Universitas Scientiarum
ISSN 0122-7483 (versión impresa)
ISSN 2027-1352 (versión en línea)

Disponible en línea en:
www.javeriana.edu.co/universitas_scientiarum
2011, Vol. 16 N° 3: 254-262

SICI: 2027-1352(201109/12)16:3<254:ADTEMEPYS>2.0.TS;2-U

Comunicación corta
doi: 10.11144/Javeriana.SC16-3.uoms

Aplicación de técnicas estadísticas multivariadas en perfilación y segmentación

Milton Januario Rueda Varon[1], Luz Marina Moya Moya[2], Moisés Aranda Silva[2]

[1]Grupo de Física Matemática, [2]Departamento de Matemáticas. Facultad de Ciencias, Pontificia Universidad Javeriana, Bogotá, D.C., **Colombia**.

* milton.rueda@javeriana.edu.co

Recibido: 01-09-2011; Aceptado: 25-10-2011

Resumen

Objetivo. Utilizar la información correspondiente a los factores determinados mediante el análisis de correspondencias, con el fin de perfilar comportamientos de las variables a analizar para luego realizar una segmentación natural, facilitando la interpretación y presentación de los resultados. **Materiales y métodos**. Se presenta un nuevo enfoque de perfilación y segmentación utilizando análisis de correspondencias y clasificación. **Resultados.** Utilizando la metodología aquí presentada se optimiza la determinación de los segmentos y la perfilación de un conjunto de variables. Este procedimiento permite a investigadores de diferentes disciplinas presentar e interpretar los resultados de una investigación de manera eficiente. **Conclusiones**. El procedimiento expuesto permite una sencilla y rápida interpretación del análisis en muchas variables, haciendo fácil su uso e implementación.

Palabras clave: perfilación, segmentación, técnicas multivariadas, correspondencias.

Abstract

Use of multivariate statistical techniques in profiling and segmentation. Objective.To use the information about the factors identified, by means of correspondence analyses to determine the behavior of the variables to be used to then make a natural segmentation, thus facilitating the interpretation and presentation of results. **Materials and methods**. We present a new profiling and segmentation approach by using correspondence analyses and classification. **Results.** By using this methodology we can optimize the determination of segments and the profiling of a set of variables. This procedure allows researchers of different disciplines to present and interpret the results of their research efficiently. **Conclusions**. The procedure here described enables a simple and quick interpretation of the analysis on many variables, rendering its use and implementation easy.

Key words: profiling, segmentation, multivariate techniques, correspondences.

Resumo

Aplicação de técnicas estatísticas multivariadas em análise de perfis e segmentação. Objetivo. Utilizar a informação que corresponde aos fatores determinados pela análise de correspondência, a fim de obter os perfis dos comportamentos das variáveis a serem analisadas e em seguida, fazer uma segmentação natural, facilitando a interpretação e apresentação dos resultados. **Materiais e métodos**. Apresenta-se uma nova abordagem para a obtenção de perfis e segmentação usando análise de correspondência e classificação. **Resultados**. Utilizando a metodologia aqui apresentada se aperfeiçoa a determinação dos segmentos e dos perfis de um conjunto de variáveis. Este procedimento permite aos pesquisadores de diferentes disciplinas apresentarem e interpretar os resultados de uma pesquisa em forma eficiente. **Conclusões**. O procedimento descrito permite uma interpretação simples e rápida da análise de muitas variáveis, tornando-o fácil de usar e de pôr em prática.

Palavras-chave: perfis, segmentação, técnicas de análise multivariada, correspondências.

Introducción

La humanidad en su evolución ha necesitado estudiar e interpretar el comportamiento del medio que la rodea. Este conocimiento se hace necesario porque dichos fenómenos afectan su desarrollo en todos los ámbitos del ser humano (social, económico, tecnológico, físico, etc.). Esta comprensión se logra mediante la construcción de modelos que puedan reproducir y explicar estos fenómenos.

La complejidad de los fenómenos de las ciencias en general ya sean puras o aplicadas, lleva a que investigadores en diferentes disciplinas se vean enfrentados a problemas donde intervienen múltiples variables y grandes volúmenes de información. Estos análisis requieren conceptos avanzados y herramientas especializadas para su tratamiento e interpretación general. Por esta razón, se han desarrollado las técnicas estadísticas multivariadas, pero sólo con la evolución de los computadores y diversos paquetes de software que procesan grandes conjuntos de datos, ha llegado a ser notoria la potencia de la estadística multivariada.

Existen diferentes métodos y en muchos casos su aplicación depende del carácter de las variables analizadas. El análisis factorial es aplicado normalmente para el análisis de variables contínuas con dos objetivos específicos: el primero de ellos de descripción o comprobación y el segundo con el fin de reducir el tamaño de un gran número de variables obteniendo nuevos factores ortogonales, que por ser combinación lineal de las variables originales recogen una gran proporción de la información suministrada. Normalmente estos factores son analizados de manera general. En el caso de variables categóricas, la técnica apropiada corresponde al análisis de correspondencias múltiples. El objetivo de este artículo es utilizar la información correspondiente a los factores determinados mediante el análisis de correspondencias, con el fin de perfilar comportamientos de las variables a analizar para luego realizar una segmentación natural utilizando como insumo los factores determinados mediante el análisis factorial. A este respecto es posible consultar aplicaciones en diversas áreas del conocimiento, algunas de ellas se pueden encontrar en Jombar et al. (1), Carranza *et al*. (2), Villarroel *et al*. (3). De igual forma algunos autores han realizado implementaciones importantes basadas en el análisis de correspondencias, entre ellos podemos enunciar a Sourial et al. (4), Wen (5) y Akiyama *et al*. (6).

Materiales y métodos

El análisis de correspondencias múltiples permite mediante la información consignada en una encuesta, instrumento o simplemente en grandes volúmenes de información, establecer relaciones entre las diferentes categorías presentes en las variables analizadas, tomando las categorías originales y estableciendo nuevos factores que resumen la información consignada en dichas respuestas. Este análisis hace más fácil la interpretación de los resultados debido a que permite una representación gráfica más potente que otros métodos y un punto de vista multivariado que garantiza la inclusión de las variables de interés. A continuación se presentan las técnicas de análisis de correspondencias simples para luego introducir su generalización a través del análisis de correspondencia múltiple, esta generalización se logra utilizando como base el análisis de componentes principales.

Análisis de correspondencia simples (ACS)

El Análisis de Correspondencias Simples (ACS), es una técnica estadística multivariada cuyo objeto es encontrar o establecer relaciones entre las filas y las columnas de una tabla de contingencia. En una tabla de contingencia las filas y las columnas están formadas por las modalidades de dos variables categóricas (*A* y *B*), de manera que los elementos de la tabla constituyen las frecuencias de ocurrencia simultánea de las categorías de la variable *A* y las categorías de la variable *B*. La metodología utilizada para establecer las relaciones entre las categorías de las dos variables en cuestión, es la misma que la utilizada en el Análisis de Componentes Principales (ACP). Es decir se trata de encontrar la mejor representación simultánea de la tabla de contingencia, como se presenta a continuación.

Sea *K* una matriz de orden (nxp) cuyo elemento ij-ésimo es K_{ij}, el cual representa el número de individuos pertenecientes a la categoría *i* de la variable *A* y simultáneamente a la categoría *j* de la variable *B*. K_{ij} se denomina frecuencia absoluta. Esquemáticamente los datos se presentan de la siguiente manera:

		Variable B					
		Categoría 1	....	Categoría *j*	....	Categoría *p*	
Variable A	Categoría 1	k_{11}		k_{1j}		k_{1p}	Frec. marginal fila
	Categoría ***i***			k_{ij}			$k_{i.}$
	Categoría ***n***	k_{n1}		k_{nj}		k_{np}	
	Frecuencia marginal col.			$k_{.j}$			k

El anterior esquema contiene, además de las frecuencias absolutas, los totales por fila y columna que se calculan de la siguiente forma:

$$k_{i.} = \sum_{j=1}^{p} k_{ij}$$ Frecuencia marginal de la fila [1]

$$k_{.j} = \sum_{j=1}^{p} k_{ij}$$ Frecuencia marginal de la columna [2]

El total de observaciones es representado por:

$$k = \sum_{i=1}^{n}\sum_{j=1}^{p} k_{ij} \quad [3]$$

Puesto que el objeto es analizar la estructura de las distancias entre categorías, no es suficiente comparar la distancia entre las frecuencias absolutas porque estas solo reflejan la diferencia entre el número de individuos por categoría y para categorías con frecuencias muy diferentes, esta diferencia pierde sentido. En cambio si se comparan los porcentajes, estas distancias representan niveles de símilaridad entre categorías, estos porcentajes son denominados perfiles y se determinan usando las siguientes relaciones:

$$f_{i.} = \frac{k_{ij}}{k_{i.}}$$ (perfil fila) y $$f_{.j} = \frac{k_{ij}}{k_{.j}}$$ (perfil columna) [4]

Estos perfiles son las frecuencias condicionales de filas y columnas respectivamente. Nótese también que la transformación hecha es la misma para las filas y para las columnas. Esto significa que filas y columnas serán tratadas de manera simétrica y que están puestas en correspondencia unas con otras.

Las proximidades entre puntos son ahora distancias entre perfiles. Así por ejemplo la distancia entre dos perfiles fila (dos categorías) de la variable *A* es:

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2 \quad [5]$$

Entre dos perfiles columna (categorías) de la variable B:

$$d^2(j,j') = \sum_{j=1}^{p} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2 \quad [6]$$

Esta distancia se llama distancia chi-cuadrado y se diferencia de la distancia euclidiana por el hecho de que cada cuadrado es ponderado por el inverso de la frecuencia correspondiente al término. Cada sumando se pondera con valores pequeños aquellas diferencias de perfiles entre los individuos *i* e *i'*, que son frecuentes en la categoría *j* de la variable *B* y con valores grandes aquellas diferencias entre perfiles de individuos que son poco frecuentes o que tienen una baja ocurrencia. Este tipo de ponderación logra dar mayor importancia a las modalidades más raras por su escasez, permitiendo diferenciar mejor las filas comparadas.

En otras palabras, esta forma de ponderar las diferencias entre individuos garantiza que se de una adecuada importancia a las diferencias de perfiles entre individuos de aquellas categorías de la variable *B* que son poco frecuentes, pero que no por esto dejan de ser importantes para establecer las similaridades o concordancia entre estos.

La distancia chi-cuadrado tiene la propiedad de que los resultados obtenidos con su aplicación son independientes de la manera como se hayan codificado las modalidades en las variables. En otras palabras: Si se funden dos categorías de cualquiera de las variables *A* ó *B*, los resultados obtenidos en análisis de correspondencias, antes y después de la fusión, son los mismos. Esta propiedad de la distancia chi-cuadrado se denomina *equivalencia distribucional* y le da robustez al análisis.

Las anteriores definiciones se pueden expresar de manera mas compacta en términos matriciales como sigue:

Sea K la matriz de las frecuencias absolutas de las variables *A* y *B*. Entonces la matriz *F* de las frecuencias relativas se calcula mediante la ecuación:

$$F = \frac{1}{k} K \quad [7]$$

Donde: $k = \sum_{i=1}^{n}\sum_{j=1}^{p} k_{ij}$

Se define la matriz de frecuencias relativas marginales de las filas:

$$D_n = diag\{f_{i.}, i = 1,...,n\} \quad [8]$$

y la matriz de frecuencias relativas marginales para las columnas:

$$D_p = diag\{f_{.j}, j = 1,...,p\} \quad [9]$$

Así, los perfiles de las filas se obtienen del producto $D_n^{-1}F$ y los perfiles de las columnas por $D_p^{-1}F$.

Cálculo de los ejes principales y las coordenadas.

Nótese que las distancias definidas entre modalidades de las filas o entre las columnas pueden escribirse también de la forma:

$$d^2(i,i') = \sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} - \frac{f_{i'j}}{f_{i'.}\sqrt{f_{.j}}}\right)^2 \qquad [10]$$

$$d^2(j,j') = \sum_{j=1}^{p}\left(\frac{f_{ij}}{f_{.j}\sqrt{f_{i.}}} - \frac{f_{ij'}}{f_{.j'}\sqrt{f_{i.}}}\right)^2 \qquad [11]$$

con lo que, las distancias son transformadas en distancias euclidianas corrientes.

Puesto que existe una completa simetría entre el manejo de filas y columnas solo se presentan los cálculos correspondientes al espacio $\Re^p$. Los cálculos correspondientes al espacio $\Re^n$ se obtienen directamente por permutación de los índices *i* y *j*. La demostración se encuentra en Lebart et. Al (7) y permite realizar análisis equivalentes en ambas direcciones.

Análisis en el espacio $\Re^p$

Análisis centrado

Para llevar a cabo un análisis centrado es necesario calcular el centro de la nube de puntos es decir la media ponderada de las coordenadas , $\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}}$ ponderadas por la masa de la fila $f_{i.}$. El centro de la nube de puntos en este caso es $g = (g_1,...,g_p)$ y su j **-** ésima componente es:

$$g_j = \sum_{i=1}^{n} f_{i.}\frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} = \sqrt{f_{.j}} \qquad [12]$$

Entonces el análisis centrado se hace sobre la matriz $\boldsymbol{T_{nxp}}$ cuyos elementos tienen la forma:

$$t_{ij} = \sum_{k=1}^{n} f_{k.}\left(\frac{f_{ki}}{f_{k.}\sqrt{f_{.i}}} - \sqrt{f_{.i}}\right)\left(\frac{f_{kj}}{f_{k.}\sqrt{f_{.j}}} - \sqrt{f_{.j}}\right) =$$

$$\sum\left(\frac{f_{ki} - f_{k.}f_{.i}}{\sqrt{f_{k.}f_{.i}}}\right)\left(\frac{f_{kj} - f_{k.}f_{.j}}{\sqrt{f_{k.}f_{.j}}}\right) \qquad [13]$$

También se puede demostrar que la matriz tiene la forma $T=Z`Z$. **Donde:** $Z = \{z_{ij}\}$, $i = 1,...,n$, $j = 1,...,p$ con elementos:

$$z_{ij} = \frac{f_{ij} - f_{i.}f_{.j}}{\sqrt{f_{i.}f_{.j}}} \qquad [14]$$

El análisis se hace análogo al realizado en componentes principales. De esta forma, los ejes factoriales se encuentran como en el *ACP*, de la ecuación

$$Tu_\alpha = \lambda_\alpha u_\alpha \qquad [15]$$

Donde: se tiene la restricción $u'_\alpha u_\alpha = 1$.

Análisis no centrado

Cuando el análisis se hace no centrado, o sea sobre las coordenadas

$$r_{ij} = \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} \qquad [16]$$

Entonces se trabaja con la matriz $\dot{Z} = \{\dot{z}_{ij}\}$ cuyos elementos son:

$$z_{ij} = \sqrt{f_{i.}}\, r_{ij} = \frac{f_{ij}}{\sqrt{f_{i.}f_{.j}}} \qquad [17]$$

La diagonalización de la matriz $\dot{T} = \dot{Z}'\dot{Z}$ conduce a los mismos resultados que el análisis centrado, pero en este caso hay que eliminar el valor propio que es igual a 1 y su correspondiente vector propio que es el centro de gravedad de la nube $g = (g_1,...,g_p)$ cuyas componentes son $g_j = \sqrt{f_{.j}}$.

Coordenadas

Las coordenadas de la proyección de los datos sobre el eje u_α son entonces $\hat{\psi}_\alpha = Xu_\alpha$ y $\hat{\phi}_\alpha = X'v_\alpha$; y explícitamente para puntos *i* y *j* tienen la siguiente forma:

$$\hat{\psi}_{i\alpha} = \sum_{j=1}^{p} \frac{f_{ij} - f_{i.} f_{.j}}{f_{i.}\sqrt{f_{.j}}} u_{j\alpha} \quad [18]$$

$$\hat{\psi}_{j\alpha} = \sum_{i=1}^{p} \frac{f_{ij} - f_{i.} f_{.j}}{\sqrt{f_{i.}} f_{.j}} v_{i\alpha} \quad [19]$$

Y si el análisis es no centrado entonces:

$$\hat{\psi}_{i\alpha} = \sum_{j=1}^{p} \frac{f_{ij}}{f_{i.}\sqrt{f_{.j}}} u_{j\alpha} \quad [20]$$

o bien

$$\hat{\psi}_{j\alpha} = \sum_{i=1}^{p} \frac{f_{ij}}{\sqrt{f_{i.}} f_{.j}} v_{i\alpha} \quad [21]$$

Las ecuaciones de transición se obtienen directamente de las matrices T y $\dot{\boldsymbol{T}}$, y tiene los mismos valores propios diferentes de cero y son análogas a las obtenidas al ACP y tienen la forma:

$$v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} T u_\alpha \quad [22]$$

$$\boldsymbol{u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \dot{T}' v_\alpha} \quad [23]$$

Generalización de ACS

El análisis de correspondencias múltiple es una generalización natural del análisis de correspondencias simples, y por lo tanto se utiliza para analizar varias variables categóricas simultáneamente. Una condición necesaria para que se pueda hacer el análisis es que a cada pregunta solo se puede contestar una categoría única. En otras palabras las respuestas a las preguntas se dan de manera disyuntiva completa. De esta manera las *k* categorías de respuesta a una pregunta dada permiten particionar la muestra en *k* grupos. De esta forma y como se mencionó anteriormente las proyecciones son análogas a las determinadas por ACP.

Tabla de Burt

Esta tabla es muy útil para realizar análisis específicos entre variables, ya que presenta las relaciones dos a dos de todo el conjunto de variables, de igual forma es posible establecer cualquier test de asociación o correlación utilizando los valores allí consignados. Asociada a la matriz *Z* se define una matriz *B* como el producto:

$$\boldsymbol{B = Z'Z} = \begin{bmatrix} \boldsymbol{Z_1'Z_1} & \boldsymbol{Z_1'Z_2} & \cdots & \boldsymbol{Z_1'Z_Q} \\ \boldsymbol{Z_2'Z_1} & \boldsymbol{Z_2'Z_2} & & \boldsymbol{Z_2'Z_Q} \\ \vdots & & \ddots & \\ \boldsymbol{Z_Q'Z_1} & \boldsymbol{Z_Q'Z_2} & \cdots & \boldsymbol{Z_Q'Z_Q} \end{bmatrix} \quad [24]$$

Esta Matriz se denomina tabla de Burt y es una tabla de contingencia asociada a la matriz *Z* en el sentido de que es una matriz por bloques con las siguientes características:

Los bloques de la diagonal $Z_q'Z_q$ contienen en la diagonal la distribución de frecuencias de la variable *q* y los bloques de fuera de la diagonal $Z_q'Z_{q'}$ son tablas de contingencia entre las categorías de las variable *q* y *q'*.

De esta forma, mediante el análisis de correspondencias múltiples es posible extraer factores que permiten la perfilación de individuos. Con estos factores es posible aplicar técnicas de segmentación convencional y lograr resultados óptimos que mejoran ostensiblemente la segmentación natural y permiten una sencilla y eficiente interpretación. Geenacre (8), Cressie (9) y Fichet (10) presentan de manera sencilla el análisis de correspondencias simples y las técnicas de clasificación por separado, usuarios avanzados pueden consultar estos autores para mayor información sobre los temas aquí expuestos.

Aplicación del modelo

Son diversas las aplicaciones que se pueden llevar a cabo utilizando esta metodología, se ha escogido el tema del perdón porque permite ver fácilmente la implementación del método antes expuesto. El tema del perdón ha sido muy estudiado en el contexto Latino-Americano. Las estructuras (eg, forgivingness y conceptualizaciones) son similares en muchos países. Esto resulta lógico debido a que muchos

estados y sus anteriores colonias tienen las mismas raíces culturales. Por otro lado la relación entre el perdón y otras variables han sido muy estudiada en Europa y estados unidos. Diferentes grupos de investigación en Europa han creado formatos para evaluar de capacidad de reconciliación y observar el comportamiento del perdón en muchos ámbitos, para resaltar las investigaciones realizadas por Akl y Mullet (11), Ballester et al. (12), Chiaramello et al. (13) y Mullet y Azar (14). Estos autores han desarrollado y perfeccionado instrumentos que permiten medir escalas de perdón y sus relaciones con variables de perfilación, sin embargo la interpretación de los resultados no es fácil. Es necesario aclarar que el objetivo del estudio es la aplicación de una metodología para la implementación y análisis, por lo tanto la interpretación concisa de los resultados se deja a expertos en el tema.

Para implementar la metodología expuesta en este artículo, es necesario distinguir tres fases de análisis, estas son:

- Determinación de factores mediante análisis de correspondencias múltiples.
- Perfilación de las variables de interés mediante la aplicación de los factores anteriormente determinados.
- Segmentación (Clasificación) de los individuos utilizando los factores obtenidos en el análisis de correspondencias Múltiple

Utilizando la metodología antes presentada se analizó la información consignada en 113 encuestas realizadas en Colombia donde se pretende analizar este tema y su relación con variables socio-demográficas como son: estrato socioeconomico, nivel educativo, edad y sexo entre otras. Como primera medida es necesario determinar los factores que permiten la interpretación y perfilación de los individuos. En la tabla 1 se presenta la composición y participación de los primeros cinco factores determinados mediante el análisis de correspondencias múltiple.

Siguiendo el método de correspondencias múltiples antes presentado, se determinan nuevos factores como combinación lineal de las variables originales (**Figura 1**), en esta aplicación se analizan solo los factores 1 y 2 y se observa como al utilizar las coordenadas estimadas para estos factores (cerca del 73% de la información original) se puede lograr una perfilación de los individuos en el tema del perdón, donde la escala esta definida entre 1 y 10 (No Perdón - Perdón), y como el comportamiento de los individuos analizados hace parte de esta trayectoria que va desde perdón hasta no perdón. Sobre esta trayectoria es posible visualizar el comportamiento de variables que permitan una mayor caracterización de los individuos como variables sociodemográficas tal como estrato, nivel educacional, sexo y nivel de estudio pueden ser interpretadas de manera sencilla dentro de la configuración establecida para la variable en estudio. Como se explico anteriormente, es posible analizar un número mayor de factores, y observar el comportamiento de las variables en cada caso.

Aunque el objetivo del presente artículo es desarrollar un método de análisis y no entrar a fondo en este tema específico, se presentan algunas relaciones entre las variables analizadas con el fin de exponer la potencia del método y su sencilla implementación. Por ejemplo, se observa como estratos 5 y 6 se encuentran asociados a no perdón, al igual que nivel de estudios de posgrado. De igual forma y como es de esperarse (de acuerdo con lo anteriormente expuesto) niveles de escolaridad como primaria y secundaria se encuentran asociadas al perdón, al igual que los estratos 2 y 3. En este tipo de análisis es posible observar variables que permanecen neutrales sobre los ejes analizados, es el caso de la variable sexo, que se encuentra en una zona de neutralidad (centro de gravedad) indicando que no presenta ninguna asociación con el eje central de estudio, en este caso el perdón. Esto permite no solo establecer segmentos poblacionales descritos a detalle, si no que permite comparar los diferentes cluster con la perfilación antes elaborada.

Una vez determinados los factores mediante análisis de correspondencias es posible utilizar técnicas de clasificación convencionales para segmentar los individuos, Everitt et al. (15) presentan los últimos avances en estos temas y hacen algunas aplicaciones prácticas si se quiere profundizar en este aspecto.

Tabla 1. Composición y participación de los primeros cinco factores en el análisis de correspondencias múltiple.

Factor	Valor propio	Contribución
1	0,5096	0,4521
2	0,3926	0,2715
3	0,2726	0,1036
4	0,2369	0,0831
5	0,2021	0,0375

Fuente: Cálculos propios

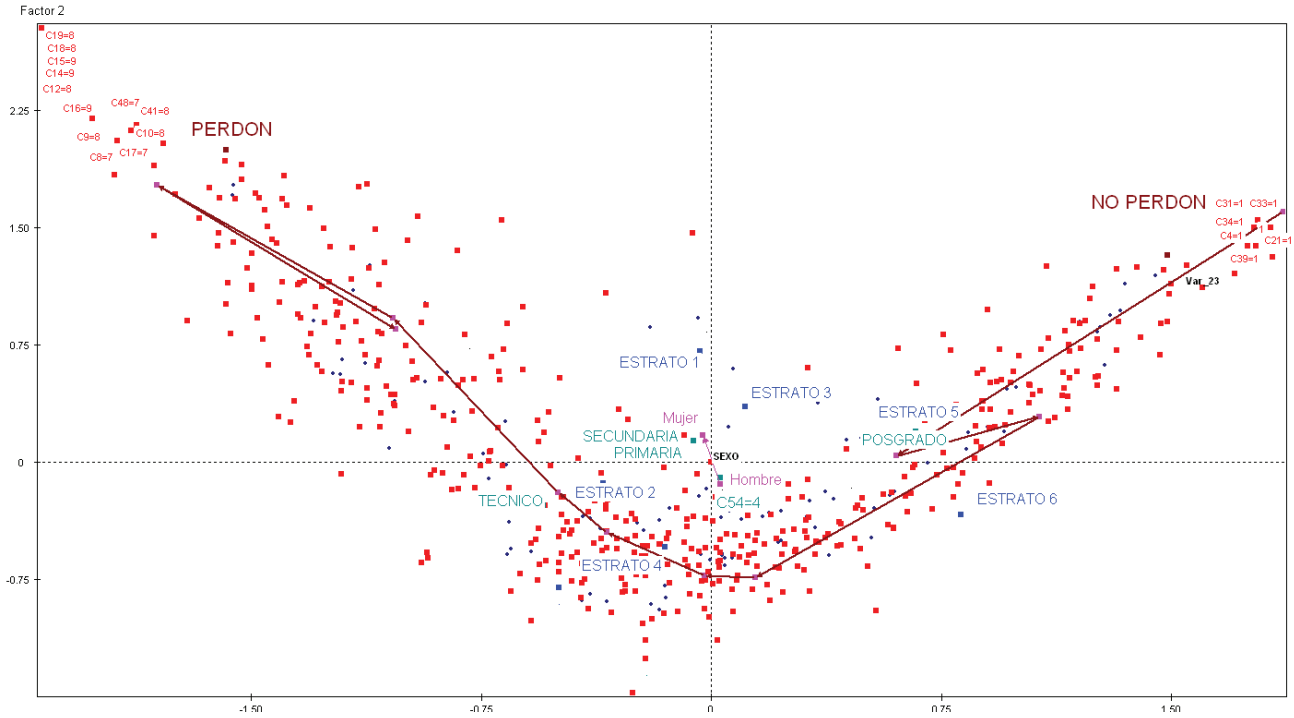

Figura 1. Perfilación de variables usando los primeros dos factores obtenidos del análisis de correspondencias múltiple, aproximadamente 72% de la variación total.

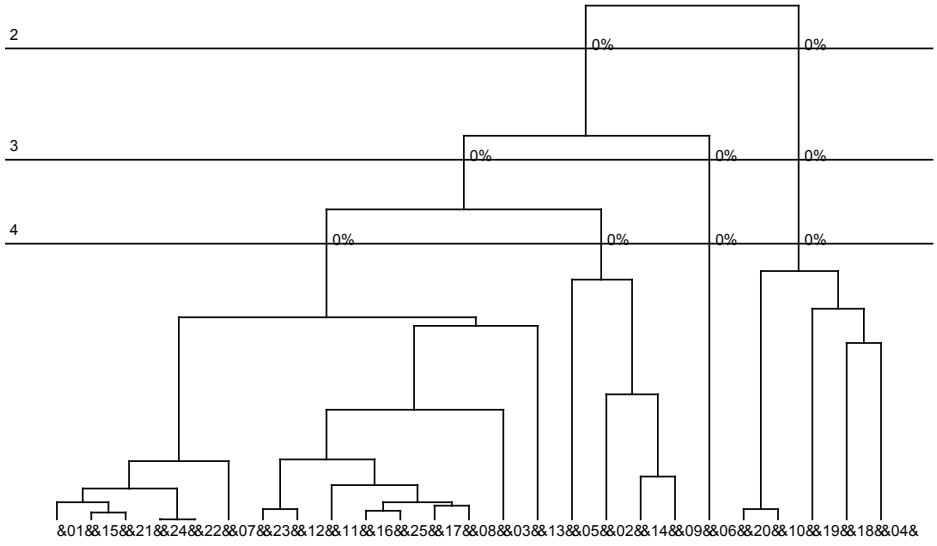

Figura 2. Dendrograma correspondiente a la clasificación jerárquica de los individuos mediante el cual se determina el número óptimo de segmentos.

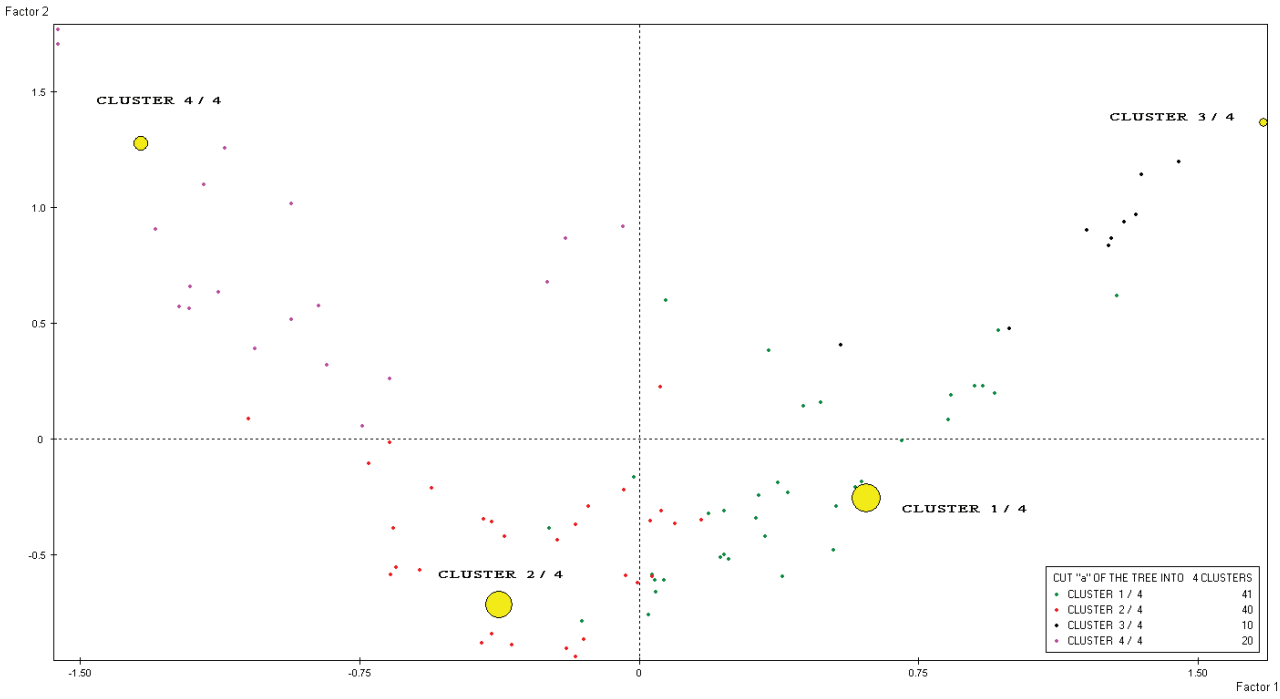

Figura 3. Distribución de los segmentos estimados (Clusters), visto en forma conjunta con el análisis de correspondencias múltiple.

El primer paso en la segmentación es determinar el número de grupos que naturalmente se deben establecer para este análisis, para este caso se utiliza un dendrograma donde es posible observar que de manera natural se deben determinar cuatro segmentos o clusters que son los que ofrecen una mayor variabilidad entre grupos y menor variabilidad al interior de cada cluster (**Figura 2**).

Para finalizar se hace una representación gráfica utilizando las coordenadas determinadas en el anterior análisis logrando extrapolar la distribución de los clusters sobre la perfilación antes presentada. Como se observa en la figura 3 esta metodología resulta óptima pues permite ver la composición de los clusters sobre la escala de análisis de la variable de interés, en este caso el perdón. De esta forma el investigador tendrá a su disposición representaciones gráficas que le permitan profundizar en el análisis y presentar resultados eficientemente. De igual forma es posible determinar el aporte de las variables analizadas en la formación de cada uno de los clusters.

Discusión

Utilizando análisis de correspondencias múltiples se pueden obtener nuevos factores ortogonales que son combinaciones lineales de las variables originales, que a su vez reúnen gran parte de la información original. El análisis de dichos factores permite examinar de manera óptima las relaciones existentes entre las variables e individuos, puesto que gran parte del análisis se lleva a cabo utilizando la exploración gráfica que facilita la labor del investigador. Esta forma de utilización del análisis de correspondencias múltiples se logra en la medida que se establecen nuevas dimensiones que corresponden a patrones de comportamiento, que a su vez son de fácil análisis e interpretación.

La segmentación de individuos permite generar clusters de manera natural acorde al comportamiento propio de las variables de interés. Los métodos tradicionales de segmentación son basados en su mayoría en técnicas de minería de datos que resultan extensas y demandan un alto costo en tiempo. Utilizando los nuevos factores generados por el análisis de correspondencias múltiples se logra mejorar la eficiencia de la segmentación y son de fácil interpretación puesto que su implementación puede ser realizada en software de distribución libre como es el caso del R (Dalgaard) (16).

Conclusiones

El interés de este artículo es presentar una nueva metodología de interpretación y análisis de grandes volúmenes de

información utilizando técnicas multivariadas conocidas. Este enfoque se obtiene logrando combinar métodos como análisis de correspondencias múltiples y análisis Cluster, utilizando sus propiedades para mejorar y optimizar su desarrollo. Estudios similares han sido implementados utilizando Análisis de componentes principales, pero como metodológicamente es bien sabido esta técnica es usada cuando las variables de interés son continuas, y en el caso de análisis de correspondencias múltiples se trabaja con variables de tipo categórico.

El procedimiento expuesto permite una sencilla y rápida interpretación del análisis en muchas variables, haciendo fácil su uso e implementación a investigadores de diferentes disciplinas interesados en este tipo de técnicas.

Financiación

Este trabajo fue realizado con recursos propios del Grupo de Física Matemática, Departamento de Matemáticas. Facultad de Ciencias, Pontificia Universidad Javeriana, Bogotá, D.C., **Colombia**.

Conflicto de intereses

No existe conflicto de intereses.

Referencias

1. Jombart T, Pontier D, Dufour A-B. Genetic markers in the playground of multivariate analysis. *Heredity*. 2009; **102**: 330-341

2. Carranza X, Fonseca G, Tellez AA. Aplicación de métodos multivariados: una respuesta a las limitaciones de los ratios financieros. Revista Contribuciones a la Economía, http://www.eumed.net/ce/2011a/cft.htm. Consultado Julio 12 de 2011.

3. Villarroel L, Alvarez J. Maldonado D. Aplicación de Análisis de Componentes Principales en el Desarrollo de Productos. *Revista Acta Nova* 2003; **2** (3): 399-408.

4. Sourial N, Wolfson C, Zhu B, Quail J, Fletcher J, Karunananthan S, Bandeen-Roche K, Béland F, Bergman H. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. *Journal of Clinical Epidemiology* 2010; **63 (**6): 638-646.

5. Wen C, Chen W. Using multiple correspondence cluster analysis to map the competitive position of airlines. *Journal of Air Transport Management*. 2010; **17**, (5), 302-304.

6. Akiyama T, Kobayashi K, Ohtsuka Y. Electroclinical characterization and classification of symptomatic epilepsies with very early onset by multiple correspondence analysis. *Epilepsy Research* 2010; **91**, (2-3): 232-239.

7. Lebart L, Morineau A, Warwick KM. Multivariate Descriptive Statistical Analysis Correspondence Analysis and Related Techniques for Large Matrices. Jhon willey & Sons. New York 1984, 304 p.

8. Greenacre M. La práctica del análisis de correspondencias. Fundacion BBVA. Espana 2008, 375 p.

9. Cressie N, Wikle C. Statistics for Spatio-Temporal Data. Jhon willey & Sons. New York 2011, 624 p.

10. Fichet B, Piccolo D, Verde R, Vichi M. Classification and Multivariate Analysis for Complex Data Structures. Springer. Berlin Heidelberg 2011, 473 p.

11. Akl M, Mullet E. Forgivingness: Relationship with conceptualizations of God's forgiveness and childhood memories. *The International Journal for the Psychology of Religion* 2010; (20): 187-200.

12. Ballester S, Muñoz MT, Mullet E. Forgivingness and lay conceptualizations of forgiveness. *Personality and Individual Differences* 2009; **47** (6): 605-609.

13. Chiaramello S, Mesnil M, Muñoz MT, Mullet E. Dispositional forgiveness among adolescents. *European Journal of Developmental Psychology* 2008; **5** (3): 326-337.

14. Mullet E, Azar F. Apologies, repentance and forgiveness: A Muslim-Christian comparison. The *International Journal for the Psychology of Religion* 2009; **19** : 275-285.

15. Everitt B, Landau S, Leese M, Stahl D. Cluster Analysis. Jhon willey & Sons. New York 2011, 330 p.

16. Dalgaard P. Introductory statistics with R. Primera edición. Springer. New York, USA. 2008, 363 p.