Revista Colombiana de Estadística
Diciembre 2009, volumen 32, no. 2, pp. 231 a 245

Una aplicación del método jerárquico de mezclas para la clasificación de los municipios venezolanos según variables socioeconómicas

An Application of Hierarchical Method of Mixtures for the Classification of the Venezuelan Counties using Socioeconomic Variables

Freddy Omar López Quintero[1,a], Rafael Eduardo Borges Peña[2,b]

[1]Departamento de Matemáticas, Instituto Venezolano de Investigaciones Científicas, Miranda, Venezuela

[2]Escuela de Estadística, Facultad de Ciencias Económicas y Sociales, Universidad de Los Andes, Mérida, Venezuela

Resumen

En este trabajo se presenta una aplicación del método propuesto por Fraley & Raftery (2002) para la obtención de grupos de municipios de Venezuela a partir de un conjunto de variables socioeconómicas. Las variables consideradas miden aspectos del hogar de las familias que viven en los municipios, la ocupación de sus miembros, la educación, aspectos demográficos, entre otros. Como datos de entrada, se decidió tomar los primeros seis componentes principales de un análisis previo realizado a estos datos. Se obtuvieron nueve grupos diferenciados entre sí marcando, principalmente, diferencias en el estatus social, en el acceso a algunos servicios, y la calidad de vida en general.

***Palabras clave*:** factor de Bayes, análisis de conglomerados, algoritmo EM, modelos mezclados.

Abstract

In this work, we present an application of the method proposed by Fraley & Raftery (2002) to obtain groups of Venezuelan counties, using the information of socio-economic variables. The variables considered in the application includes some aspects related with the families that live in counties, such as occupation of its members, education, demographic aspects and others. For the analysis, we use the first six principal components taken from a previous analysis. A classification on nine groups was obtained, and the difference between these groups was influenced by the socioeconomic status, the access to some basic services and quality of life.

***Key words*:** Bayes factor, Cluster analysis, EM algorithm, Mixture models.

[a]Estudiante de maestría. E-mail: folopez@ivic.ve
[b]Profesor agregado. E-mail: borgesr@ula.ve

1. Introducción

El análisis de conglomerados (*cluster analysis*) es un conjunto de técnicas que permite la ubicación de unos objetos, ítems, individuos, etc., dentro de unos grupos denominados conglomerados, de forma tal que, en cada grupo, los objetos sean semejantes entre sí y, entre los diversos grupos, diferentes.

Principalmente se busca: la identificación de tales grupos, la confirmación de sus diferencias y la explicación de su formación, en cuanto a las variables medidas en ellos.

La manera en que se forman los conglomerados puede variar. Entre los métodos más populares están los jerárquicos, de partición, gráficos, y de conglomerados difusos (Díaz 2002).

Fraley (1998), Fraley & Raftery (2002) añaden un supuesto importante para la búsqueda de los conglomerados: la normalidad multivariante. Este supuesto, junto a la utilización de información previa de los individuos (forma presumible en la que pueden estar formados los conglomerados), ayuda a la definición de los grupos resultantes.

El modelo propuesto por Fraley y Raftery no es el único, ni el más reciente; existen en la literatura diversas propuestas, algunas de las cuales mencionaremos a continuación. Gallegos & Ritter (2005) presentan un método robusto que permite trabajar incorporando los valores atípicos (*outliers*), aunque con limitaciones respecto a las familias paramétricas considerada. Oh & Raftery (2007) plantean un método de clusters basados en modelos que admiten disimilaridades en el espacio euclídeo de distancias. Gnanadesikan et al. (2007) presentan un interesante trabajo donde se plantean algunas alternativas para la identificación de la estructura de los clusters. Bouveyron et al. (2007) proponen una generalización del método basado en mezclas o mixturas para datos de alta dimensión. Y en otro contexto, Leisch (2006) plantea una interesante discusión del análisis de conglomerados basado en los centroides.

Sin embargo, el método propuesto por Fraley & Raftery (2002) sigue siendo una excelente alternativa, debido a la disponibilidad de software y a la versatilidad en cuanto a las distribuciones admitidas, las cuales, según los mismos autores, no necesariamente tienen que ser gaussianas. Una revision actualizada de algunos programas disponibles se encuentra en el trabajo de Haughton et al. (2007).

Este trabajo se enmarcó básicamente bajo la metodología de estos últimos y no representa un avance metodológico del tema pero sí muestra nuevos hallazgos que ayudan a conocer el país. El objetivo fundamental es determinar territorios sociales (municipales) en Venezuela partiendo de una base de datos censal cuyas variables son principalmente socioecónomicas (véanse INE 2005*a*, INE 2005*b*, INE 2005*c*).

Las variables pueden dividirse en: equipamiento del hogar (porcentaje de familias con nevera, proporción de familias con internet, etc.), acceso a servicios del hogar (porcentaje de familias sin servicio de electricidad, porcentaje de familias que no tiene servicio de aseo urbano, etc.), ocupación (porcentaje de familias que trabaja principalmente en el sector público, porcentaje de familias que trabaja principalmente en el sector informal, etc.), educación (alfabetismo, personas titu-

ladas, etc.), aspectos demográficos (índice de masculinidad del municipio, tasa de natalidad, etc.), otros indicadores y ciertos activos.

La búsqueda de estos grupos no es nueva en la literatura. Por ejemplo, Bergonzoli (2006) sugiere una forma de estratificar zonas geográficas (municipios, parroquias, cantones, estados, etc.) con el método de la razón proporcional de brechas (RPB) que toma en cuenta la tasa de mortalidad, porcentaje de analfabetismo, vacuna antisarampión en niños menores de un año, y el porcentaje de riqueza; y lo ejemplifica con estados guatemaltecos.

De forma alternativa, da una serie de pasos para la estratificación a través de otras variables: el producto interno bruto (PIB), el porcentaje de personas pertenecientes a una etnia indígena y el porcentaje de ruralidad. Una vez conocidos estos estratos, y conocidos qué estados pertenecen a cuáles grupos, realiza varios análisis de varianzas para cerciorarse que son realmente distintos.

Lago et al. (2000), en su trabajo sobre la conformación de subregiones argentinas, proponen la utilización de métodos estadísticos multivariados con una cantidad mayor de variables. A este efecto, dicen, utilizaron veintiséis variables, y, básicamente, siguieron dos pasos: a) realizaron un análisis de componentes principales sobre la matriz de datos para resumir esta información y b) clasificaron los setentaiún departamentos en una cantidad reducida de estratos. Para este último punto, utilizaron el método de k-means con los puntajes obtenidos en a). Para verificar el resultado obtenido, realizaron una serie de análisis de varianzas.

López et al. (2002), en nuestro país, señalan los pasos para crear los estratos nacionales mediante, únicamente, el análisis de conglomerados. Es bueno advertir que López et al. (2002) no dan resultados sobre este asunto, sino que indican cómo realizarlo. El tipo de análisis de conglomerado que utilizan es de k-means.

Este trabajo se divide de la siguiente manera: la sección 2 expone el análisis de conglomerados desde la perspectiva de Fraley y Raftery, considerando sus etapas principales; la sección 3 muestra una aplicación de la técnica sobre unos datos de tipo socioeconómico, y en la sección 4 se presentan algunas conclusiones.

2. Análisis de conglomerados según Fraley y Raftery (2002)

Sea X una variable p-dimensional observada en el conjunto de datos y sea $f(x)$ su función de densidad. Sean $\{x_i; i = 1, \ldots, n\}$ las observaciones de X correspondientes a una muestra aleatoria simple de la población objeto de estudio.

Fraley y Raftery suponen que la densidad f viene dada por la mezcla de la forma

$$f(\mathbf{x}) = \sum_{k=1}^{m} \pi_k f_k(\mathbf{x} \mid \theta_k) \tag{1}$$

donde cada $\pi_k > 0$ y $\sum_{k=1}^{m} \pi_k = 1$.

Por su parte, m es el número de componentes en la mezcla (número de grupos presentes en la población estudiada), $\{f_k(x \mid \theta_k); k = 1, \ldots, m\}$ son los modelos

distribucionales para cada uno y $\{\pi_i; k = 1, \ldots, m\}$ son los pesos dentro de la mezcla (tamaño del grupo k-ésimo).

Además, para Fraley y Raftery, el k-ésimo conglomerado se puede representar por un modelo gaussiano de la forma

$$\phi_k(x_i \mid \mu_k, \Sigma_k) = \frac{e^{-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)}}{\sqrt{|2\pi\Sigma_k|}} \tag{2}$$

cuyas medias y varianzas son, respectivamente: μ_k y Σ_k.

Cada matriz de covarianzas puede parametrizarse por su descomposición espectral en la forma

$$\Sigma_k = \lambda_k D_k A_k D_k^T \tag{3}$$

donde D_k es la matriz ortogonal de vectores propios de Σ_k y sirve para determinar la orientación de los elipsoides de equidensidad de Σ_k; A_k es una matriz diagonal en la que se verifica que $|A_k| = 1$ y determina la forma de la distribución; además, sus elementos son proporcionales a los valores propios de Σ_k, y λ_k es un escalar que especifica el volumen del correspondiente elipsoide, el cual es proporcional al escalar $\lambda_k^d |A_k|$, donde d es la dimensión de los datos.

Las características (orientación, volumen, y forma) de las distribuciones son generalmente estimadas de los mismos datos, y puede permitirse variación entre los conglomerados, o ser forzados a tener las mismas medidas (véanse Murtagh & Raftery 1984, Bandfield & Raftery 1993, Celeux & Govaert 1995).

Utilizando esta (re)parametrización de cada uno de los modelos componentes, en términos de μ_k, λ_k, D_k y A_k, se pueden construir hasta 27 familias de modelos de mezclas que surgen de la combinación de las variantes del

- Volumen (λ_k): I ($\lambda_k = 1, \forall k$); E ($\lambda_k = \lambda, \forall k$) o V ($\lambda_k$ diferente para cada k),
- Forma (A_k): I ($A_k = I_p, \forall k$); E ($A_k = A, \forall k$) o V (A_k diferente para cada k), y
- Orientación (D_k): I ($D_k = I_p, \forall k$); E ($D_k = D, \forall k$) o V (D_k diferente para cada k).

Así, el modelo `EVI` denota un modelo en el cual el volumen de todos los conglomerados es igual (`E`, de *equal*, en inglés), la forma de los conglomerados puede variar (`V`, de *varying*) y la orientación es la identidad (`I`, de *identity*).

Una vez hallada la mejor representación (de las 27) para nuestro conjunto de datos, la metodología de Fraley & Raftery (2002) consiste en 3 etapas:

2.1. Agrupamiento jerárquico

Entonces, en el enfoque de la verosimilitud de clasificación (Fraley 1998) los parámetros de θ y γ son escogidos tales que ellos maximicen

$$L(x; \theta, \gamma) = \prod_{i=1}^{m} f_{\gamma_i}(x; \theta) \tag{4}$$

En su trabajo, Fraley (1998) y Fraley & Raftery (2002) se centran en el caso donde $f_k(x;\theta_k)$ es del tipo normal (gaussiana). Nótese que una vez maximizada la ecuación (4) se obtendrá una variable que nos dirá a qué grupo pertenece el individuo x_i.

Cuando $f_k(x;\mid \theta_k)$ es una función normal multivariante, la función (4) toma la forma

$$L(x;\mu_1,\ldots,\mu_m,\Sigma_1,\ldots,\mu_m,\gamma) = \prod_{k=1}^{m}\prod_{i\in I_k}(2\pi)^{-\frac{\pi}{2}}\,|\Sigma_k|^{-\frac{1}{2}}\,e^{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)} \quad (5)$$

donde $I \in \{i:\gamma_i = k\}$ es el conjunto de índices correspondientes a las observaciones provenientes del k-ésimo grupo.

2.2. Algoritmo EM

El algoritmo EM (Dempster et al. 1977) se utiliza en estadística para hallar el máximo de una función de verosimilitud en un modelo probabilístico, donde el modelo depende de unas variables no observadas. El algoritmo alterna entre la realización de un paso de expectación (E), el cual calcula una esperanza de la verosimilitud incluyendo la variable latente como si ella fuese observada, y un paso de maximización (M), el cual calcula el máximo de la función de verosimilitud utilizando los valores de los parámetros hallados en el paso E. Los parámetros encontrados en el paso M se utilizan para comenzar otro paso E, y el proceso se repite hasta la convergencia.

En el algoritmo EM para modelos mezclados se considera "datos completos" a $x_i = (y_i, z_i)$, donde $z_i = (z_{i1},\ldots,z_{im})$ (m es el número de grupos) es la porción de datos no observados con

$$z_{ik} = \begin{cases} 1 & \text{si } x_i \text{ pertenece al grupo } k \\ 0 & \text{en otro caso} \end{cases} \quad (6)$$

Ahora, se tiene que cada z_i es independiente e idénticamente distribuido de acuerdo con una distribución multinomial de m categorías. Es decir

$$z_i|\pi_1,\ldots,\pi_m,\theta_1,\ldots,\theta_m,x_1,\ldots,x_n \sim Multinomial(1,\alpha_{i1},\ldots,\alpha_{im}) \quad (7)$$

donde

$$\alpha_{ik} = \frac{\pi_k f_k(x_i \mid \theta_k)}{\sum_{k=1}^{m}\pi_k f_k(x_i \mid \theta_k)} \quad (8)$$

es la probabilidad a posteriori que el individuo i-ésimo pertenezca al grupo k-ésimo para $k = 1,\ldots,m$, y tomándose como probabilidades a priori de cada grupo los pesos $\{\pi_i; k = 1,\ldots,m\}$. La densidad de una observación y_i dado z_i está dada por $f(x_i|z_i) = \prod_{k=1}^{m} f_k(x_i \mid \theta_k)^{z_{ik}}$ (Fraley & Raftery 2002).

En efecto, como señala Peña (2004), en z_i solo un componente z_{ik} es distinto de cero y ese componente definirá cuál es la función de densidad de las observaciones. Análogamente, la función de probabilidades de la variable z_i será (Peña 2004)

$$p(z_i) = \prod_{k=1}^{m} \pi_k^{z_{ik}} \tag{9}$$

Por otro lado, la función de densidad conjunta es (Peña 2004)

$$f(x_i, z_i) = f(x_i \mid z_i) p(z_i) \tag{10}$$

que, por (9) y (10), se puede escribir

$$f(x_i, z_i) = \prod_{k=1}^{m} (\pi_k f_k(x_i \mid \theta_k))^{z_{ik}} \tag{11}$$

Y así, la función de logverosimilitud conjunta es

$$L_C(\theta \mid x, z) = \sum_{i=1}^{n} \log f(x_i, z_i) = \sum_{i=1}^{n}\sum_{k=1}^{m} z_{ik} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{m} z_{ik} \log f_k(x_i \mid \theta_k) \tag{12}$$

El algoritmo comenzará con una estimación inicial de los parámetros, $\widehat{\theta}^{(0)}$.

En el paso E se calculará el valor esperado de las observaciones ausentes en la verosimilitud completa (12) condicionando los parámetros iniciales y os datos observados. Como la verosimilitud es lineal en z_{ik}, esto equivale a sustituir las variables ausentes por sus esperanzas. Entonces

$$E\left(z_{ik} \mid x, \widehat{\theta}^{(0)}\right) = p\left(z_{ik} = 1 \mid x_i, \widehat{\theta}^{(0)}\right) = \alpha_{ik}^{(0)} \tag{13}$$

Al sustituir estos valores en (12) se obtiene

$$L_C^*(\theta \mid x) = \sum_{i=1}^{n}\sum_{k=1}^{m} \alpha_{ik}^{(0)} \log \pi_k + \sum_{i=1}^{n}\sum_{k=1}^{m} \alpha_{ik}^{(0)} \log f_k(x_i \mid \theta_k) \tag{14}$$

En la etapa M se debe maximizar la función (14) respecto a los parámetros $\theta = (\pi_1, \ldots, \pi_m, \theta_1, \ldots, \theta_m)$.

Y la solución a este problema (en el caso que el modelo sea VVV) viene dada por

$$\widehat{\mu}_k = \frac{\sum_{i=1}^{n} \alpha_{ik} \mathbf{x}_i}{\sum_{i=1}^{n} \alpha_{ik}}$$

$$\widehat{\Sigma}_k = \frac{\sum_{i=1}^{n} \alpha_{ik} \left(\mathbf{x}_i - \widehat{\mu}_k\right)\left(x_i - \widehat{\mu}_k\right)}{\sum_{i=1}^{n} \alpha_{ik}}$$

y

$$\widehat{\pi}_k = \frac{\sum_{i=1}^{n} \alpha_{ik}}{n}$$

para $k = 1, \ldots, m$.

En el resto de los casos la forma de estimar $\widehat{\pi}_k$ y $\widehat{\mu}_k$ no varía; sin embargo, la forma de estimar $\widehat{\Sigma}_k$ debe obtenerse por medio de procedimientos iterativos (Celeux & Govaert 1995).

La resolución de estas ecuaciones conduce a un nuevo vector de parámetros $\widehat{\theta}^{(1)}$, y el algoritmo debe iterar hasta obtener la convergencia.

2.3. Selección del modelo

Para seleccionar un modelo se calcula el BIC (Schwarz 1978) para cada $m = 1, \ldots, M$ y para cada una de las hipótesis hechas sobre las matrices Σ_k de los modelos componentes y se elige la combinación que maximice dicho criterio.

2.4. Construcción de los grupos

Con la información de las secciones anteriores, se puede definir la siguiente estrategia:

- Determinar un número máximo de conglomerados (M) a trabajar y un conjunto de modelos mezclados a considerar.
- Aglomerar jerárquicamente los conglomerados para maximizar la verosimilitud de clasificación de cada modelo y obtener la clasificación hasta M grupos.
- Implementar el algoritmo EM para cada modelo y para cada número de conglomerados $2, \ldots, M$, comenzando con la aglomeración jerárquica.
- Calcular el BIC para cada modelo y para cada cantidad de conglomerados.

3. Aplicación práctica

La aplicación que se presenta es parte de la búsqueda de similitudes y patrones de la totalidad de los municipios de Venezuela (parroquias, en el caso del Distrito Capital), en cuanto a una serie de variables socioeconómicas. El objetivo del estudio era encontrar grupos de municipios semejantes entre sí (véase López 2007).

La matriz de entrada es de tamaño (366×6: 366 municipios $\times$ 6 componentes), que se corresponde a los 6 componentes retenidos en un Análisis de Componentes Principales no Paramétrico (ACPnP) (Lebart et al. 1984) sobre un conjunto de datos de más de treinta variables. Se retuvo el número de componentes que tuvieran autovalor mayor que la unidad. Eso se ve cumplido con el autovalor sexto. Este autovalor explica aproximadamente el 3 % de la varianza total y hasta él se explica el 67 % de la variabilidad total.

Tabla 1: Autovalores.

	ACPnP				
	Autovalor	Lím. Inf.	Lím. Sup.	% Var.	% Var. Acum.
1	12.56	11.19	14.29	39.24	39.24
2	3.33	2.97	3.79	10.40	49.64
3	2.03	1.81	2.32	6.36	56.00
4	1.54	1.37	1.75	4.82	60.81
5	1.16	1.03	1.32	3.61	64.43
6	0.94	0.84	1.08	2.95	67.38
7	0.79	0.70	0.90	2.46	69.84
⋮	...	...	...	...	...

Al primer componente le dan la misma contribución, más o menos, todas las variables, a excepción de unas pocas que, entre ellas, no aportan ni el 1 %. Este componente ubica del lado positivo algunas variables que de alguna u otra forma denotan una calidad de vida superior (porcentaje de familias que poseen carro, alfabetismo, IDH, camas por hospital, etc.) y del lado negativo están variables asociadas a problemas sociales: índice de masculinidad, porcentaje de familias sin acceso a aseo urbano, porcentaje de familias con casa propia, etc. Así, se bautizará este componente como 'Factor Estatus Social'.

Para el segundo componente solo dos variables le contribuyen en más de un 10 %: porcentaje de personas que asiste a una institución educativa y déficit funcional de viviendas de la entidad a la que pertenece el municipio. Este componente coloca del lado positivo variables de desarrollo humano en general y del lado negativo la tasa de fecundidad, tasa de natalidad, etc. Se nombrará 'Factor Expectativa de Vida'.

El tercer componente tiene influencia clara de porcentaje de personas con casa propia (16.99 %), porcentaje de familias que viven en una vivienda nuclear (14.3 %), porcentaje de extranjeros en el municipio (10.11 %), tasa de actividad (12.7 %) y sector público (17.61 %). Entre todas suman 71.71 %. Este componente se denominará 'Factor de Viviendas'.

Lo primero que se mostrará, con relación al análisis de Fraley y Raftery, será la tabla 2. Esta tabla muestra el BIC calculado para cada parametrización supuesta de los datos con cada número de conglomerado introducido. Se ha permitido que el programa actúe con grupos desde dos hasta nueve. En la tabla aparece en negrilla el valor BIC más alto.

Una vez entendida esta salida, se averigua qué conglomerado le corresponde a cada uno de los municipios. Para dar respuesta a esta pregunta nuevamente se siguieron los pasos dados por los autores (ver sección 2.4) y se utilizó el programa por ellos proporcionado (Fraley & Raftery 2006).

Al revisar la tabla 2 se observa que se establecen nueve conglomerados partiendo del supuesto que los datos siguen una distribución normal elipsoidal de igual volumen, igual forma e igual orientación (`EEE`).

Tabla 2: Criterio BIC.

	Características de la distribución					
Estratos	EII	VII	EEI	VVI	EEE	VVV
			ACPnP			
1	−9082.27	−9082.27	−8125.51	−8125.51	−8208.14	−8208.14
2	−8378.57	−8362.42	−8122.85	−8065.05	−8060.31	−8049.93
3	−8270.93	−8232.93	−8111.74	−8076.89	−8120.42	−8060.73
4	−8187.76	−8176.40	−8097.53	−8120.82	−8039.16	−8068.69
5	−8141.82	−8105.30	−8075.15	−8094.53	−8056.85	−8075.51
6	−8151.56	−8077.89	−8087.12	−8120.20	−8073.43	−8219.51
7	−8126.25	−8073.36	−8086.76	−8152.46	−8070.67	−8248.18
8	−8100.34	−8067.95	−8062.97	−8187.15	−8024.77	−8254.13
9	−8088.74	−8053.88	−8041.99	−8213.47	**−7992.37**	−8380.45

La figura 1 muestra que el mejor modelo que representa los datos es aquel cuyas matrices de covarianza estimadas son del tipo `EEE` y se maximiza con nueve grupos.

Se puede observar, en la figura 2, cómo se agrupan los conglomerados para los componentes primero y segundo, ubicando en los extremos del componente primero los grupos 7, 1 y 6, 9. Para el componente segundo, es claro que el grupo 2 y el 4 están ubicados hacia los valores más negativos y positivos, respectivamente. La figura 3 muestra un gráfico de dispersión matricial para todos los componentes. En esta figura destaca que los tres primeros componentes separan mejor los grupos. Se harán más comentarios al respecto en la sección 4.

3.1. Identificación de valores atípicos

Si bien Fraley & Raftery (2002) sugieren un método para encontrar atípicos, en este trabajo se optó, por comodidad, por seguir otro procedimiento: Johnson & Wichern (1998) utilizan una serie de pasos para detectar valores atípicos multivariantes: además de realizar las inspecciones gráficas de rutina, proponen calcular la distancia cuadrada generalizada

$$d^2 = (x_j - \overline{x})^T S^{-1} (x_j - \overline{x}) \qquad j = 1, \ldots, n \tag{15}$$

y examinarla para valores grandes. Esos valores deben compararse con un valor crítico específico. El valor está dado por la distribución chi-cuadrado $\chi^2_{p,0.005}$, donde p es la dimensión de los datos.

La tabla 3 contiene los valores d_i^2 que resultaron ser más grandes que el valor crítico establecido.

Es importante notar que los siete municipios que conforman el estado Amazonas y los cuatro que conforman el estado Delta Amacuro resultan ser atípicos. Existen varias maneras de explicar esto: en primer lugar, son estados desprovistos de muchas condiciones que otros ostentan. Los municipios del estado Amazonas y del estado Delta Amacuro son municipios sin grandes ciudades y sin vistosas infraestructuras y donde la mayor fuente de trabajo está de la mano del sector

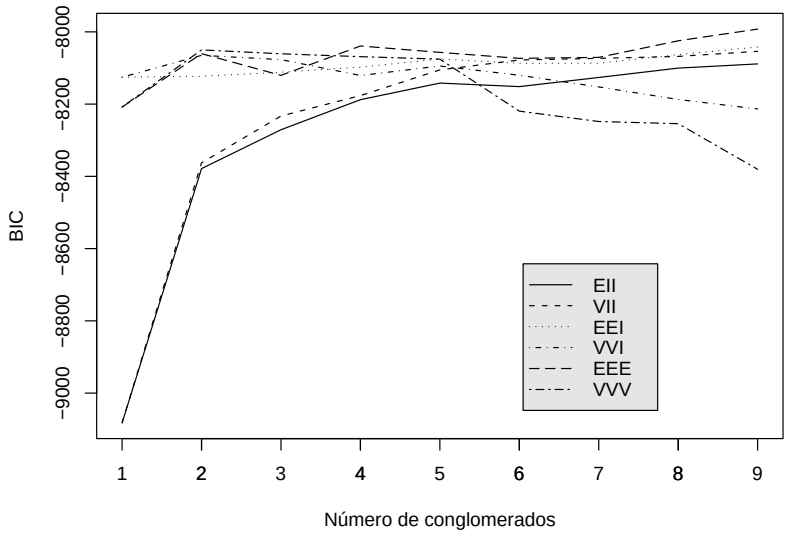

Figura 1: ACPnP. BIC. Criterio bayesiano. El BIC se maximiza para el modelo EEE y para nueve grupos.

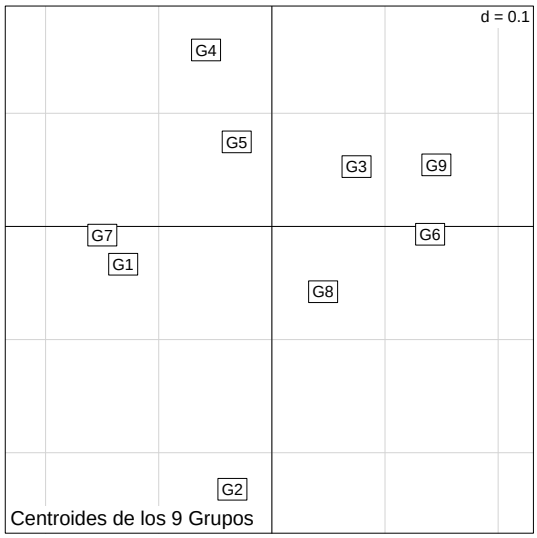

Figura 2: ACPnP. Centroides para los 9 grupos. Componente 1 y componente 2.

público (61.15 % para Amazonas y 44.95 % para Delta Amacuro; la media total de toda Venezuela es 20.19 %). Además, son poblaciones mayoritariamente indígenas, cuestión que los diferencia claramente de las demás entidades.

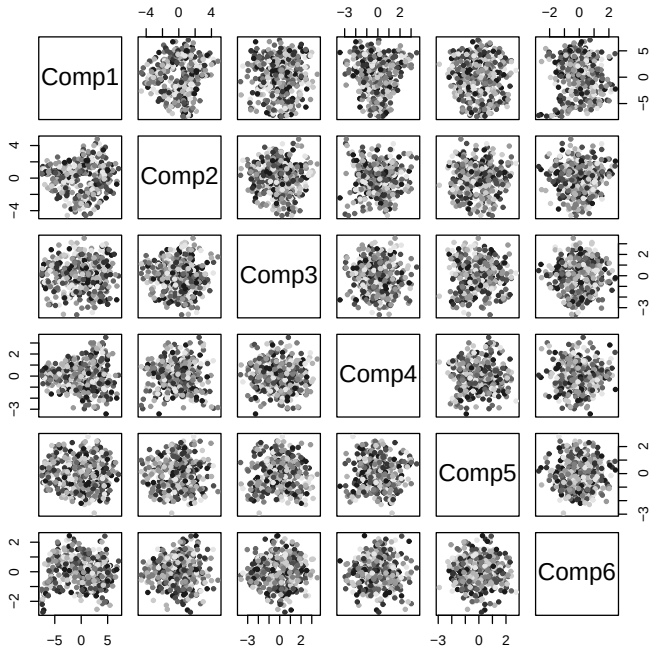

FIGURA 3: Gráfico de dispersión matricial para los seis componentes retenidos.

En ningún caso se removió ningún municipio porque la búsqueda de los grupos era a nivel nacional y debían considerarse todos.

4. Conclusiones

El análisis exploratorio previo puso de manifiesto la relación clara entre los bienes materiales (televisor, computadora, nevera, carro, ...). Esto es fácil de comprender debido a que cuanto más ingresos tenga una persona, más gastará en tales objetos. Si está en capacidad de tener internet en su casa, un individuo, naturalmente, tendrá facilidad para mantener, al menos, una computadora. Además, los bienes medidos son de amplio acceso para la mayoría de las familias.

Se evidenció también una relación directa con las variables referentes a los servicios del hogar (electricidad, tuberías de agua, excretas, ...). Cuando una familia no tiene acceso a un sistema de cañerías de calidad, una de las posibles causas es la ausencia de tuberías funcionales.

Los grupos más beneficiados en términos socioeconómicos son el noveno y sexto, los cuales están formados principalmente por los municipios de Caracas, Miranda, Carabobo, Nueva Esparta y Aragua. Barinas, Cojedes y Portuguesa son la cara contraria.

Tabla 3: Municipios atípicos multivariantes.

Municipio	d^2	Municipio	d^2
AmAltoOrinoco	195.61	MeAChacon	89.95
AmAtabapo	59.78	MeLibertador	58.14
AmAtures	59.87	MiBaruta	62.97
AmAutana	108.63	MiChacao	60.54
AmMaroa	104.22	MiElHatillo	109.75
AmManapiare	68.13	MiLSalias	52.92
AmRioNegro	111.49	MoAcosta	57.95
AnFranCarmCarv	52.35	MoUracoa	57.69
AnTDBUrbaneja	74.28	PoSRosalia	60.50
ApPCamejo	53.83	TaPMUrena	64.59
ArTovar	55.37	TaRUrdaneta	56.23
BaArismendi	67.30	VaCaruao	53.04
DelAntDiaz	214.03	VaElJunko	87.81
DelCasacoima	84.20	YaBolivar	102.30
DelPedernales	99.52	ZuAPadilla	72.63
DelTucupita	81.21	ZuColon	82.78
FaUrumaco	59.74	ZuJMSemprun	57.50
GuSGdGuayabal	58.37	ZuPaez	119.47
MeAricagua	115.79	DCCatedral	61.55

Nota: Teniendo en cuenta que son veintisiete variables, el valor crítico de la distribución χ^2 es 49.64492.

Los municipios de Trujillo, Cojedes, Amazonas, Guárico, Portuguesa y Delta Amacuro (que, en general, son del grupo segundo) tienen una esperanza de vida menor, altas tasas de natalidad y un mayor número de personas trabajando para el sector público.

Los grupos cuarto y quinto están en una posición media en cuanto a nivel socioeconómico, pero, a diferencia del grupo segundo, muestran una mayor esperanza de vida, más camas por hospital (en promedio), más calidad de vida (reflejada en el IDH) y un mayor número de personas trabajando en el sector informal.

Por su parte, el grupo octavo, con altas condiciones de vida, está en el cuadrante cuarto porque presenta mayor número de viviendas nucleares que sus contrapartes del primer cuadrante. Esto quiere decir que son municipios donde las casas se utilizan como núcleos familiares. Esto tal vez sugiera que en otros municipios, con un mismo nivel socioeconómico que el grupo octavo (como el grupo tercero), las casas no se utilicen como viviendas familiares. La ubicación de los grupos, según los municipios, en un mapa de Venezuela se muestra en la figura 4.

Como conclusiones referidas al conjunto de datos se tiene que la distribución normal no se muestra en la mayoría de las variables involucradas. La distribución multinormal es difícilmente sostenida en el conjunto de datos total; no obstante, una vez formados los grupos, esta se sostiene en la mayoría de ellos. Esto, además de ser valiosísimo para la aplicación de la técnica central de este trabajo, el análisis de conglomerados de Fraley & Raftery (2002), supone la posibilidad de utilización de otras técnicas que requieren normalidad en cada grupo.

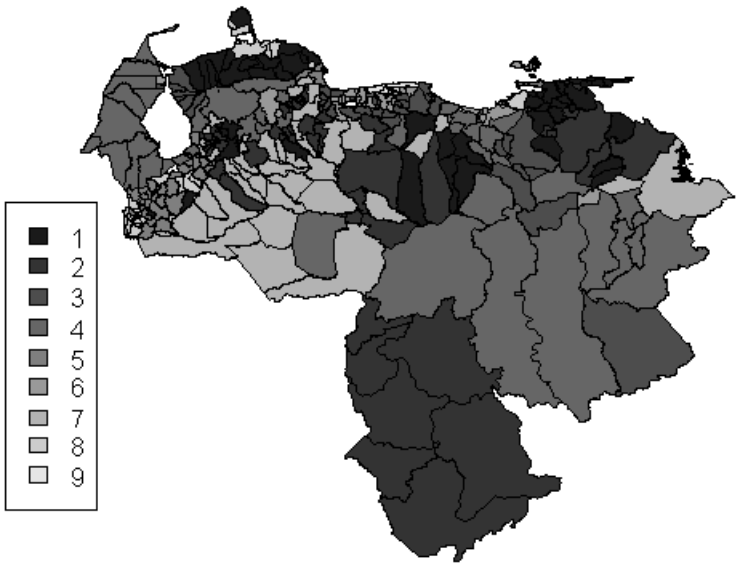

Figura 4: Relación geográfica de los 9 grupos encontrados.

Agradecimientos

Al profesor Segundo Quiroz por la ayuda prestada al comienzo de esta investigación y a los árbitros anónimos que contribuyeron a enriquecer y mejorar este trabajo.

[Recibido: septiembre de 2008 — Aceptado: octubre de 2009]

Referencias

Bandfield, J. & Raftery, A. (1993), 'Model-based Gaussian and Non-Gaussian Clustering', *Biometrics* **49**, 803–821.

Bergonzoli, G. (2006), *Sala situacional*, IAESP. Instrumento para la vigilancia de salud pública.

Bouveyron, C., Girard, S. & Schmid, C. (2007), 'High-Dimensional Data Clustering', *Computational Statistics & Data Analysis* **52**(1), 502–519.

Celeux, G. & Govaert, G. (1995), 'Gaussian Parsimonious Clustering Models', *Pattern Recognition* **28**, 781–793.

Díaz, L. (2002), *Estadística multivariada: inferencia y métodos*, 1 edn, McGraw-Hill, Bogotá, Colombia.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum Likelihood Estimation from Incomplete Data Via the EM Algorithm', *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38.

Fraley, C. (1998), 'Algorithms for Model-Based Gaussian Hierarchical Clustering', *SIAM Journal on Scientific Computing* **20**(1), 270–281.

Fraley, C. & Raftery, A. E. (2002), 'Model-Based Clustering, Discriminant Analysis, and Density Estimation', *Journal of the American Statistical Association* **97**.

Fraley, C. & Raftery, A. E. (2006), 'MCLUST: Model-based cluster analysis'. R port by Ron Wehrens. R package version 2.1-14.

Gallegos, M. T. & Ritter, G. (2005), 'A Robust Method for Cluster Analysis', *The Annals of Statistics* **33**, 347–380.

Gnanadesikan, R., Kettenring, J. R. & Maloor, S. (2007), 'Better Alternatives to Current Methods of Scaling and Weighting Data for Cluster Analysis', *Journal of Statistical Planning and Inference* **137**, 3483–3496.

Haughton, D., Legrand, P. & Woolford, S. (2007), 'Review of Three Latent Class Cluster Analysis Packages: Latent Gold, poLCA, and MCLUST', *The American Statistician* **63**(1), 81–91.

INE (2005*a*), *Censo 2001 por municipios y parroquias. Tabulados prioritarios*, C.D Instituto Nacional de Estadística.

INE (2005*b*), 'Instituto nacional de estadística'.
*http://www.ine.gov.ve

INE (2005*c*), *Venezuela: estadísticas vitales, 2004*, Instituto Nacional de Estadística.

Johnson, R. & Wichern, D. (1998), *Applied Multivariate Statistical Analysis*, 4 edn, Prentice Hall.

Lago, S., Mauro, M. & Álvarez, G. (2000), 'Análisis exploratorio multivariado. La conformación de subregiones al interior de cuatro provincias argentinas según el impacto del desarrollo en las condiciones de vida', *Cinta de Moebio* (9), 1–18.

Lebart, L., Morineau, A. & Warwick, K. M. (1984), *Multivariate Descriptive Statistical Analysis*, John Wiley & Sons, New York, United States.

Leisch, F. (2006), 'A Toolbox for K-Centroids Cluster Analysis', *Computational Statistics and Data Analysis* **51**(2), 526–544.

López, F. (2007), 'Búsqueda de estratos socieconómicos a nivel nacional. Caracterización estadística de los municipios de Venezuela'. Tesis para optar al título de Licenciado en Estadística. Universidad de Los Andes. Mérida, Venezuela.

López, N., Moreno, A., Medina, E., García, J., Rivera, E., Díaz, Y., Porcio, G., Sánchez, O., Aguirre, J., Ponce, X., Arias, J., Vivas, J. & Bergonzoli, G. (2002), *Identificación y representación de necesidades sociales. Módulo II*, Ministerio de Salud, Dirección de Análisis Estratégico.

Murtagh, F. & Raftery, A. (1984), 'Fitting Straight Lines to Point Patterns', *Pattern Recognition* **17**, 479–483.

Oh, M. S. & Raftery, A. (2007), 'Model-Based Clustering With Dissimilarities: A Bayesian Approach', *Journal of Computational and Graphical Statistics* **16**(3), 559–585.

Peña, D. (2004), *Análisis de datos multivariantes*, McGraw-Hill Interamericana.

Schwarz, G. (1978), 'Estimating the Dimension of a Model', *Annals of Statistics* **6**(2), 461–464.