UNA CLASIFICACIÓN DE LAS ZONAS RURALES DE MUNICIPIOS SIN PREDIALIZACIÓN CATASTRAL.

Hernando E. Mutis G.*

ABSTRACT

The main purpose of this paper is developing and applying a methodology based on multivariate statistical procedures to endorse a classification of certain rural regions in Colombia. In order to achieve the aims, cluster analyses were the basic procedures made. The techniques of principal components, canonical variates, discriminant analysis and Andrews´ plots were used to validate the cluster analyses results.

PALABRAS CLAVE

Estratificación rural, formación predial catastral, análisis de conglomerados, análisis de componentes principales, análisis de variable canónica, análisis discriminante, curvas de Andrews.

1. PRESENTACIÓN

La metodología para la estratificación de fincas y viviendas dispersas en la zona rural que se realiza en el país la ha desarrollado el Departamento Nacional de Planeación (DNP) y se basa en la escogencia de uno de dos criterios: los puntajes de la calidad de la vivienda rural y los propios de la capacidad productiva potencial de su predio.

Para el primer criterio, el Instituto Geográfico Agustín Codazzi (IGAC) ha desarrollado –y fundamentado– una ficha predial catastral para obtener un puntaje que mide la calidad de la vivienda. Respecto al segundo criterio, el DNP ha realizado una metodología que se fundamenta en dos fuentes. Una, en los estudios del IGAC de las llamadas Zonas Homogéneas Geoeconómicas, las cuales expresan una jerarquización de la capacidad de explotación económica del suelo y que se formula en valores catastrales. Otra, los estudios que el IICA del Ministerio de Agricultura ha efectuado sobre la Unidad Agrícola Familiar: ésta representa[1] el número de hectáreas requeridas para que una explotación campesina (agrícola, pecuaria, forestal, acuícola o mixta) genere ingresos anuales equivalentes a 1080 salarios mínimos diarios legales en condiciones de productividad promedio en el municipio.

Los criterios básicos parten, entonces, de los procesos de formación predial catastral. Existe un grupo de cerca de 60 municipios en el país a los cuales el procedimiento de estratificación rural del DNP no puede aplicarse precisamente por la carencia –por diversas razones– de la predialización catastral de las regiones rurales.

El propósito de este material es desarrollar y aplicar una metodología basada en procedimientos de métodos estadísticos multivariados

* Profesor Asociado, Departamento de Ingeniería Industrial, Universidad de Los Andes, hmutis@uniandes.edu.co

1 SINTAP (1992) e IICA–MINAGRICULTURA (1994).

que sirva de sustentación para una clasificación de las áreas rurales del grupo mencionado de municipios.

Para lograr tal propósito se utilizaron diversas técnicas de análisis de conglomerados y se convalidaron los resultados con diferentes métodos: componentes principales, análisis de variables canónicas y curvas de Andrews, entre otros.

Este documento está dividido en varias secciones: en la segunda se aborda el problema de la especificación de las variables y de la disponibilidad de la información. En la tercera se presentan los procedimientos y métodos utilizados y los técnicas de validación con una breve discusión sobre cada uno de ellos y termina con la descripción del diseño metodológico del conjunto de procedimientos utilizado. Finalmente, en la cuarta sección, se analizan los principales resultados, se verifican los procesos de validación de la clasificación definitiva obtenida y se acompañan las conclusiones pertinentes.

2. La información

La especificación de las variables es un problema inicial básico en cualquier tipo de análisis estadístico. El conjunto inicial de municipios para los cuales sus zonas rurales deben ser clasificadas contó, en principio, con un acervo de información basado en variables de tipo demográfico, de cobertura de servicios públicos, de indicadores de calidad de vida rural, sobre tamaños de predios y sobre la Unidad Agrícola Familiar promedio municipal[2]. La deficiencia de la información es de dos tipos: (i) cubrimiento: inexistencia de todos los rubros para cada uno de los municipios; (ii) calidad: algunas de las variables existentes presentaban inconsistencias y se cuestionaba la confiabilidad de su utilización.

El segundo tipo de problemas afectó fundamentalmente a las variables demográficas (población del resto rural de 1999, su proporción con respecto al total, población dispersa, densidad, número de hogares rurales y número de personas por hogar). El segundo tipo de deficiencias influyó en la disminución del número de municipios (de 60 a 43) para los cuales se aplicó la metodología desarrollada en la sección tercera.

En conclusión, las variables que fueron finalmente utilizadas correspondieron a los porcentajes de cobertura de energía rural y de acueducto rural, el Índice de Calidad de Vida Rural, la Unidad Agrícola Familiar y el porcentaje de hogares con Necesidades Básicas Insatisfechas[3].

Con el objeto de lograr una clasificación rural de las zonas no urbanas de los municipios, se propone una técnica que está diseñada para la obtención de agrupaciones en un conjunto de p variables en n observaciones: es la llamada análisis de agrupaciones. Este será el tema que se tratará en la siguiente sección.

3. Procedimientos y métodos

En esta sección se presenta una síntesis de las técnicas que se conocen como análisis de conglomerados o análisis de agrupaciones. La presentación está guiada por una fundamentación metodológica que indique los procedimientos que se podrían seguir con el propósito de lograr la clasificación rural de los municipios. La presentación está acompañada por una discusión sobre las técnicas de validación de los resultados y forma parte de ese entorno

2 Las fuentes fueron el DANE, el DNP y el Ministerio de Agricultura.

3 Los datos se refieren a 1998 y 1999.

metodológico. Termina la sección con la presentación del diseño metodológico utilizado para la obtención de la clasificación de las regiones rurales de los municipios.

El propósito de los análisis de conglomerados es la detección de aglomerados de un conjunto de n elementos que están descritos por un grupo de valores de p variables de acuerdo a una medida de distancia o disimilaridad. Un subconjunto de tales elementos forma una agrupación si éstos exhiben valores similares en las medidas de distancia o similaridad en tanto otras agrupaciones difieren de ese subconjunto. Es decir, se trata de obtener agrupaciones en las cuales la variabilidad al interior **de los grupos o aglomerados sea la mínima posible y la variabilidad con respecto a otras agrupaciones –o variabilidad** entre **grupos–, sea la máxima posible.**

Los problemas básicos que se debe enfrentar en análisis de conglomerados son: (i) la selección de variables de entrada, (ii) la escogencia apropiada de la medida de distancia, (iii) los procedimientos numéricos de agrupación y, (iv) el número de grupos.

3.1 La selección de variables.

Este es parte del problema básico de especificación en los métodos y modelos estadísticos. Se asume que se escoge el grupo de variables que es, de alguna manera, "explicativo" del proceso de clasificación que se persigue. Pero, más de las ocasiones el problema tiene que ver con las limitación en la obtención de las variables deseables en el análisis. Una alternativa muy popular es la utilización de los puntajes de componentes principales aplicados sobre el grupo de variables. El asunto de la selección tiene que ver, también, con las unidades de medición de las variables y, cuando la variabilidad de una o de algunas de ellas pesa o pesan fuertemente sobre el conjunto, se acostumbra homogeneizarlas a través de un proceso como la estandarización. Los puntajes de componentes principales poseen, también, una propiedad de uniformización.

3.2 Las medidas de distancia.

Una distancia, o mejor, una función distancia d(x,y) se define sobre un conjunto de puntos como una función numérica que satisface:

i) $d_{(i,j)} = d_{(j,i)}$ i.e. la distancia entre el elemento i y el elemento j es la misma que la distancia entre j e i.

ii) $d_{(i,j)}$ e•0 i.e. la distancia entre i y j no debe ser negativa.

iii) $d_{(i,i)} = 0$ i.e. la distancia de un elemento a sí mismo debe ser cero

iv) $d_{(i,j)} = 0$ si y sólo si $i = j$

v) $d_{(i,j)}$ d»#ð $d_{(i,k)} + d_{(k,j)}$ lo cual es conocido como desigualdad triangular..

Una de las medidas de distancia más comunes es la Euclidiana y su popularidad se debe, en parte, a sus propiedades geométricas: el cuadrado de $d_{(i,j)}$ puede interpretarse como una hipotenusa en el espacio p dimensional y es invariante bajo transformaciones ortogonales. Si la variabilidad de un grupo de variables pesa excesivamente sobre todo el conjunto seleccionado para analizar, la medida de distancia aconsejable es la distancia de Mahalanobis, la cual corresponde a la Euclidiana cuando se hace el ajuste por variabilidad a todo el conjunto.

3.3 Métodos de agrupamiento y algoritmos.

Existen varios métodos de agrupamiento, los cuales, hablando en un sentido amplio, corresponden a diferentes criterios de clasificación. Siguiendo a Everitt (1993), y a pesar de que no hay una única tipología, se podrían distinguir métodos jerárquicos y de optimización.

Los agrupamientos jerárquicos se distinguen por una sucesión de subdivisiones o subclasificaciones de la información en, por ejemplo, g grupos, con g variando de 1 hasta n, el total de los elementos. Bien puede empezarse con todos los elementos en un solo grupo y, secuencialmente, de acuerdo al criterio de distancia empleado, se procede a encontrar grupos más reducidos terminando en que cada elemento se comporta como un grupo. Otra forma parte a la inversa: cada elemento se consi-

dera un grupo separado y se forma un primer conglomerado con aquellos dos elementos más cercanos, de acuerdo a la medida de distancia; el proceso se continúa hasta obtener un conjunto o grupo conformado por todos los elementos. En el camino se va construyendo un cierto número de conglomerados.

Dentro de los métodos jerárquicos más conocidos están el del vecino más cercano, el del vecino más lejano, el del promedio de todos los pares de elementos, el del centroide, el de la mediana, el flexible, el EML y los de mínima varianza. El método de Ward corresponde a estos últimos. Los métodos jerárquicos comparten su énfasis en la estructura de la información a diferentes niveles y en la forma en que la estructura cambia. Estas relaciones se sintetizan en un diagrama de árbol llamado dendograma el cual muestra la secuencia de fusiones sucesivas en dos dimensiones: una representa a los conglomerados y la otra la medida de distancia o disimilaridad empleada. Diferentes niveles de agrupamiento se obtienen a través de cortes en la medida de distancia.

En general, diferentes métodos conducen a diferentes agrupamientos. Es por esto que se aconseja precaución en la adopción de los resultados de los análisis de conglomerados (Chatfield & Collins, 1980, p. 229; Everitt 1993; Krzanoswki, 1990, pp. 94-96; Marriot, 1974, p. 780). Cada método tiene sus propias características, pero algunos métodos parecen comportarse mejor que otros. En los estudios sobre el tema, Everitt (1979) comenta que el método del vecino más cercano es menos exitoso que otros métodos, particularmente cuando se lo comparaba con el método de Ward. Al considerar siete métodos aglomerativos Spath (1980, p. 193) recomienda el de Ward[4].

Los métodos que se basan en la optimización de una función objetivo reasignan elementos entre grupos. Una función objetivo muy utilizada es la conformación de conglomerados a través de la minimización de las sumas de cuadrados dentro de ellos. Regularmente estos procedimientos empiezan haciendo una partición de los elementos en K grupos y en la siguiente etapa se asigna un elemento al centroide (vector de medias de las variables del grupo) más cercano, éste se recalcula con cada nueva asignación y el proceso se detiene cuando nuevas asignaciones no conducen a nuevos cambios en los conglomerados. Cuando el algoritmo encuentra las agrupaciones finales se considera que la partición es localmente óptima. No obstante, no hay garantía de obtención de óptimos globales. El método más utilizado es el conocido como el de K-Medias. Debe mencionarse, además, que los algoritmos son sensibles a la definición inicial (semillas) de los grupos.

En conclusión sobre los procedimientos que se deben utilizar, quizás el mejor consejo es dado por Spath (1980) quien dice:

> «We ourselves and many others (Hodson 1970, Inglis and Johnson 1969, Lance and Williams 1976b, Wallace 1968) have found the KMEANS principle applied to data or distance matrices to be the most satisfactory. In contrast to the hierarchical methods where a branching or merging tree is optimally configured, in KMEANS or related methods an optimizing objective function exist for any given n.
> It would seem reasonable to use the results for a desired n of a hierarchical method, especially an agglomerative method, as an initial partition for an algorithm based on the KMEANS principle (see Anderberg

4 Chatfield y Collins sintetizan el procedimiento de Ward de la siguiente manera: «Ward's hierarchical clustering method is based on within-group sums of squares rather than on ways of linking groups. An agglomerative algorithm is used. At each stage the number of groups is reduced by one, by combining the two groups which give the smallest possible increase in the total within-group sum of squares. Of course, when we start with n groups of just one individual, the total within-group sum of squares is zero» (Chatfield & Collins, 1980, p. 224)

1973), since less computer time might be taken than is necessary with the heuristic algorithms in section 3.1 in each case. The effect seems to depend on the particular data set and on the value of m.» (Spath 1980, p. 194)[5].

Milligan (1980) también aconseja el procedimiento de K-Medias dado una información previa proveniente de algún algoritmo jerárquico.

Adicionalmente a los procedimientos recomendados, se sugiere que los resultados de análisis de conglomerados sean comparados y revalidados utilizando otras técnicas multivariadas. Esto es claramente indicado por Marriot:

«I suggest further that the results of any cluster analysis, whatever method has been used, should be accompanied by scatter diagrams, based on principal components or canonical variates, to illustrate how the clusters found relate to the distribution of the points. A reader sceptical about the numerical analysis (as we well might be) can then judge for himself how well the proposed clusters reflect the structure of the data» (Marriot, 1974, p. 78).

3.4 Sobre el número de grupos.

Con el fin de determinar el número de agrupaciones que se debe formar se han sugerido varias técnicas. Una usa el criterio de graficar la razón de trazas de las matrices de covarianzas entre grupos (B) y dentro de grupos (W), ponderadas por sus respectivos grados de libertad, para los diferentes aglomeraciones formadas. Este criterio es el llamado de la pseudo-F[6] (Everitt, 1974, página 60) y es prestado de la razón F del análisis de varianza. Aquí se debe recordar que la traza de una matriz de covarianzas corresponde al concepto de varianza total. El máximo valor del pseudo-F sugeriría el número de grupos, porque correspondería a la mayor razón entre varianzas totales.

Otra aproximación usa el criterio de relacionar el determinante de la matriz de covarianzas dentro de grupos –ponderada por el cuadrado del número de grupos– (g^2 |W|) con el número de conglomerados y se conoce como el criterio de Marriot. Como el determinante de una matriz de covarianzas se refiere al concepto de varianza generalizada, entonces el número aconsejado de grupos correspondería a aquel número en el que se alcance el menor valor del criterio de Marriot por cuanto un grupo se caracteriza por contar con una mínima variabilidad dentro de grupos.

Otro enfoque se conoce como el CCC por sus iniciales en inglés (Cubic Clustering Criterion) (Sarle, 1983). Milligan and Cooper (1985) informaron que el criterio CCC se comportó mucho mejor que otros en un estudio al respecto. El CCC se basa «on the assumption that clusters obtained from a uniform distribution on a hyperbox are hypercubes of the same size. The hypercube assumption is obviously false in most cases, but is generally conservative unless the number of clusters is very large in two or more dimensions.» (Sarle, 1983, p. 4) El número de grupos recomendado es aquel en que el valor del CCC alcanza un máximo.

No obstante, no hay una medida exclusiva que determine sin lugar a dudas el número de agrupaciones: estos criterios han fallado en la determinación del número correcto de grupos usando simulaciones (Everitt 1974, pp. 73-76, Everitt (1979) y Milligan and Cooper (1985).

Es posible usar dos enfoques diferentes para revisar y verificar el número y la consistencia

5 En el párrafo n es el número de conglomerados y m es el número de observaciones.

6 $Pseudo\ F = \frac{traza(B)/g-1}{traza(W)/N-g}$ Donde: B: matriz de covarianzas entre grupos; W: matriz de covarianzas dentro de grupos; g: número de grupos y N: número total de observaciones.

de las agrupaciones: análisis de variable canónica y gráficas de Andrews, sin volver a mencionar las gráficas de dispersión de los puntajes de componentes principales por grupos.

Krzanowski (1990, p. 44, pp. 291-306), Everitt & Dunn (1993), y Stevens (1992, p. 410) recomiendan la graficación de las dos o tres primeras variables canónicas como un método para detectar y verificar un conjunto de agrupaciones. La técnica se basa en la derivación de los eigenvectores de la matriz $W^{-1}B$ (con W y B como se definieron anteriormente). Se trata de obtener nuevas variables con el criterio de la maximización de la razón de la variabilidad entre grupos (B) con respecto a la variabilidad dentro de grupos (W). La idea es construir unas nuevas variables que enfaticen las diferencias entre g grupos en un conjunto de datos de p variables tomadas sobre n elementos. Si llamamos a y_i a la iésima variable canónica, ésta se obtendría de $y_i = a^t X_i$, donde el vector a es el correspondiente eigenvector de la matriz $W^{-1}B$. Si los grupos están bien conformados, en las gráficas de variables canónicas se encontraría una clara separación entre ellos.

Las gráficas de Andrews (Andrews, 1972) son funciones matemáticas basadas en transformaciones de Fourier. El vector p-dimensional de observaciones $x^t = [x_1, x_2, ..., x_p]$ es transformado por el vector $f_{(t)} = [\ 2^{-1/2}, \text{Sin } t_o, \text{Cos } t_o, \text{Sin } t_o, ...\ ]$ dando $f_x(t) = f_{(t)}\ x$. La función $f_x(t)$ posee algunas propiedades importantes: preserva las medias, las distancias euclidianas y las varianzas de las variables X.

Por estas razones, las gráficas de $f_x(t)$ sobre el rango $-p < t < p$ son útiles en la detección de la configuración de agrupaciones de una información multivariada. Si los patrones de las gráficas distinguen comportamientos diferentes para los grupos, eso daría lugar a considerar que la estructura de agrupaciones está bien concebida. Si los patrones se confunden, se consideraría que la estructura encontrada no corresponde grupos bien conformados.

Otro posible criterio que ayuda en la coherencia de la clasificación obtenida es el análisis discriminante utilizando rutinas particulares orientadas a detectar el número de elementos incorrectamente clasificado a partir de procesos de validación utilizando técnicas "bootstrapping" para evitar el sesgo de clasificación debido a la aplicación de una regla basada en el mismo conjunto de datos. Una limitación del análisis discriminante es la exigencia de distribución normal multivariada, la cual no se cumple para el conjunto de datos analizado en este documento[7].

3.5 Conclusión.

El procedimiento para realizar la clasificación de los sectores rurales de los municipios de interés sigue la metodología expresada en esta sección. Utilizando como entradas bien a las variables en sus unidades estandarizadas o bien a los puntajes de componentes principales se aplicaron los distintos algoritmos de agrupación. De cada resultado se obtenía el dendograma, se capturaban los centroides, y, con el fin de refinar el proceso, éstos fueron las semillas para el algoritmo de K-Medias. En el paso siguiente se analizaban las gráficas de dispersión de los conglomerados obtenidos con respecto a las variables canónicas, los componentes principales, las curvas de Andrews y se examinaba el comportamiento de medidas como el CCC, el pseudo F y el pseudo T^2, de acuerdo al número diferente de grupos sugerido y esto se hacía tanto para los resultados de los algoritmos jerárquicos como para los correspondientes del algoritmo de K-medias. Estos pasos se repitieron, alterando el nivel de corte del árbol jerárquico –y por consiguiente el número de centroides del algoritmo de K-

[4] Para la prueba de distribución normal multivariada se utilizó el enfoque de Mardia (1974 y 1979).

Medias– con el fin de obtener 2, 3, ..., hasta 10 grupos.

Los resultados finales son presentados en la siguiente sección.

4. Análisis de resultados.

En esta sección se sintetizan los resultados básicos obtenidos de la aplicación de la metodología esbozada en la sección tercera. Se presentarán, primero, los grupos finales resultantes, luego se realizará una caracterización de los mismos, más adelante se validarán los grupos obtenidos y, finalmente, se resumirán las principales conclusiones.

4.1 Los Grupos.

Los municipios a los que se les aplicó la metodología desarrollada en la sección anterior fueron 43. Las variables utilizadas son las reportadas en la sección 2 y se detectaron dos agrupaciones. La clasificación final de los municipios resultó idéntica para los métodos jerárquicos de Ward, del vecino más cercano, del EML[8] y del método llamado Flexible. Así mismo aconteció con las salidas del método de K-Medias, cuando las semillas eran los centroides de las salidas anteriores. En la Gráfica No. 1 se presenta el árbol jerárquico producto del algoritmo de Ward y la clasificación resultante se resume en la siguiente tabla:

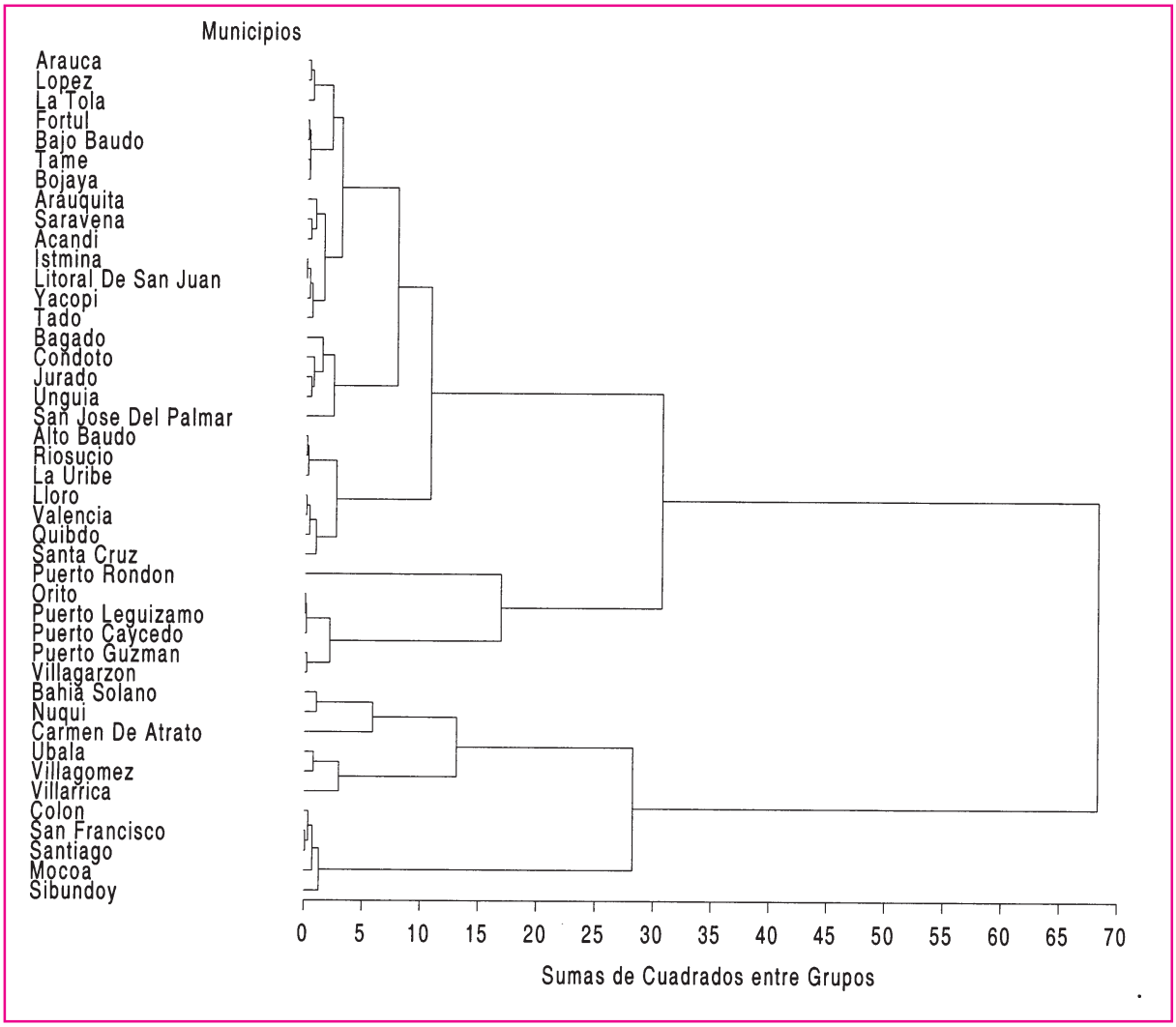
Tabla No. 1. Clasificación de la Zona Rural de los Municipios

	Grupo 1 (32 municipios):
Departamento	**Municipio**
ARAUCA	Fortul, Tame, Puerto Rondón, Arauca, Arauquita y Saravena
CAUCA	López
CHOCÖ	Alto Baudó, Bojayá, Riosucio, Bagadó, Unguía, Juradó, Bajo Baudó, San José del Palmar, Acandí, Lloro, Condoto, Quibdó. Istmina, Litral de San Juan y Tado.
CÓRDOBA	Valencia
CUNDINAMARCA	Yacopí
META	La Uribe
NARIÑO	Santa Cruz
NARIÑO	La Tola
PUTUMAYO	Puerto Caycedo, Orito, Puerto Guzmán, Villagarzón y Puerto Leguízamo
	Grupo 2 (11 municipios):
Departamento	**Municipio**
CHOCÓ	Bahía Solano, Nuquí y Carmen de Atrato
CUNDINAMARCA	Ubalá y Villagómez
PUTUMAYO	Mocoa, Santiago, Colón, San Francisco y Sibundoy
TOLIMA	Villarrica

[8] El método EML junta grupos mediante la maximización de la verosimilitud a cada nivel jerárquico asumiendo una mezcla normal multivariada, matrices de covarianzas esféricas iguales y probabilidades muestrales desiguales. (SAS Institute 1989, p. 534). Para el método llamado Flexible, ver también (SAS Institute 1989, p. 534).

El dendograma muestra la estructuración de la progresiva clasificación finalmente resultante. En el eje vertical están las ramas terminales del árbol y en el eje horizontal las sumas de cuadrados entre grupos y mide el nivel en el que se amalgamaron los elementos. Las zonas rurales de los municipios del grupo No. 2 se aprecian en la parte inferior del árbol jerárquico empezando en Bahía Solano, seguida de Nuquí y terminando en Sibundoy. Bahía Solano y Nuquí se juntan en uno de los primeros niveles del árbol, esto es indicativo de la gran similaridad que existe entre ellos. También Ubalá y Villagómez conforman un subgrupo desde la primera etapa. Análoga situación se presenta con San Francisco, Santiago y Colón, y después se les junta la zona rural de Mocoa. Siguiendo las sucesivas uniones se describe el mapa de cercanías o similaridades de la estructuración de las rurales de los municipios.

Gráfica No. 1. *Dendograma*



4.2 Las características de los grupos.

El grupo No. 2, compuesto por once municipios, se caracteriza porque comparte niveles relativamente altos –en comparación a los del primer grupo– del Índice de Calidad de Vida Rural y porcentajes, también relativamente altos –en relación al primer grupo– de las cobertura de energía rural y/o de acueducto rural. Siete de estos municipios poseen –como subgrupo– los niveles más bajos de los porcentajes de hogares con Necesidades Básicas Insa-

tisfechas, en tanto los otros cuatro presentan coberturas altas –relativamente– de energía y acueducto rurales.

Esta interpretación se facilita por una revisión de las correlaciones de las variables con los componentes principales. La tabla No. 2 presenta esa información. La primera columna menciona a las variables y las siguientes a cada uno de los tres primeros componentes. La columna etiquetada Prin1 se refiere al primer componente y así sucesivamente. Las correlaciones mencionadas se indican en la primer línea al frente del nombre de cada variable. El renglón que les sigue indica la probabilidad asociada a la hipótesis nula de que la correlación poblacional es cero.

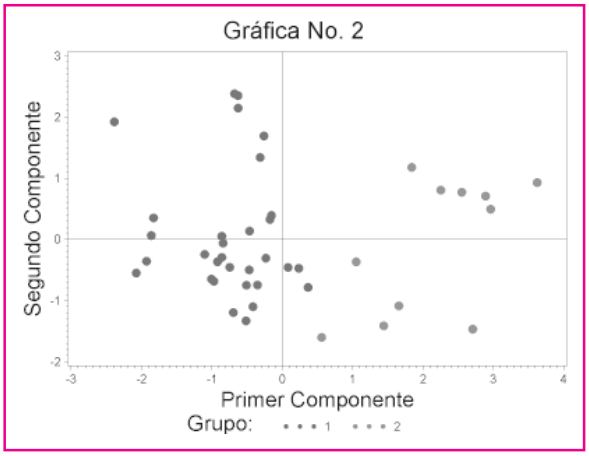
El primer componente está asociado a los porcentajes de cobertura de energía y acueducto rurales y al Índice de Calidad de Vida Rural. El segundo componente se relaciona con el porcentaje de hogares con NBI y la Unidad Agrícola Familiar y con el Índice de Calidad de Vida Rural. El tercer componente se refiere a un contraste entre la Unidad Agrícola Familiar y el porcentaje de hogares con NBI.

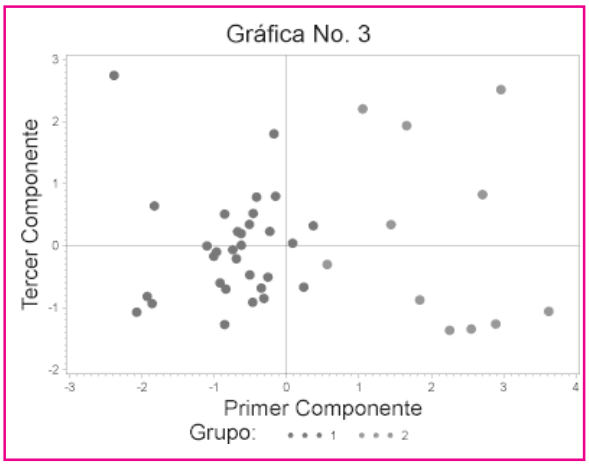
Tabla No. 2. *Correlaciones de las Variables con los Componentes Principales*

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations: 43			
	Prin1	Prin2	Prin3
Porcentaje de Cob. Energía Rural	0.82283 <.0001	-0.09529 0.5433	-0.24465 0.1139
Porcentaje de Cob. Acueducto Rural	0.86186 <.0001	0.00099 0.9950	0.27245 0.0771
Porcentaje de Hogs, con NBI	-0.18857 0.2259	0.65596 <.0001	-0.70556 <.0001
Ind. Calidad de Vida Rural	0.76736 <.0001	0.51161 0.0005	0.04438 0.7775
Unidad Agrícola Familiar	-0.30357 0.0478	0.63029 <.0001	0.66086 <.0001

4.3 La validación de los grupos

En cuanto a la determinación del número de grupos los criterios Pseudo F, Pseudo T^2, CCC y Marriot no arrojaron resultados concluyentes. Por esta razón la validación de los grupos se hace, acorde con la metodología esbozada, en primer lugar, con las gráficas de dispersión de los componentes, los cuales, de paso nos muestran las características mencionadas de los municipios. La Gráfica No. 2 presenta la dispersión de las zonas rurales de los municipios –cada uno está representado por un punto– de acuerdo a los dos primeros componentes. El primer componente se encuentra en el eje horizontal y el segundo componente en el eje vertical. Los grupos de municipios se encuentran en colores, correspondiendo el azul al grupo 1 y en rojo al grupo 2. Es clara la distinción entre los dos grupos que se exhibe en la gráfica.



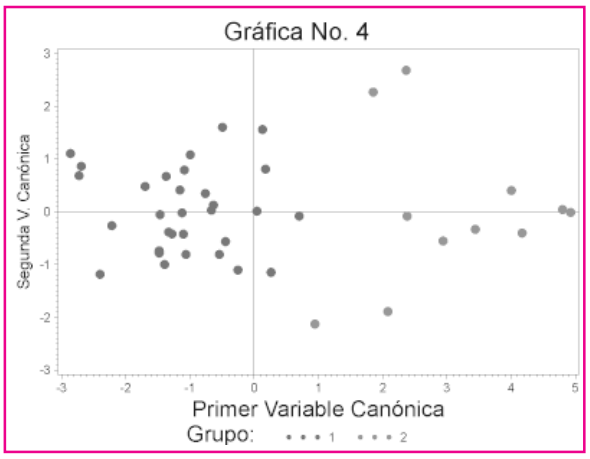


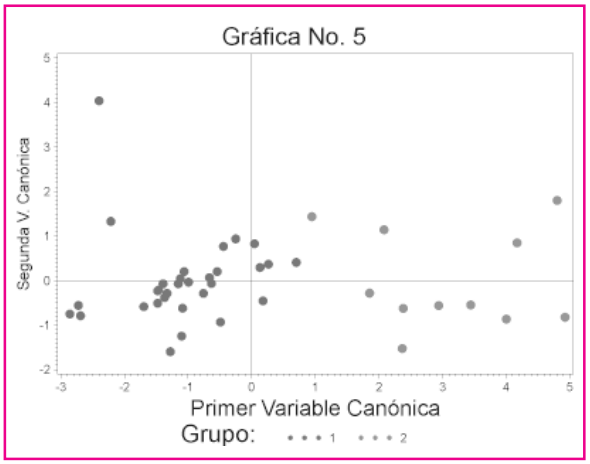
La Gráfica No. 3 se diferencia de la anterior en que el eje vertical es ahora ocupado por el tercer componente principal. En tanto la dispersión varía, la representación grupal muestra la coherencia de los grupos formados.

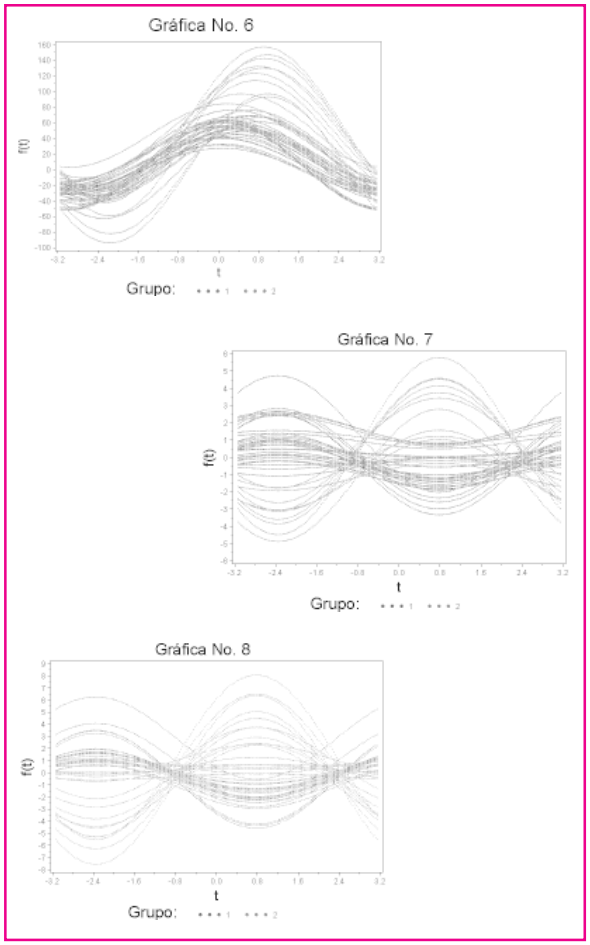
Para continuar con la validación en la construcción de los dos grupos, el turno es ahora para los diagramas de dispersión de las zonas rurales con respecto a las variables canónicas. Las Gráficas No. 4 y No. 5 presentan esos diagramas. En ambos casos la primera variable canónica se encuentra en el eje horizontal y la segunda en el eje vertical. Los grupos uno y dos están coloreados como en el caso de las gráficas de dispersión de los componentes: el azul para el grupo 1 y el rojo para el otro. La diferencia entre las dos gráficas se debe a las variables de entrada: en la No. 4, se usó como entrada al conjunto original de variables, mientras que en la No. 5 el conjunto de entrada corresponde a los cinco primeros componentes.

Análogamente a las representaciones basadas en componentes principales, la separación de los grupos es notoria, pero un énfasis mayor se puede apreciar en las Gráficas 4 y 5. Además, de estas últimas, a pesar de la diferencia en los diagramas de dispersión, se aprecia una nítida diferenciación de los dos grupos.

Las curvas de Andrews confirman la organización de la información en dos grupos. Las Gráficas Nros. 6, 7 y 8 son la representación de los dos grupos usando como entrada a las variables originales (Gráfica No. 6), a los componentes principales (Gráfica No. 7) y a las variables canónicas (Gráfica No. 8). Como en las gráficas anteriores, el grupo 2 se representa por el color rojo y el grupo 1 por el azul. En los tres casos en el eje horizontal se representa el ángulo entre -p y p y en el eje vertical la función de Andrews. Nótese que esta última tiene un rango diferente de valores en cada una de las gráficas







La configuración de la Gráfica No. 6 difiere de las otras dos, pero es evidente la conformación de los dos grupos. Las Gráficas No. 7 y 8 manifiestan una estructura muy similar y se acusan las diferencias entre los grupos, pero obsérvese que el eje vertical tiene un rango de valores distintos, con las variables canónicas exhibiendo una mayor diferenciación entre los grupos.

4.4 Conclusiones

Se fundamentó una metodología que sirviera para lograr la clasificación rural de los municipios que carecían de predialización catastral y los resultados de la aplicación de la metodología son coherentes en términos de los requerimientos exigidos por diferentes especialistas en la materia. Los resultados podrían considerarse robustos luego de haber pasado por el juicio crítico de otras técnicas que pueden orientarse a evaluar la efectividad de unos resultados. A través del documento se muestra también la gran versatilidad y utilidad de algunas de los métodos multivariados presentados, en los que está inmerso, incluso más que un estudio estadístico, un análisis numérico de la información.

Si bien es cierto se obtuvo una agrupación, no todos los 60 municipios fueron incluidos en el proceso. El camino a seguir sería aplicar análisis análogos a los aquí presentados con un conjunto menor de variables, combinando observaciones de los dos grupos con las que están sin clasificar y analizar las cercanías de los faltantes con los ya clasificados; esto daría un elemento técnico para sustentar la agrupación de las zonas rurales de los municipios aún por agrupar.

BIBLIOGRAFÍA.

Andrews, D.F. (1972). "Plots of high dimensional data", *Biometrics*, 28, 125-136.

Chatfield, C. and Collins, A.J. (1980). Introduction to Multivariate Analysis, Chapman and Hall, London.

Everitt, B.S. (1974). Cluster Analysis, Heinemann Educational Books, London.

Everitt, B.S. (1979). Unresolved problems in cluster analysis', *Biometrics*, 35, 169-182.

Everitt, B.S. (1993). Cluster Analysis, Third Edition, Edward Arnold, Great Britain.

Everitt, B.S. and Dunn, G. (1991), Applied Multivariate Data Analysis, Edward Arnold, London.

Hartigan, J.A. (1975). Clustering Algorithms, John Wiley and Sons, New York.

IICA–MINAGRICULTURA (1994). Proyecto Denso del Minifundio. Documento de trabajo No. 1 Metodología. Instituto Colombiano de Cooperación para la Agricultura – Ministerio de Agricultura. Santafé de Bogotá, D.C.

Krazanowski, W.J. (1990). Principles of Multivariate Analysis. A User's Perspective, Clarendon Press, Oxford.

Marriot, F.H.C. (1974). The Interpretation of Multiple Observations, Academic Press, London.

Mardia, K. V. (1974). "Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies", *Sankhya B*, 36, pp 115-128.

Mardia, K. V. (1975). «Assessment of Multinormality and the Robustness of Hotelling's T-squared Test», *Applied Statistics*, 1975, 24(2).

Mardia, K. V., Kent, J.T. & Bibby, J.M. (1979). Multivariate Analysis, Academic Press, New York

Milligan, G.W. (1980). "An examination of the effect of six types of error perturbation on fifteen clustering algorithms", *Psychometrika*, 45, No. 3, 325-342.

Milligan, G.W. (1981). "A Monte Carlo study of thirty internal criterion measures for cluster analysis", *Psychometrika*, 46, 187-99.

Milligan, G.W. and Cooper, M. (1985). "An examination of procedures for determining the number of clusters in a data set", *Psychometrika*, 50, 159-79.

Sarle, W.S (1983). Cubic Clustering Criterion. SAS Institute Inc., Cary, NC, USA.

SINTAP (1992). METODOLOGÍA PARA DETERMINAR LA UNIDAD AGRICOLA FAMILIAR. Sistema Nacional de Transferencia de Tecnología, Ministerio de Agricultura. Documento MA-STT-002-2-92. Santafé de Bogotá, D.C.

SAS Institute Inc. (1989). SAS/STAT User´s Guide, Version 6, Fourth Edition, Volumen 1, SAS Institute Inc., Cary, NC.

Spath, H. (1980), Cluster Analysis algorithms for data reductions and classifications of objects, John Wiley and Sons, West Sussex.

Stevens, J. (1992). Applied Multivariate Statistics for the Social Sciences Lawrence Erlbaum Ass,Inc., Hillsdale, New Jersey.