$\begin{aligned}
\mathbf{C} &= \frac{\sum_{k=1}^{T} \left(\rho_k \mathbf{X}_k \mathbf{P}_k^t\right)}{\sum_{k=1}^{T} \rho_k^2}, \\
\mathbf{P}_k &= \left(\mathbf{C}^t \mathbf{X}_k \mathbf{X}_k^t \mathbf{C}\right)^{1/2} \left(\mathbf{X}_k^t \mathbf{C}\right)^{-1}, \\
\rho_k &= \frac{trace\left(\mathbf{C}\mathbf{P}_k \mathbf{X}_k^t\right)}{trace\left(\mathbf{C}\mathbf{C}^t\right)}.
\end{aligned}$$

Se obtiene un compromiso óptimo y se trazan las trayectorias por medio de un programa implementado en MATLAB (Lacourly y Lera [13]).

2.5. El Doble Análisis de Componentes Principales

Este método (GERI [7]) se aplica a matrices cúbicas, cuyos datos no están centrados inicialmente. El objetivo principal del método es comparar globalmente la evolución de las relaciones entre las diferentes variables, así como la de los individuos.

Consta de tres etapas:

- La primera analiza un fenómeno de evolución global, llamado "desplazamiento de la nube de puntos a lo largo del tiempo". Esta evolución es estudiada por un ACP de los centros de gravedad de las nubes y corresponde al estudio de la interestructura.
- La segunda consiste en estudiar la deformación de la nube alrededor de su centro de gravedad; para esto se realizan T ACP a T nubes de puntos, centrados en su centro de gravedad, para eliminar el fenómeno de evolución global.
- La tercera consiste en representar en un mismo espacio las evoluciones de los diferentes individuos a lo largo del tiempo; el problema mayor de esta etapa es precisamente buscar un espacio de representación común, objetivo fundamental de todos los análisis de datos evolutivos: encontrar un espacio en el cual puedan ser representadas las trayectorias de los individuos. Bouroche [2] propone el uso de cuatro procedimientos diferentes, que permiten determinar un referencial común. Aquí nos referimos sólo a uno de los procedimientos y es el que se presenta a continuación.

El criterio consiste en maximizar la inercia del conjunto de nubes proyectadas resolviendo el problema de optimización siguiente:

$$\text{máx}_{v_i,\ldots,v_2} \sum_{l=1}^{q} v_l^t \mathbf{V} v_1,$$

donde

$$\mathbf{V} = \sum_{t=1}^{T} V_t.$$

Estos métodos se encuentran implementados en MATLAB y SAS.

Los métodos analizados siguen un enfoque descriptivo, centrándose en la estructura algebraica de los datos, sin tener en cuenta supuestos distribucionales.

3. Un ejemplo de aplicación

Se obtuvieron muestras de sedimentos superficiales de 11 estaciones, en 5 ocasiones diferentes (abril de 1988, septiembre de 1988, marzo de 1989, septiembre de 1989 y abril de 1990), en la bahía de Cienfuegos, Cuba. Se analizaron los metales `Cu`, `Pb`, `Zn` como indicadores de contaminación urbano–industrial y `Co`, `Fe`, `Mn` y `Ni` como indicadores de contaminación industrial, con el objetivo de buscar una estructura común de la contaminación así como analizar su evolución.

Para el procesamiento estadístico se autonormalizaron los datos con relación al aluminio.

3.1. Resultados y discusión

En las tablas 1 y 2 se tienen las medias y desviaciones estándar, respectivamente, de las variables para cada uno de los muestreos. Las concentraciones de los metales están dadas en μg/g.

Tabla 1: Medias de las variables analizadas

	Co	Cu	Fe	Mn	Ni	Pb	Zn
Abril/1988	7, 8636	60, 3636	3, 8909	638, 8182	40, 7273	25, 4000	85, 1818
Sept/1988	7, 5182	64, 0909	4, 3400	543, 1818	34, 1818	11, 7818	94, 7273
Marzo/1989	8, 1455	50, 4545	4, 3309	479, 7273	30, 7273	11, 2273	110, 0909
Sept/1989	7, 3091	46, 6364	3, 1782	410, 6364	35, 1818	24, 5455	83, 3636
Abril/1990	10, 9545	65, 9091	4, 0427	468, 1818	37, 0000	22, 4091	107, 9091

Tabla 2: Desviación Estándar de las variables analizadas

	Co	Cu	Fe	Mn	Ni	Pb	Zn
Abril/1988	2, 6662	18, 9646	0, 7462	529, 1015	5, 9681	43, 7827	21, 0752
Sept/1988	2, 8712	31, 3192	1, 2382	338, 6481	9, 3469	12, 8014	29, 0692
Marzo/1989	2, 1929	13, 7285	0, 6297	169, 3081	7, 0724	8, 5312	24, 8775
Sept/1989	2, 4353	18, 3481	0, 7600	176, 0700	8, 3405	10, 3862	24, 6750
Abril/1990	2, 2174	30, 1213	0, 8947	206, 6799	8, 2219	28, 3867	34, 6423

3.1.1. `STATIS`

La figura 1, que es una representación euclidiana de las tablas, muestra una estructura ligeramente diferente para los datos de la tabla 1 (muestreo de abril de 1988) y de la tabla 2 (septiembre, 1988) con relación al resto. La imagen euclidiana del compromiso indica que la estación 11 y las estaciones 7, 8 y 10, tienen un comportamiento diferente al resto de las estaciones (figura 2), presentando gran contaminación industrial y urbano–industrial la 11 y un grado

de contaminacion intermedio las 7, 8 y 10, predominando la contaminación urbano–industrial.

Figura 1: Imagen euclidiana de las tablas.

Figura 2: Imagen euclideana del compromiso de las estaciones.

En la tabla 3 se observa que los coeficientes de la matriz RV son mayores que 0,6, lo que nos sugiere que las estaciones presentan una estructura común en

los muestreos efectuados.

Tabla 3: Matriz RV

1,0000	0,6094	0,7044	0,6236	0,7102
0,6094	1,0000	0,7189	0,7769	0,6585
0,7044	0,7189	1,0000	0,7546	0,7587
0,6236	0,7769	0,7546	1,0000	0,6904
0,7102	0,6585	0,7587	0,6904	1,0000

Al analizar las correlaciones de las variables con los ejes (tabla 4, figura 3) se puede ver que el `Co`, `Cu`, `Fe`, `Mn` y `Ni` son los metales más correlacionados con el eje 1, que representa la contaminación industrial, el porcentaje (%) de varianza, que explica este eje es del 45,84 %. El `Pb` y `Zn` están correlacionados con el eje 2, representando la contaminación urbano–industrial.

Al analizar las posiciones de compromiso de las observaciones (figura 2) se pueden diferenciar los siguientes grupos de estaciones:

- La estación más contaminada, que corresponde a la estación 11;
- Las estaciones con una contaminación intermedia, que corresponden a las estaciones 7, 8 y 10;
- Las estaciones menos contaminadas, que corresponde al resto de las estaciones.

Tabla 4: Correlaciones de las variables con los ejes

Metales	**Grupo 1 (3/88)**		**Grupo 2 (9/88)**		**Grupo 3 (3/89)**	
`Co`	0,8838	−0,1954	0,9212	−0,0888	0,8047	0,1590
`Cu`	0,6546	−0,6552	0,7102	−0,6221	0,7171	−0,6212
`Fe`	0,8258	0,1909	0,9057	−0,0908	0,6947	0,5178
`Mn`	0,8131	−0,4487	0,8412	−0,4474	0,8636	−0,4029
`Ni`	0,6563	0,4374	0,7356	0,3748	0,6751	0,6566
`Pb`	−0,3974	−0,5942	−0,5160	−0,5713	−0,3180	−0,6397
`Zn`	−0,2563	−0,6879	−0,4480	−0,5226	−0,2441	−0,7313

Metales	**Grupo 4 (9/89)**		**Grupo 5 (3/90)**	
`Co`	0,9045	−0,0879	0,9099	0,0564
`Cu`	0,6250	−0,7428	−0,1303	−0,8708
`Fe`	0,8849	0,0515	0,7531	0,1517
`Mn`	0,8756	−0,3670	0,7226	−0,5388
`Ni`	0,5549	0,1812	0,0791	0,1193
`Pb`	−0,4359	−0,7322	−0,5176	−0,6792
`Zn`	0,0883	−0,9311	−0,3900	−0,7748

Figura 3: Correlaciones de las variables con los ejes de compromiso.

En la figura 4 se observan las trayectorias de las estaciones que se interpretan como la evolución de una estación ficticia promedio, que asume como valores los promedios de las variables por muestreo. Como puede verse, tienen diferentes formas, por ejemplo, la trayectoria de la estación 8 es como si se enquistara sobre sí misma, lo que indica en este caso ausencia de evolución, y la estación 10 tiene una trayectoria de gran amplitud, reflejando una evolución en el tiempo muy diferente de la evolución promedio.

3.1.2. Análisis de Componentes Principales Triple

Se conformaron las matrices **A**, **B** y **C** con los 3 primeros componentes principales de la matriz formada por las variables, los 2 primeros componentes principales de la matriz formada los muestreos y los 3 primeros componentes principales de la matriz formada por las estaciones.

La tabla 5 muestra los 3 primeros vectores y valores propios de la matriz **A**, que explican el 77,86 % de la variabilidad total. Se observa que la variación más importante está dada en un primer eje caracterizado por `Co`, `Fe`, `Mn` y `Ni` (`Fe` con signo contrario); indicadores de contaminación industrial, fundamentalmente. El segundo eje se caracteriza por `Cu`, `Pb` y `Zn`, indicadores de contaminación

Figura 4: Trayectorias de las estaciones.

Figura 5: Representación de las estaciones en el compromiso.

urbano–industrial.

Tabla 5: ACP Variables

	$-0,513$	$-0,0152$	$0,0307$
	$-0,3721$	$0,5313$	$-0,1125$
	$0,4935$	$-0,1023$	$0,2388$
	$-0,4672$	$0,1850$	$-0,3627$
	$-0,3008$	$-0,2542$	$0,1374$
	$0,2061$	$0,5696$	$-0,4053$
	$0,0601$	$0,5326$	$0,7840$
V, P,	$42,6\,\%$	$23,7\,\%$	$11,52\,\%$

La tabla 6 muestra los 2 primeros vectores y valores propios de la matriz **B**. La variación más importante está caracterizada por los muestreos.

Tabla 6: ACP Muestreos

	$0,4120$	$-0,7433$
	$0,4646$	$0,2955$
	$0,4377$	$0,5702$
	$0,4625$	$0,0103$
	$0,4571$	$-0,1869$
V.P.	$60,55\,\%$	$13,37\,\%$

El `ACP` de las estaciones determinó la selección de 3 componentes principales que explican un 70,99 % de la variabilidad total. El primer componente está caracterizado por las estaciones menos contaminadas, el segundo componente por las estaciones de contaminación intermedia y el tercero, la estación 11, que es la más contaminada.

La Tabla 7 muestra la matriz central **G**. Esta matriz da las relaciones entre los diferentes modos. La forma más sencilla de analizarla es describir los sujetos característicos.

Tabla 7: Matriz Central (G)

	C_1		C_2		C_3	
	B_1	B_2	B_1	B_2	B_1	B_2
A_1	$-1,9847$	$6,0384$	$-0,9268$	$-1,9856$	$-1,1070$	$-3,7467$
A_2	$-0,5287$	$-5,0186$	$-8,0835$	$6,6919$	$-3,5811$	$4,3682$
A_3	$-1,7346$	$1,2983$	$4,1807$	$3,5271$	$0,9771$	$2,8055$

Se observa que las combinaciones de componentes más importantes con relación a los sujetos son:

- **Sujeto 1**. A_1, B_2, C_1: representa las estaciones menos contaminadas con respecto a contaminación industrial en todos muestreos.
- **Sujeto 2**. A_2, B_1, C_2 : representa contaminación urbano–industrial para las estaciones intermedias en todos los muestreos.

A_2, B_2, C_2 : representa contaminación urbano–industrial para las estaciones intermedias en los muestreos de abril/1988 y marzo/1989.

- **Sujeto 3**. A_2, B_2, C_3 : representa contaminación urbano–industrial para las estaciones intermedias en los muestreos de abril/1988 y marzo/1989; para la estación más contaminada en los muestreos de abril/1988 y marzo/1989.

  Se tiene que el modelo tiene un ajuste del 71,28 %.

3.1.3. Método de Comparación de Grupos de Componentes Principales

Al analizar los resultados obtenidos por este método (tabla 8), se observa que con 2 componentes principales se logra una buena aproximación del espacio común, siendo el `Co`, `Cu`, `Fe`, `Mn` y `Ni`, indicadores de contaminación industrial, fundamentalmente los metales que relacionan a los 5 muestreos realizados.

Se ve que $\sum_{i=1}^{5} \cos^2 \theta_i$ se aproxima a 5 para el primer componente.

Tabla 8: Análisis del espacio común de la bahía de Cienfuegos en los muestreos de abril y septiembre de 1988, marzo y septiembre de 1989 y abril de 1990

Metales	**Vectores y valores propios**	
`Co`	$-0,5045$	$0,1218$
`Cu`	$-0,4479$	$0,3745$
`Fe`	$-0,4937$	$-0,0061$
`Mn`	$-0,4773$	$0,3884$
`Ni`	$-0,4452$	$-0,2526$
`Pb`	$0,3906$	$0,4735$
`Zn`	$0,1619$	$0,4650$
Valores propios	$4,8975$	$4,2987$

	CP 1		**CP 2**	
	$\cos^2\theta$	θ	$\cos^2\theta$	θ
Abril/1988	$0,9931$	$0,0832$	$0,5020$	$0,7834$
Septiembre/1988	$0,9703$	$0,1732$	$0,9394$	$0,2487$
Marzo/1989	$0,9947$	$0,0729$	$0,9642$	$0,1904$
Septiembre/1989	$0,9888$	$0,1060$	$0,9458$	$0,2350$
Abril/1990	$0,9505$	$0,2244$	$0,9474$	$0,2314$
$\sum \cos^2\theta$	$4,8974$		$4,2988$	

Figura 6: Representación de las estaciones en el compromiso de Procrusto.

Figura 7: Trayectorias en el compromiso de Procrusto para las estaciones.

3.1.4. MPG

La figura 5 muestra la representación euclidiana de las estaciones en el compromiso. Puede observarse que se mantiene la misma relación entre las estaciones que para el método STATIS (ver figura 2).

La figura 6 da el compromiso de Procrusto con relación a las tablas; se observa que éste se encuentra prácticamente en el centro de las tablas, por lo que el compromiso representa un promedio para las tablas. La figura 7, de las trayectorias de las estaciones en el compromiso, muestra que, al igual que en el caso del STATIS, la estación 11 y las estaciones 7, 8 y 10 presentan un comportamiento diferente del resto de las estaciones.

Luego, el MPG, aunque se puede utilizar para comparar dos o más tablas de datos a través de la obtención de un compromiso, no da una medida de asociación como la da el STATIS para analizar la relación entre las tablas.

Figura 8: Correlaciones de las variables con los 2 primeros ejes.

3.1.5. DACP

El análisis de la interestructura, en este método, es un análisis de componentes principales de la tabla de centros de gravedad de cada una de las tablas iniciales

centradas y reducidas (tabla 9).

Tabla 9: Centros de Gravedad de las Tablas

Tiempos	Co	Cu	Fe	Mn	Ni	Pb	Zn
Abril/88	7,863	60,363	3,8909	638,81	40,727	25,4	85,181
Sept/88	7,518	64,090	4,34	543,18	34,181	11,781	94,727
Marz/89	8,145	50,454	4,330	479,72	30,727	11,227	110,09
Abril/89	7,309	46,636	3,178	410,63	35,181	24,545	83,363
Sept/89	10,954	65,909	4,042	468,18	37	22,409	107,90

Tabla 10: Valores propios de la interestructura

Componentes	**Valor propio**	**Diferencia**	**% de Inercia**	**% acumulado**
CP1	2,99478	1,26797	0,427826	0,42783
CP2	1,72682	0,91084	0,246688	0,67451
CP3	0,81597	0,12176	0,116568	0,79108
CP4	0,69421	0,28235	0,099173	0,89025
CP5	0,41186	0,22499	0,058837	0,94909
CP6	0,18686	0,01737	0,026695	0,97579
CP7	0,16950	0,0	0,024214	1,00000

Los dos primeros ejes explican el 67 % de la inercia

Tabla 11: Componentes Principales de la Tabla de Centros de Gravedad

Variables	**CP1**	**CP2**	**CP3**
Co	0,199438	0,336889	0,647992
Cu	0,150276	0,591490	0,016525
Fe	0,494439	0,260724	−0,277410
Mn	−0,077832	0,467868	−0,534597
Ni	−0,404540	0,467905	0,043460
Pb	−0,497193	0,169133	0,354135
Zn	0,525604	0,047195	0,299638

Tabla 12: Tiempos con relación a las componentes principales

TIEMPOS	**CP1**	**CP2**	**CP3**
Abril/88	−1,68625	1,52256	−0,86134
Sept/88	0,97643	0,30937	−1,21493
Marzo/89	1,93592	−1,23661	−0,27094
Sept/89	−1,94609	−1,90840	0,53696
Abril/89	0,71999	1,31307	1,81025

Las tres primeras componentes principales explican el 79 % de la inercia total. En la figura 8 se observa que la primera componente es un contraste entre Zn y Fe, en el lado positivo, con Pb y Ni en el lado negativo.

La segunda componente es una ponderación de todos los metales.

Los meses de marzo/89 y septiembre/88 son altos y positivos en la primera componente y abril/88 con septiembre de 89 con valores negativos.

Con relación a la segunda componente, los momentos de más contaminación son los meses de abril, para ambos años.

Para analizar la intraestructura, vamos a utilizar el segundo criterio propuesto por Bouroche [2], para determinar el sistema de ejes en los cuales se puedan representar las trayectorias de los individuos en el tiempo. El sistema de ejes buscado está constituido por los vectores propios de la matriz

$$V = \sum_{t=1}^{5} V_t,$$

donde V_t es la matriz de varianza covarianza de la tabla (t).

El resultado de la diagonalización de la matriz V es equivalente a la diagonalización de la matriz

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_5 \end{bmatrix},$$

donde $\mathbf{Y}_i$ son las tablas centradas con relación a la respectiva media de cada tabla.

Tabla 13: Valores Propios de la matriz $\mathbf{Y}$

Ejes	Valor Propio	Diferencia	% de Inerca	% acumulado
CP1	3, 24301	1, 44824	0, 463288	0, 46329
CP2	1, 79478	0, 91995	0, 256396	0, 71968
CP3	0, 87483	0, 33814	0, 124976	0, 84466
CP4	0, 53669	0, 26922	0, 076669	0, 92133
CP5	0, 26747	0, 08938	0, 038210	0, 95954
CP6	0, 17809	0, 07296	0, 025442	0, 98498
CP7	0, 10514	0, 00000	0, 015019	1, 00000

Tabla 14: Vectores Propios de la Matriz $\mathbf{Y}$

Variables	CP1	CP2	CP3
Co	0, 531240	0, 010172	−0, 057076
Cu	0, 302536	0, 578022	−0, 022012
Fe	0, 500119	−0, 063658	0, 068113
Mn	0, 443523	0, 278812	−0, 334094
Ni	0, 319015	−0, 230963	0, 701710
Pb	−0, 239734	0, 534134	−0, 063499
Zn	−0, 142064	0, 495334	0, 619331

Con relación a la intraestructura, podría decirse que los tres primeros ejes interpretan el 84,47 %. Pueden además adicionarse gráficas sobre la trayectoria de las estaciones en los años.

4. Conclusiones

- Se presentaron algunos métodos factoriales destinados al análisis de tablas de datos triples o cúbicos, donde una o más de las direcciones puede ser el tiempo.
- El objetivo fundamental de los métodos factoriales analizados es la búsqueda de una estructura común a las tablas de datos y/o la obtención de un compromiso.
- El uso combinado de los métodos factoriales permite alcanzar una mejor interpretación de los resultados, así como complementar un análisis con otro.
- La aplicación de los métodos permitió la obtención de un espacio común entre los 5 muestreos, siendo las variables que los unen `Co`, `Cu, Fe`, `Mn` y `Ni`, lo que muestra que la zona en estudio ha mantenido una contaminación prácticamente estable a lo largo del período analizado, lo que coincide con lo esperado.
- Se encontraron 3 grupos de estaciones: uno formado por la estación 1l, que presenta gran contaminación industrial y urbano–industrial; otro formado por las estaciones 7, 8 y 10, que pueden ser consideradas como intermedias, predominando la contaminación urbano–industrial fundamentalmente, y el último formado por el resto de las estaciones, que son las menos contaminadas.

Bibliografía

[1] Bore G. & Di Ciaccio A. [1989], "Comparisons among three factorial methods for analysing three–mode data", in: R. Coppy et S. Bolasco (eds.), *Multiway Data Analysis*, Elsevier Science Publishers B.V.

[2] Bouroche J. M. [1975], *Analyse des donnés ternaires: la double analyse en composantes principales.* Thése.

[3] Carlier A., Lavit C., Pages M., Pernin M.O. & Turlot J.C. [1989], "A comparative review of methods which handle a set of methods which handle a set of indexed data tables", in: R. Coppy et S. Bolasco (eds.), *Multiway Data Analysis*, Elsevier Science Publishers B.V. North , Amsterdam; p. 85–102.

[4] Cox T.F. & Cox M.A.A. [1994], *Multidimensional Scaling.* Monographs on Statistics and Applied Probability 59. Chapman & Hall.

[5] Flury B. & Riedwyl H. [1988], *Multivariate Statistics. A practical approach.* Chapman and Hall, London.

[6] Gower J. C. [1975], "Generalized Procusto Analysis", *Psychometrika*, **40**, p. 33–51.

[7] Groupe GERI [1996], *L'anayse des donnés évolutives. Méthods et applications.* Editions TECHNIP.

[8] Jackson J.E. [1991], *A User's Guide to Principal Components.* John Wiley & Sons, New York.

[9] Krzanowski W. J. [1979], "Between–Groups Comparison of principal Components". *Journal of the American Statistical Association*, **74**, no. 374, p. 703–707.

[10] Kroonenberg P. M. [1983], *Three–Mode Principal Componentes Analysis*, Leiden, DSWO Press.

[11] Kroonenberg P. M. & De Leeuw J. [1980] "Principal Component Analysis of Three–Mode Data by Means of Alternating Least Squares Algorithms". *Psychometrika*, 45, p. 69–97.

[12] Lacourly N. [1996], "Panorama de Métodos para el Análisis de Datos Longitudinales". *Memorias del Seminario de Capacitación e Investigación. Recolección y Análisis de Datos Longitudinales*, p. 165–177.

[13] Lacourly N. & Lera L. [1997]: "Estudio comparativo entre los métodos STATIS y Procrustes". *Publicaciones Técnicas Universidad de Chile*, MA–97–B–443.

[14] Lacourly N. Lera L. & Farías [1999]: "An Algorithms for the Generalized Procrustes Analysis". Enviado a *Biometrical Journal.*

[15] Lavit Ch. [1988], *Analyse Conjointe de Tableaux Quantitatifs.* Masson, Paris.

[16] Lechevalier F. [1990], "L'Analyse en Composantes conJointes d'une famille de triplets indexés". *Statistique et Analyse des Données,* **15**, no. 2, p. 35–75.

[17] Lera L. & González H. [1999], "Búsqueda de una estructura común en el estudio de la contaminación marina por metales pesados en la bahía de Cienfuegos", Cuba. Enviado a publicar como un reporte de investigación.

[18] Mardia K.V., Kent J.T. y Bibby J.M. [1979], *Multivariate Analysis.* Academic Press, London.

[19] Qannari E. M., Courcoux M., Lejeune M. & Maystre O. [1997]: "Comparaison de Trois Stratégies de Détermination D'un compromis en évaluation sensorielle". *Rev. Statistique*, **XLV** (1), p. 61–74.

[20] Rao C. R. [1964], "The Use and Interpretation of Principal Component Analysis in Applied Research", *Sankhya*, **26**, Ser. A, p. 329–385.

[21] Sánchez S. [1995], *El Análisis de Componentes Principales Comunes.* Tesis de Maestría. Universidad de Guadalajara, México.

[22] Tucker L. R. [1966], "Some Mathematical Notes on Three Mode Factor Analysis", *Psychometrika,* **31**, no. 3, p. 279–311.