Revista Colombiana de Estadística
Volumen 29 N° 1. pp. 17 a 34. Junio 2006

El doble análisis en componentes principales para datos categóricos y su aplicación en un estudio de migración

Double Principal Components Analysis for Categorical Data and its Application to a Migration Study

Raúl Alberto Pérez[1*], Lydia Lera[2†], Ana Boquet[3‡]

[1]Universidad Nacional de Colombia, Escuela de Estadística, Medellín

[2]ICIMAT, La Habana, Cuba y Universidad de Chile, INTA, Santiago

[3]Instituto de Planificación Física, La Habana, Cuba

Resumen

Se hace una adaptación del método doble análisis en componentes principales (DACP) (Bouroche 1975), creado para el análisis de datos cuantitativos de tipo cúbico, a datos categóricos mediante la utilización de la distancia Chi-cuadrado entre perfiles fila y columna de una tabla de contingencia y se realiza una aplicación a un estudio de migración interna en Cuba.

Palabras clave: Doble análisis en componentes principales, datos categóricos, estructura común.

Abstract

We adapted the double principal component analysis (DACP) (Bouroche 1975), developed for the analysis of three-dimensional quantitative data, to categorical data by mean of the Chi-squared distance between rows and columns profile of a contingency table and we carry out an application to a study of internal migration in Cuba.

Key words: Double principal component analysis, Categorical data, Common structure.

*Profesor asistente. E-mail: raperez1@unalmed.edu.co
†E-mail: llera@uec.inta.uchile.cl
‡E-mail: aboquet@icimat.cu

1. Introducción

El doble análisis en componentes principales (DACP) fue introducido por Bouroche (1975) para datos cuantitativos de tipo cúbico, en los que se tienen las mismas variables y los mismos individuos en diversas ocasiones, es decir, se tienen T tablas de datos de orden $n \times p$ donde n es el número de individuos, p es el número de variables y T es el número de ocasiones.

El objetivo principal del método es comparar globalmente las relaciones entre las diferentes variables y la evolución de los individuos.

El método está formado por las siguientes fases:

- El análisis de un fenómeno de evolución global.
- El estudio de la deformación de la nube de puntos alrededor de su centro de gravedad.
- La representación de las evoluciones de los diferentes individuos en un mismo espacio a lo largo del tiempo.

A partir de lo anterior se hace una adaptación del método al caso en que los datos son categóricos, mediante la utilización de la distancia Chi-cuadrado entre perfiles fila y columna de una tabla de contingencia y mediante una recodificación binaria de los datos (Pérez & Lera 2001). Además, se realiza una aplicación del DACP para datos categóricos en un estudio de migración interna en Cuba.

Ramos (1996) hizo una adaptación de otro método de tipo factorial, el método Statis (Structuration des Tableaux a Trois Indices de la Statistique) (Lavit 1988) y lo aplicó a datos provenientes de encuestas. Es bien conocido que muchas investigaciones de corte social, económico, médico, etc., analizan variables cualitativas de tipo longitudinal, de ahí la importancia de adaptar métodos que puedan utilizarse para este tipo de datos.

2. El doble análisis en componentes principales

El DACP se creó para el análisis de datos cuantitativos a los que se les miden las mismas variables sobre los mismos individuos en diferentes instantes.

En el caso en que la tercera dimensión no sea el tiempo, el resto del análisis es posible pero la interpretación de los resultados es mucho más difícil. El dominio de aplicación de este método es entonces más restrictivo que el del Statis, pero se encuentra frecuentemente en la práctica. El objetivo principal es, como en el Statis, comparar globalmente la evolución de los "ligamentos" entre las diferentes variables, como también la evolución de los individuos.

Inicialmente se tienen T-tablas de estudios de orden $n \times p$, X_t con $t = 1, 2, \ldots, T$, formadas por n-individuos a los cuales se les van a medir p-variables en T-instantes diferentes.

Las entradas en una de las tablas anteriores se denotan por $(x_i^j)^{(t)}$, las cuales representan la medida de la variable j-ésima sobre el individuo i-ésimo en el instante t, para $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$ y $t = 1, 2, \ldots, T$.

En el instante t, la variable j-ésima será denotada por el vector de R^n dado por:

$$(x^j)_{n\times 1}^{(t)} = \begin{bmatrix} (x_1^j)^{(t)} \\ (x_2^j)^{(t)} \\ \vdots \\ (x_n^j)^{(t)} \end{bmatrix}_{n\times 1} \quad \text{para} \quad j = 1, 2, \ldots, p \tag{1}$$

y el individuo i-ésimo se denota por el vector de R^p dado por:

$$(e_i)^{(t)'} = \begin{bmatrix} (x_i^1)^{(t)} & (x_i^2)^{(t)} & \cdots & (x_i^p)^{(t)} \end{bmatrix} \quad \text{para} \quad i = 1, 2, \ldots, n \tag{2}$$

Se ponderan los individuos por $p_1, p_2, \ldots, p_n$ y se define la matriz de los pesos de los individuos, como sigue:

$$D_{n\times n} = \begin{bmatrix} p_1 & 0 & \ldots & 0 \\ 0 & p_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & p_n \end{bmatrix}_{n\times n} \quad \text{tal que} \quad \sum_{i=1}^{n} p_i = 1 \tag{3}$$

En el instante t $(t = 1, 2, \ldots, T)$, el centro de gravedad de la tabla X_t, asociada a la matriz de pesos D, es el vector de medias ponderadas de las p-variables definido por:

$$g_{p\times 1}^{(t)} = \begin{bmatrix} (\overline{x}^1)^{(t)} \\ (\overline{x}^2)^{(t)} \\ \vdots \\ (\overline{x}^p)^{(t)} \end{bmatrix}_{p\times 1} \quad \text{donde} \quad (\overline{x}^j)^{(t)} = \sum_{i=1}^{n} p_i (x_i^j)^{(t)} \tag{4}$$

para $j = 1, 2, \ldots, p$.

Para cada instante t $(t = 1, 2, \ldots, T)$ se tiene una nube de puntos definida por la tabla X_t, la cual se denota por:

$$N_I^{(t)} = \{(e_i)^{(t)} : i = 1, 2, \ldots, n\} \tag{5}$$

A continuación se hace una breve explicación de las fases del DACP.

2.1. Estudio de la interestructura (estudio de la nube de centros de gravedad)

El objetivo de la primera fase del DACP es describir la evolución global de la población de individuos estudiados. Esta fase puede mirarse en paralelo con la

primera fase del Statis, es decir, el estudio de la interestructura. Por otra parte, el enfoque es ligeramente diferente puesto que el Statis estudia las semejanzas y las diferencias entre tablas centradas, mientras que el DACP estudia la evolución de las tablas por intermedio de su centro de gravedad.

En esta fase se realiza un ACP a la nube de puntos definida por los centros de gravedad de cada tabla, obteniéndose una imagen euclidiana de las tablas en un espacio dimensional deseado; luego se puede verificar que el primer eje de esta imagen se explica en términos de la evolución global de los tiempos. Los centros de gravedad g^t varían de manera continua en los tiempos a lo largo de este eje.

2.2. Estudio de las T nubes de individuos

En esta fase se estudia la deformación de la nube alrededor de su centro de gravedad. Para ello se efectúan T ACP a cada una de las tablas de datos centradas en los centros de gravedad, con el fin de eliminar el fenómeno de evolución global.

La tabla centrada está dada por la siguiente expresión:

$$Y_t = X_t - 1_n g_t = (I_n - 1_n 1_n' D) X_t$$

Los T ACP permiten interpretar cada uno de los análisis con la ayuda de las representaciones gráficas y estos T ACP proporcionan los dos sistemas de ejes ortogonales. Este análisis evidentemente tiene la dificultad del número de tablas.

Si se denota por q el número de ejes retenidos en los ACP ($q < \text{mín}(p, n)$), se tienen:

- T sistemas de factores principales $((\mathbf{u}_l)^{(t)})_{l=1,2,\ldots,q}$, (vectores de tamaño p) para $t = 1, 2, \ldots, T$
- T sistemas de componentes principales $((\mathbf{c}^l)^{(t)})_{l=1,2,\ldots,q}$, (vectores de tamaño n) para $t = 1, 2, \ldots, T$

2.3. Estudio de la intraestructura

La última fase del método responde a su objetivo principal, la representación de los individuos en un espacio común a través del tiempo.

Bouroche (1975) propone 4 criterios para la selección de los ejes, que miden la proximidad entre los sistemas de ejes. Nos referiremos al segundo criterio, cuyo objetivo es maximizar la inercia de la muestra de nubes proyectadas, que se traduce en la resolución del problema de optimización siguiente:

$$\max_{v_1, v_2, \ldots, v_q} \sum_{t=1}^{T} \sum_{l=1}^{q} v_l' V_t v_l = \max_{v_1, v_2, \ldots, v_q} \sum_{l=1}^{q} v_l' V v_l = \sum_{l=1}^{q} V_t \tag{6}$$

La solución de este problema se basa en un ACP.

Estos criterios de selección de ejes se basan en 2 índices que describen la calidad de la imagen euclidiana compromiso (Groupe Geri 1996).

2.4. Compromiso e interpretación de las trayectorias de los individuos

Para determinar el compromiso y los ejes se seleccionó el criterio definido anteriormente.

El sistema de ejes está formado por los vectores propios de la matriz,

$$V = \sum_{l=1}^{q} V_t$$

donde V_t es la matriz de varianzas covarianzas de la tabla t.

El compromiso representa la suma de las correlaciones entre variables de una misma tabla. Las trayectorias de los individuos se representan proyectando los individuos definidos por las tablas sobre el sistema de ejes determinado por el criterio seleccionado.

3. El doble análisis en componentes principales cuando los datos son categóricos

En este caso, las tablas de datos son tablas de contingencia formadas por el cruce de dos variables cualitativas con K_1 y K_2 categorías.

Dada una serie de tablas de contingencias $C_1, C_2, \ldots, C_T$, formadas por individuos que poseen las características i y j, se denotan por $1, 2, \ldots, K_1$, las categorías de la primera variable y por $1, 2, \ldots, K_2$, las categorías de la segunda variable. Sea C_t la tabla dada por:

$$C_t = \begin{bmatrix} k_{11}^t & k_{12}^t & \ldots & k_{1K_2}^t \\ k_{21}^t & k_{22}^t & \ldots & k_{2K_2}^t \\ \vdots & \vdots & \ddots & \vdots \\ k_{K_1 1}^t & k_{K_1 2}^t & \ldots & k_{K_1 K_2}^t \end{bmatrix} \tag{7}$$

donde k_{ij}^t es el número de individuos que satisfacen simultáneamente la característica i de la primera variable y la característica j de la segunda variable, para $i = 1, 2, \ldots, K_1$ y $j = 1, 2, \ldots, K_2$.

Se denota por $k_{\cdot\cdot}^t$ el número total de individuos, es decir,

$$k_{\cdot\cdot}^t = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} k_{ij}^t$$

y por $k_{i\cdot}^t$ y $k_{\cdot j}^t$, el número total de individuos en la categoría i de la primera variable (filas) y el número total de individuos en la categoría j de la segunda variable (columnas), respectivamente, para $i = 1, 2, \ldots, K_1$ y $j = 1, 2, \ldots, K_2$, es decir, en el instante t se tiene:

$$K_{i\cdot}^t = \sum_{j=1}^{K_2} k_{ij}^t \quad \text{y} \quad K_{\cdot j}^t = \sum_{i=1}^{K_1} k_{ij}^t$$

Ahora se denotan por:

$$D_1^t = \begin{bmatrix} k_{1\cdot}^t & 0 & \dots & 0 \\ 0 & k_{2\cdot}^t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & k_{K_1\cdot}^t \end{bmatrix} \quad \text{y por} \quad D_2^t = \begin{bmatrix} k_{\cdot 1}^t & 0 & \dots & 0 \\ 0 & k_{\cdot 2}^t & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & k_{\cdot K_2}^t \end{bmatrix} \tag{8}$$

las matrices diagonales de los efectos marginales de las dos variables.

Sea F_t la matriz cuyas entradas son las frecuencias relativas de cada casilla (i, j) de la tabla t, es decir,

$$F_t = [f_{ij}^t] = \left[\frac{1}{k_{\cdot\cdot}} C_t\right] = \left[\frac{k_{ij}^t}{k_{\cdot\cdot}}\right] = \begin{bmatrix} f_{11}^t & f_{12}^t & \cdots & f_{1K_2}^t \\ f_{21}^t & f_{22}^t & \cdots & f_{2K_2}^t \\ \vdots & \vdots & \ddots & \vdots \\ f_{K_1 1}^t & f_{K_1 2}^t & \cdots & f_{K_1 K_2}^t \end{bmatrix}_{K_1 \times K_2}$$

y se denota por $f_{i\cdot}^t$ y $f_{\cdot j}^t$ las frecuencias marginales tanto por filas como por columnas en la ocasión t, es decir,

$$f_{i\cdot}^t = \sum_{j=1}^{K_2} f_{ij}^t = \sum_{j=1}^{K_2} \frac{k_{ij}^t}{k_{\cdot\cdot}} = \frac{k_{i\cdot}^t}{k_{\cdot\cdot}} \quad \text{y} \quad f_{\cdot j}^t = \sum_{j=1}^{K_1} f_{ij}^t = \sum_{j=1}^{K_1} \frac{k_{ij}^t}{k_{\cdot\cdot}} = \frac{k_{\cdot j}^t}{k_{\cdot\cdot}} \tag{9}$$

3.1. Fases del método

3.1.1. Estudio de la interestructura

En esta fase se calculan los centros de gravedad de las nubes.

El centro de gravedad de la nube de K_1 puntos formada por los perfiles fila, g_F, es el vector

$$g_F^t = \frac{1}{k_{\cdot\cdot}}((D_1^t)^{-1} C_t)' D_1^t 1 = \left[\frac{k_{\cdot 1}^t}{k_{\cdot\cdot}} \quad \frac{k_{\cdot 2}^t}{k_{\cdot\cdot}} \quad \dots \quad \frac{k_{\cdot K_2}^t}{k_{\cdot\cdot}}\right]' = \left[f_{\cdot 1}^t \quad f_{\cdot 2}^t \quad \cdots \quad f_{\cdot K_2}^t\right]' \tag{10}$$

que son los perfiles marginales de las filas, $g_F \in R^{K_2}$.

Recíprocamente, el centro de gravedad de la nube de K_2 puntos formada por los perfiles columna, g_C, es el vector

$$g_C^t = \frac{1}{k_{\cdot\cdot}}((D_2^t)^{-1} C_t)' D_2^t 1 = \left[\frac{k_{1\cdot}^t}{k_{\cdot\cdot}} \quad \frac{k_{2\cdot}^t}{k_{\cdot\cdot}} \quad \dots \quad \frac{k_{K_1\cdot}^t}{k_{\cdot\cdot}}\right]' = \left[f_{1\cdot}^t \quad f_{2\cdot}^t \quad \cdots \quad f_{K_1\cdot}^t\right]' \tag{11}$$

que son los perfiles marginales de las columnas, $g_C \in R^{K_1}$.

De la independencia estadística de las características i y j se tiene:

$$\frac{k_{ij}^t}{k_{i\cdot}^t} = \frac{k_{\cdot j}^t}{k_{\cdot\cdot}^t} \quad \text{y} \quad \frac{k_{ij}^t}{k_{\cdot j}} = \frac{k_{i\cdot}^t}{k_{\cdot\cdot}^t}$$

Las nubes son reducidas cada una a un punto correspondiente a los centros de gravedad respectivos.

En el caso particular en que se tienen dos variables cualitativas con K_1 y K_2 categorías respectivamente, se usará la distancia Chi-cuadrado para definir la proximidad entre dos filas o dos columnas, como se hace en el análisis factorial de correspondencias (AFC).

La distancia Chi-cuadrado entre dos categorías i , i', de una variable categórica se define como:

$$d^2(i,i') = \sum_{j=1}^{K_2} \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{i'j}}{f_{i'\cdot}} \right)^2$$

para $i, i' = 1, 2, \ldots, K_1$, que representa la suma de los cuadrados de las diferencias entre las coordenadas de los puntos i, i', ponderadas por su respectiva frecuencia marginal. De manera similar, se define la distancia Chi-cuadrado entre dos categorías j, j' de la segunda variable.

Como el ACP utiliza la distancia euclidiana, para que estas dos distancias sean equivalentes se utiliza la transformación de los datos a:

$$x_{ij} = \frac{f_{ij}}{f_{i\cdot}\sqrt{f_{\cdot j}}}$$

Con los centros de gravedad calculados, se conforman las matrices de centros de gravedad:

$$G_F = \begin{bmatrix} g_F^1 & g_F^2 & \ldots & g_F^T \end{bmatrix} \quad \text{y} \quad G_C = \begin{bmatrix} g_C^1 & g_C^2 & \ldots & g_C^T \end{bmatrix}$$

y la matriz de varianzas y covarianzas de una nube de puntos centrada en los centros de gravedad es:

$$V = X'DX - g_F g_F' \quad \text{donde} \quad D = \begin{bmatrix} \frac{D_1}{k_{\cdot\cdot}} \end{bmatrix}$$

Ahora, de la descomposición espectral de V se obtienen U y L tales que:

$$V = ULU' \quad \text{con} \quad U'U = I_p$$

y de ahí los puntajes para los componentes principales ($Z = XU$). Luego el análisis se prosigue como en el ACP. De igual forma se procede para los perfiles columna.

Como se parte de T tablas de contingencia es posible realizar 2 ACP a cada una de las tablas, para los perfiles fila y para los perfiles columna. A cada una de estas matrices se les efectúa un ACP para estudiar la evolución de las tablas por intermedio de los centros de gravedad.

3.1.2. Estudio de las T nubes de puntos

En esta fase se estudia la deformación de la nube alrededor de su centro de gravedad, se efectúa un ACP a cada una de las T tablas centradas en los centros de gravedad, con el fin de eliminar el fenómeno de evolución global. Se procede como en el epígrafe anterior para cada una de las tablas.

De aquí se tienen $2T$ sistemas de ejes ortogonales:

- T sistemas de factores principales (de tamaño K_2), $t = 1, 2, \ldots, T$, que son los vectores propios de la matriz MV_t, asociados a los q mayores valores propios, $q < \text{mín}(K_2, K_1)$, ($V_t = X_t' X_t$, es la matriz de productos internos entre perfiles columna que representan la estructura interna entre las columnas).
- T sistemas de componentes principales $((c^l)^t)_{l=1,\ldots,q}$ (vectores de tamaño K_1), $t = 1, \ldots, T$, son los vectores propios de la matriz $W_t D$, asociados a los q mayores valores propios ($W_t = X_t X_t'$, es la matriz de productos internos entre perfiles fila que representan la estructura interna entre las filas).

Los factores principales de los perfiles fila se obtienen calculando los vectores propios de la matriz $D_2^{-1} C_t' D_1^{-1} C_t$ y los componentes principales son los vectores propios de la matriz $D_1^{-1} C_t D_2^{-1} C_t'$ normalizados para $a'\left(\frac{D_1}{k_{..}}\right)a = \lambda$. Similarmente se realiza para los perfiles columna.

3.1.3. Estudio de la intraestructura

Esta fase, como se dijo en el epígrafe anterior, responde al objetivo principal del DACP, la representación de los individuos en un espacio común a través del tiempo. Bouroche propone buscar dos sistemas de q vectores ortogonales que resuman lo mejor posible (según ciertos criterios) las semejanzas o diferencias entre los sistemas de ejes.

Los dos sistemas de ejes óptimos serán:

- $(V_l)_{l=1,\ldots,q}$; los factores principales
- $(d_l)_{l=1,\ldots,q}$; las componentes principales

Las trayectorias se obtienen proyectando los puntos fila (o columna) sobre el nuevo sistema de ejes.

3.1.4. Compromiso e interpretación de las trayectorias

Al igual que en el Statis (Ramos 1996), para dos de los criterios de selección de los ejes, el compromiso será equivalente a una tabla de contingencia promedio.

La correlación variable-factor actuará como la posición de las columnas en el plano-compromiso y los individuos-compromiso tendrán la posición de las filas en el plano.

Las trayectorias serán las distintas posiciones de los puntos fila a través de la serie de tablas. El primer eje se interpreta en general en términos de evolución en el tiempo.

4. Aplicación del doble análisis de componentes principales para datos categóricos en un estudio de migración interna en Cuba

El movimiento migratorio dentro de los países es un tema de gran interés de los especialistas de las más diversas esferas en todas partes del mundo, por sus efectos en la distribución y composición de la población y por su sensibilidad a los cambios socioeconómicos. La realización de estudios sobre los movimientos migratorios de la población se dificultan porque no siempre se dispone de estadísticas seguras sobre éstos.

En Cuba se ha mantenido un estudio sistemático de las migraciones a lo largo del período revolucionario y se han realizado estudios de migraciones a diferentes escalas, desde nacionales hasta estudios de detalles en zonas de interés, utilizándose como fuentes de información (censos, registros de población y encuestas levantadas para estudios específicos) las que se tengan disponibles en el momento del estudio (Boquet 1997).

El estudio de las migraciones a escala municipal es importante ya que en esa unidad territorial se pueden determinar las causas de los movimientos con bastante certeza, a la vez que se pueden tomar decisiones de planeamiento si se considera modificar un comportamiento migratorio indeseable para el territorio.

La migración interna se encuentra íntimamente ligada a los procesos de transformaciones económicas y sociales de los territorios, ya sea atrayendo migrantes hacia donde se dan mejores condiciones, o con la salida de migrantes desde los territorios más deprimidos. Una medida del efecto de la migración en la población de un territorio está dada por la tasa migratoria promedio de entrada y de salida.

Para la aplicación del DACP para datos categóricos se utilizarán como fuentes las tasas migratorias de entrada y salida, calculadas a partir de las bases de datos de la Oficina Nacional de Estadística de Cuba. Para el estudio, las tasas se calcularon por trienios, 1986-1988, 1989-1991, 1992-1994, 1995-1997 y 1998-2000, para cada municipio. Se utilizaron los 169 municipios del país.

La tasa migratoria de entrada se define como el cociente entre el número de personas que entran a un territorio y el número total de habitantes de ese territorio dividido por mil, y la tasa migratoria de salida se define como el cociente entre el número de personas que salen de un territorio y el número total de habitantes de ese territorio dividido por mil.

En las migraciones, en ocasiones, el dato categórico tiene un significado más útil que el dato continuo ya que con este último se mezclan casos que no son convenientes para los objetivos trazados. En este trabajo se categorizaron las tasas de entrada y salida de los municipios en tres categorías: 1-Baja, 2-Media y 3-Alta.

Se aplicó el DACP adaptado a datos categóricos mediante el uso de la distancia Chi-cuadrado entre perfiles fila y columna, para lo que se elaboró un algoritmo en el sistema estadístico SAS. Las matrices de datos se transformaron en tablas de contingencia.

4.1. Resultados y discusión

4.1.1. Fase 1

En esta fase se realizan ACP a las matrices de centros de gravedad tanto para perfiles fila (tasa de entrada) como para perfiles columna (tasa de salida).

En la tabla 1 se tienen los vectores y valores propios del ACP de la tabla de centros de gravedad para los perfiles fila (GPF), que representan la tasa de entrada, con su respectivo porcentaje de varianza explicada. Se observa que la tasa de entrada baja se opone a la tasa de entrada media y alta.

Del DACP de la matriz GPF se obtienen las coordenadas, contribuciones y cosenos cuadrados de los individuos (filas de GPF, que representan a las 5 tablas de perfiles fila) sobre los ejes factoriales, las cuales aparecen en la tabla 2, para los dos primeros ejes.

Tabla 1: Valores y vectores propios de GPF.

Variables	Vectores propios	
Variables - Vector	Vector-1	Vector-2
1	−0.59	0.20
2	0.58	−0.57
3	0.56	0.80
Valor propio	2.83	0.17
Porcentaje	0.942	0.055
Porc. acumulado	0.942	0.997

Tabla 2: Contribución de los individuos, perfiles fila.

Individuos		Coordenadas		Contribución		Cosenos cuadrados	
Número	Distancia	1	2	1	2	1	2
T1	1.81	1.31	−0.27	12.1	8.7	0.95	0.04
T2	6.3	1.26	0.21	11.2	5.1	0.97	0.03
T3	1.35	1.06	0.47	7.9	26.4	0.83	0.16
T4	0.76	-0.56	−0.66	2.2	52.1	0.42	0.57
T5	9.46	-3.07	0.25	66.5	7.6	0.99	0.01

Similarmente se tienen todos los resultados para la matriz de centros de gravedad de perfiles columna (GPC). Las coordenadas, contribuciones y cosenos cuadrados de las columnas de la matriz GPC, que representan a las 5 tablas de datos de perfiles columna, sobre los ejes factoriales, aparecen en las tablas 3 y 4.

De las tablas 1 y 2 del ACP de la tabla GPF, se observa lo siguiente:

Los trienios 86-88, 89-91 y 92-94 se caracterizan por presentar una tasa alta de entrada en los municipios en general, es decir, son grandes receptores. El primer eje representa hacia la derecha tasa alta de entrada y hacia la izquierda tasa baja.

El trienio 95-97 presenta una tasa media de entrada puesto que está cerca del origen, mientras que el trienio 98-2000 presenta una tasa de entrada baja.

Tabla 3: Valores y vectores propios de GPC.

Variables	Vectores propios	
Variables-Vector	Vector-1	Vector-2
1	−0.58	0.02
2	0.57	0.72
3	0.57	−0.69
Valor propio	2.91	0.08
Porcentaje	0.971	0.026
Porc. acumulado	0.9971	0.997

Tabla 4: Contribución de los individuos, perfiles columna.

Individuos		Coordenadas		Contribución		Cosenos cuadrados	
Número	Distancia	1	2	1	2	1	2
T1	0.34	0.56	0.17	2.1	7.7	0.91	0.09
T2	2.71	1.59	−0.43	17.3	47.1	0.93	0.07
T3	1.91	1.37	0.09	12.9	2.1	0.98	0.00
T4	0.30	-0.40	0.36	1.1	33.4	0.54	0.44
T5	9.73	-3.11	−0.20	66.5	9.7	1.00	0.00

De los resultados del ACP de la tabla GPC, se observa lo siguiente:

Los trienios 89-91 y 92-94 se caracterizan por presentar globalmente una tasa alta de salida en los municipios, lo que los convierte en trienios con municipios que son grandes emisores fundamentalmente. Los trienios 86-88 y 95-97 presentan una tasa media y el trienio 98-2000 una tasa baja.

4.1.2. Fase 2

En esta fase se realizan 5 ACP a las matrices centradas con relación a sus centros de gravedad para perfiles tanto fila como columna.

En las tablas 5 y 6 se tienen los dos primeros vectores correspondientes a los ACP de las 5 matrices, centradas con relación a sus centros de gravedad, para perfiles fila y para perfiles columna.

En las tablas 7 y 8 se tienen los valores propios de los 5 ACP correspondientes, tanto para las matrices de perfiles fila como para perfiles columna centradas con relación a su centro de gravedad, y su respectivo porcentaje de varianza explicada.

En las tablas 9 y 10 se tienen las coordenadas, contribuciones y cosenos cuadrados de los individuos (tanto para tasas de entrada como de salida) sobre los ejes factoriales, para los 5 instantes diferentes.

De los resultados obtenidos se corrobora que el comportamiento de los periodos analizados por separado es bastante similar, predominando una componente principal representada por una alta y media tasa de entrada en el caso de los perfiles fila, y en el caso de los perfiles columna los periodos 89-91 y 98-2000 se comportan de modo diferente al resto, predominando una tasa media y baja de salida.

Tabla 5: Vectores propios de los 5 ACP para perfiles fila.

	T1		T2		T3		T4		T5	
Vbles.	Vec–1	Vec–2	Vec–1	Vec–2	Vec–1	Vec–2	Vec–1	Vec–2	Vec–1	Vec–2
1	-0.50	-0.71	0.44	0.74	0.69	-0.37	-0.65	0.02	0.58	-0.54
2	0.50	0.71	0.51	-0.67	0.06	0.88	0.53	0.73	0.56	0.814
3	-0.71	0.00	-0.74	-0.02	-0.72	-0.30	0.55	-0.68	-0.59	0.24

Tabla 6: Vectores propios de los 5 ACP para perfiles columna.

	T1		T2		T3		T4		T5	
Vbles	Vec–1	Vec–2	Vec–1	Vec–2	Vec–1	Vec–2	Vec–1	Vec–2	Vec–1	Vec–2
1	-0.66	0.05	0.62	-0.51	-0.74	-0.14	-0.63	0.08	0.58	-0.43
2	0.49	0.78	0.30	0.86	0.24	0.85	0.54	0.76	0.56	0.82
3	0.56	-0.63	-0.73	-0.07	0.62	-0.51	0.56	-0.64	-0.59	0.31

Tabla 7: Valores propios de los 5 ACP para los perfiles fila.

Instante	Número	Valor propio	Porcentaje	Porc. acumulado
1	1	1.982	66.1	66.1
1	2	1.018	33.9	100.0
2	1	1.836	61.2	61.2
2	2	1.164	38.8	100.0
3	1	1.721	57.4	57.4
3	2	1.279	42.6	100.0
4	1	2.353	78.4	78.4
4	2	0.647	21.6	100.0
5	1	2.844	94.8	94.8
5	2	0.156	5.1	100.0

Tabla 8: Valores propios de los 5 ACP para los perfiles columna.

Instante	Número	Valor propio	Porcentaje	Porc. acumulado
1	1	2.260	75.3	75.3
1	2	0.739	24.6	100.0
2	1	1.871	62.6	62.6
2	2	1.121	37.4	100.0
3	1	1.767	58.9	58.9
3	2	1.233	41.1	100.0
4	1	2.546	84.8	84.8
4	2	0.453	15.1	100.0
5	1	2.827	94.2	94.2
5	2	0.173	5.7	100.0

Tabla 9: Contribuciones de los individuos, perfiles fila.

	Individuos		Coordenadas		Contribuciones		Cosenos cuadrados	
Instan.	Núm.	Dist.	1	2	1	2	1	2
1	1	2.72	1.19	-1.14	23.8	42.9	0.52	0.48
1	2	2.34	0.79	1.31	10.4	56.2	0.26	0.73
1	3	3.94	-1.98	-0.17	65.8	0.9	0.99	0.01
2	1	2.68	0.98	1.31	17.3	49.3	0.35	0.64
2	2	2.65	0.94	-1.33	16.0	50.7	0.33	0.66
2	3	3.67	-1.92	0.02	66.7	0.0	1.00	0.00
3	1	3.13	1.49	-0.95	43.0	23.7	0.71	0.29
3	2	2.57	0.21	1.59	0.9	65.8	0.02	0.98
3	3	3.30	-1.71	-0.63	56.2	10.5	0.87	0.12
4	1	4.31	-2.04	-0.38	58.9	7.4	0.96	0.03
4	2	1.40	0.38	1.12	2.0	64.6	0.10	0.89
4	3	3.29	1.66	-0.73	39.0	27.6	0.84	0.16
5	1	1.23	0.98	-0.51	11.3	55.3	0.79	0.21
5	2	2.14	1.39	0.45	22.6	44.0	0.90	0.09
5	3	5.64	-2.37	0.05	66.0	0.6	0.99	0.01

Tabla 10: Contribuciones de los individuos, perfiles columna.

	Individuos		Coordenadas		Contribuciones		Cosenos cuadrados	
Instan.	Núm.	Dist.	1	2	1	2	1	2
1	1	4.04	-1.95	-0.48	56.1	10.5	0.94	0.06
1	2	1.52	0.24	1.21	0.9	65.8	0.04	0.96
1	3	3.44	1.71	-0.72	43.0	23.7	0.85	0.15
2	1	2.86	1.24	-1.15	27.2	39.5	0.54	0.46
2	2	2.43	0.67	1.41	8.0	58.7	0.19	0.81
2	3	3.71	-1.91	-0.25	64.8	1.9	0.98	0.02
3	1	3.40	-1.76	-0.56	58.3	8.4	0.91	0.09
3	2	2.50	0.30	1.55	1.7	64.9	0.04	0.96
3	3	3.11	1.46	-0.99	40.0	26.7	0.68	0.32
4	1	4.15	-1.98	-0.45	51.6	15.1	0.95	0.05
4	2	0.91	0.06	0.95	0.1	66.6	0.01	0.99
4	3	3.94	1.92	-0.50	48.4	18.3	0.94	0.06
5	1	1.14	0.92	-0.54	9.9	56.8	0.74	0.25
5	2	2.30	1.44	0.47	24.5	42.2	0.91	0.09
5	3	5.57	-2.36	0.07	65.6	1.1	0.99	0.01

4.1.3. Fase 3

En esta fase se realiza un ACP a los compromisos de perfiles fila y perfiles columna, para obtener un espacio de representación común de los individuos y de las variables; estos compromisos están dados por las tablas $V = \sum[V(t)]$ y $W = \sum[W(t)]$. Igualmente en esta fase se tienen las dos nubes de K_1T-puntos en $\mathbf{R}^{K_2}$ (15 puntos en $\mathbf{R}^3$) y de K_2T-puntos en $\mathbf{R}^{K_1}$ (15 puntos en $\mathbf{R}^3$), a las cuales se les realizan los respectivos ACP para obtener las trayectorias de los respectivos individuos (perfiles fila y perfiles columna) en el espacio de representación común obtenido.

En la tabla 11 se tienen los dos primeros vectores y valores propios del ACP de la nube de puntos formada por las 5 tablas para las categorías de la variable tasa de entrada, con su porcentaje de varianza explicada, y la tabla 12 muestra las coordenadas, contribuciones y cosenos cuadrados de los individuos (tasa de entrada) sobre los dos primeros ejes factoriales, para los 5 instantes diferentes.

Tabla 11: Valores y vectores propios de la tasa de entrada.

Variables	Vectores propios	
Variables - Vector	Vector-1	Vector-2
1	0.56	−0.61
2	0.40	0.80
3	−0.72	−0.02
Valor propio	1.72	1.07
Porcentaje	0.59	0.36
Porc. acumulado	0.59	0.95

En el gráfico de las trayectorias (figura 1) se observan 3 grupos formados por los receptores altos, medios y bajos respectivamente. En el grupo de los receptores altos, la trayectoria muestra que del primer periodo al segundo periodo hay un crecimiento de la tasa de entrada; del segundo al tercero hay un decrecimiento; del tercero al cuarto vuelve a haber un incremento, y del cuarto al quinto la tasa de entrada decrece nuevamente. Un análisis similar se puede hacer para las otras dos trayectorias.

Tabla 12: Contribuciones de los individuos, tasa de entrada.

	Individuos		Coordenadas		Contribuciones		Cosenos cuadrados	
Instan.	Núm.	Dist.	1	2	1	2	1	2
1	1	3.61	0.90	-1.03	3.1	6.6	0.44	0.56
1	2	2.27	0.55	1.12	1.2	7.9	0.19	0.80
1	3	3.12	-1.26	-0.65	6.0	2.7	0.66	0.18
2	1	3.35	1.65	-1.45	10.3	13.1	0.50	0.39
2	2	2.45	0.37	0.58	0.5	2.1	0.22	0.52
2	3	3.20	-1.05	0.17	4.2	0.2	0.97	0.03
3	1	4.07	0.80	-0.86	2.4	21.5	0.16	0.84
3	2	2.32	0.75	1.28	2.2	10.2	0.26	0.74
3	3	2.62	-1.17	-0.36	5.1	0.8	0.70	0.07
4	1	5.47	0.60	-0.99	1.4	6.1	0.26	0.71
4	2	0.99	0.68	0.36	1.7	11.7	0.20	0.79
4	3	2.54	-0.57	1.29	1.2	10.3	0.17	0.83
5	1	1.59	0.61	-0.08	1.4	0.0	0.96	0.02
5	2	1.45	0.96	0.98	3.5	6.0	0.47	0.49
5	3	5.96	-3.84	-0.38	55.8	0.9	0.95	0.01

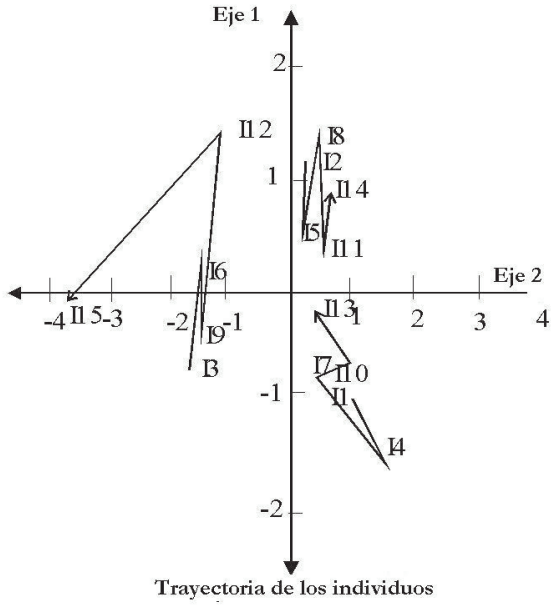

Figura 1: Gráfica de la tabla 12, tasas de entrada.

En la tabla 13 se tienen los dos primeros vectores y valores propios del ACP de la nube de puntos formada por las 5 tablas para las categorías de la variable tasa de salida, con su porcentaje de varianza explicada, y en la tabla 14 se muestran las coordenadas, contribuciones y cosenos cuadrados de los individuos (tasa de salida) sobre los dos primeros ejes factoriales, para los 5 instantes diferentes.

Tabla 13: Valores y vectores propios de la tasa de salida.

Variables	Vectores propios	
Variables - Vector	Vector-1	Vector-2
1	−0.73	−0.08
2	0.31	0.84
3	0.60	−0.53
Valor Propio	1.72	1.14
Porcentaje	0.57	0.38
Porc. Acumulado	0.57	0.95

Tabla 14: Contribuciones de los individuos, tasa de salida.

	Individuos		Coordenadas		Contribuciones		Coseno cuadrados	
Instan.	Núm.	Dist.	1	2	1	2	1	2
1	1	4.31	-1.85	-0.95	13.2	5.3	0.79	0.21
1	2	1.28	0.01	1.13	0.0	7.4	0.00	1.00
1	3	1.77	1.28	-0.04	6.4	0.0	0.93	0.00
2	1	5.86	-2.25	-0.79	19.7	3.7	0.87	0.11
2	2	1.04	-0.00	1.02	0.0	6.1	0.00	1.00
2	3	3.64	0.83	-1.53	2.7	13.6	0.19	0.64
3	1	3.89	-1.85	-0.67	13.2	2.7	0.88	0.12
3	2	2.14	0.32	1.43	0.4	12.0	0.05	0.95
3	3	1.33	0.69	-0.74	1.9	3.2	0.36	0.41
4	1	1.54	-1.18	-0.35	5.4	0.7	0.91	0.08
4	2	1.81	0.28	1.31	0.3	10.0	0.04	0.95
4	3	3.68	1.80	0.67	12.6	2.6	0.88	0.12
5	1	0.42	-0.42	0.49	0.7	1.4	0.43	0.57
5	2	1.21	-0.13	1.08	0.1	6.8	0.01	0.96
5	3	11.08	2.46	-2.05	23.6	24.6	0.55	0.38

En el gráfico de las trayectorias (figura 2) se observan 3 grupos formados por los emisores altos, medios y bajos respectivamente. En el grupo de los emisores bajos, la trayectoria muestra una evolución, es decir, hay un ligero aumento de un periodo a otro. Un análisis similar se puede hacer para las otras dos trayectorias.

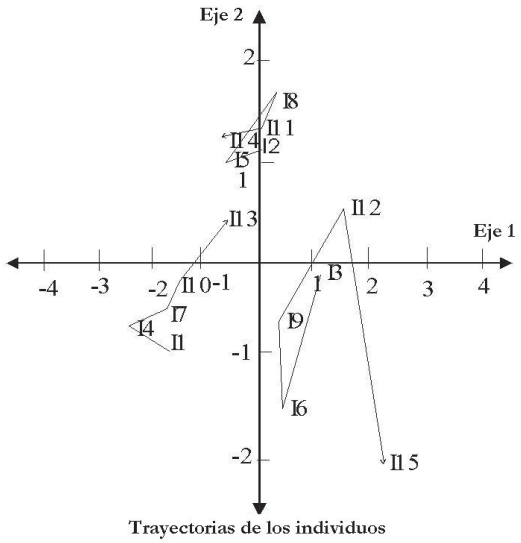

FIGURA 2: Gráfica de la tabla 14, tasas de salida.

5. Conclusiones

1. Se hizo una adaptación del método DACP, creado para datos cuantitativos, a datos de tipo categórico, mediante una transformación simple de la información original, lo cual permite utilizar los criterios, la geometría y la interpretación del DACP.

2. Se emplea el método del DACP, creado para el análisis de datos cuantitativos de tipo cúbico, en el caso en que los datos son categóricos, mediante la utilización de la distancia Chi-cuadrado entre perfiles fila y columna de una tabla de contingencia con un conjunto de datos reales de un estudio migratorio en Cuba, lo cual permitió analizar el comportamiento migratorio de los 169 municipios, comparar globalmente las diferentes categorías con relación a las tasas de entrada y salida durante los 5 trienios estudiados.

3. Se obtuvo que en los periodos 86-88, 89-91 y 92-94 la tasa de entrada en los municipios fue alta; en el periodo 95-97 fue media y en el periodo 98-2000 fue baja.

4. Se obtuvo que en los periodos 89-91 y 92-94 se presentaron las tasas más altas de salida en los municipios; en los periodos 86-88 y 95-97 fueron medias, y en el periodo 98-2000 fueron bajas.

Recibido: mayo de 2005
Aceptado: abril de 2006

Referencias

Boquet, A. (1997), Migraciones internas. Estudio descriptivo de las migraciones internas de Cuba de 1989 a 1996, Technical report, Instituto de Planeación, La Habana, Cuba.

Bouroche, J. (1975), Analyse des donnés ternaires: Le double Analyse en composantes principales, Thèse de 3ème cycle, Université de Paris VI.

Groupe Geri (1996), 'L'analyse des donnés évolutives. Méthodes et applications', *Editions Technip* .

Lavit, C. (1988), *Analyse Conjointe de Tableaux Quantitatifs*, Masson, Paris.

Pérez, R. A. & Lera, L. (2001), Doble análisis de componentes principales para datos categóricos, *in* 'Memorias de la IV ITLA, Fourth Italian-Latin American Conference on Applied and Industrial Mathematics'.

Ramos, J. (1996), Una aplicación del método Statis a datos longitudinales, *in* O. Barbary, ed., 'Memorias del Seminario de Capacitación e Investigación. Recolección y Análisis de Datos Longitudinales', Universidad Nacional de Colombia. Departamento de Estadística, Orstom & Presta, Bogotá, pp. 179–202.