PERFIL SOCIECONÓMICO DE LOS MUNICIPIOS DE PUERTO RICO: APLICACIÓN DE TÉCNICAS ESTADÍSTICAS MULTIVARIADAS

Por

Yovani Correa Prieto

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

Maestría en Ciencias
en
Matemáticas (Estadísticas)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGUEZ
2011

Aprobada por:

Dámaris Santana Morant, Ph.D Miembro, Comité Graduado	Fecha
Edgardo Lorenzo González, Ph.D Miembro, Comité Graduado	Fecha
Julio Quintana Díaz, Ph.D Presidente, Comité Graduado	Fecha
Carlos Quiñones Padovani, Ph.D Representante de Estudios Graduados	Fecha
Omar Colon Reyes, Ph.D Director interino Departamento de Ciencias Matemáticas	Fecha

ABSTRACT

Classifying a particular society in similar socioeconomic groups allows to adoption of strategic conditions in those groups that are poor and at the same time to strengthen those in other groups.

In the present study we applied multivariate statistical techniques (principal component, factor and cluster analyses) to fourteen socio-economic variables available in Puerto Rico census data for the year 2000 and to four indexes calculated by us. The main objective was to develop a socioeconomic profile of the municipalities of Puerto Rico. By applying the method of principal components we reduced the number of variables to four, which summarize about 80% of the initial information from the data. Subsequently, these new variables were used as classification criteria from which emerged a division into five clusters for the 78 municipalities of Puerto Rico. Then a comparison was made of the similarities of the municipalities that belong to their respective clusters and also the differences between the resulting groups.

In general we found that some municipalities located at the central region of the Island belong to a particular cluster characterized by low economic development and low percentage of population obtaining at least a high school educational level. However, these municipalities show better social characteristics, such as a low proportion of homes with a woman as householder and the number of grandparents living with their grandchildrens and being responsible for them. It is also interesting that most of municipalities located at

the metropolitan area of Puerto Rico show better economic conditions and simultaneously have some problems related to unequal distribution of income among the population and higher rates of mortality. The remaining clusters also have particular characteristics that are discussed in detail in the findings chapter and presented in our conclusions.

RESUMEN

Clasificar determinada sociedad en grupos con características socioeconómicas similares permite adoptar estrategias para mejorar las condiciones en aquellos grupos en las que éstas sean precarias y al mismo tiempo fortalecerlas en los demás grupos.

En el presente trabajo el cual tiene un propósito estadístico y cuyos resultados pudiesen ser analizados posteriormente por sociólogos, economistas, psicólogos o cualquier otro experto interesado en ellos; se aplicaron técnicas estadísticas multivariadas (Análisis de Componentes principales, análisis factorial y análisis de conglomerados) a catorce variables socioeconómicas disponibles en los datos del censo realizado en Puerto Rico en el año 2000 y a cuatro índices que se calcularon como parte de la metodología. El objetivo principal fue descubrir un perfil socioeconómico de los municipios de Puerto Rico. A partir de la aplicación del método de componentes principales se redujo el número de variables a cuatro, las cuales resumen aproximadamente el 80% de la información inicial de los datos. Posteriormente estas nuevas variables se utilizaron como criterio de clasificación de donde surgió una división en cinco grupos para los 78 municipios de Puerto Rico. Se hizo entonces una comparación de las similitudes de los municipios que pertenecen a un mismo grupo y de las diferencias entre los conglomerados resultantes. De manera general se evidenció que algunos municipios que están ubicados en el centro de la isla y forman uno de los conglomerados, tienen un bajo desarrollo económico y algunos problemas en cuanto al porcentaje de población que logra un grado de escolaridad por lo menos de escuela

superior. Al mismo tiempo estos municipios presentan mejores condiciones en cuanto a ciertas características sociales como la baja proporción de mujeres cabeza de hogar o la cantidad de abuelos que son responsables de sus nietos. Los municipios que están ubicados en otro conglomerado, los cuales están ubicados básicamente en el área metropolitana de Puerto Rico, tienen mejores condiciones económicas y a su vez presentan algunas características negativas tales como mucha desigualdad en la distribución de los ingresos y la tasa de mortalidad elevada. Los demás conglomerados también tienen determinadas características que los identifican, las cuales se pueden encontrar en más detalle en las conclusiones.

A mi familia, especialmente a mi madre Rosa Prieto;

Por su amor y apoyo incondicional

AGRADECIMIENTOS

A Dios por su amor y sabiduría.

Al Dr. Julio C. Quintana, Presidente del Comité Graduado, por su acompañamiento durante el desarrollo de esta tesis.

A todas las personas que forman parte del Departamento de Ciencias Matemáticas del Recinto Universitario de Mayagüez, por brindarme la oportunidad de aprender y compartir.

A todos mis amigos y compañeros con los cuales compartí experiencias de vida.

Tabla de Contenido

Lista de Tablas

Tablas Página

Lista de Gráficas

Gráficas Página

1 INTRODUCCIÓN

1.1 Justificación

Los conjuntos de datos multivariados son aquellos en los cuales hemos medido más de una característica. Los métodos estadísticos multivariados son técnicas que nos permiten estudiar estas características simultáneamente para establecer relación entre ellas. La elección del método adecuado en ocasiones no es fácil y depende exclusivamente de las características que tengan los datos a estudiar. Los orígenes del Análisis Factorial suelen atribuirse a Spearman (1904) en su trabajo clásico sobre la inteligencia. Anteriormente Galton y K. Pearson (1901) presentaron algunos trabajos sobre el método de componentes principales.

Por su parte el análisis de conglomerados (clusters) tiene como propósito agrupar objetos tales que los elementos que pertenezcan a un mismo grupo, tengan características similares y que éstos grupos difieran significativamente. “El objetivo básico en el análisis de conglomerados es descubrir agrupaciones naturales de los objetos (o de variables). Al respecto, lo primero que hay que desarrollar una escala cuantitativa sobre la cual medir la asociación (similitud) entre objetos” [12]. La obtención de estos conglomerados (clúster) depende del criterio de distancia que se utilice para tal fin.

La clasificación de cualquier índole ha sido y seguirá siendo una actividad necesaria en la vida del ser humano ya que ésta permite apreciar características comunes de grupos de

elementos, lo cual facilita su estudio y tratamiento. En el campo de la actividad socioeconómica, el tener una clasificación en grupos de cierta entidad permite crear políticas conjuntas entre éstos, con el propósito de mejorar sus condiciones en este campo. Son escasos los estudios a nivel socioeconómico que se han realizado en Puerto Rico, en parte debido a la falta de información disponible para tal efecto. En este estudio se tomó información de los datos censales, una de las pocas fuentes con las que se cuentan para realizar este tipo de investigaciones. Se exploraron los métodos mencionados anteriormente, los que posteriormente se usaron para analizar datos tomados directamente para 15 variables seleccionadas correspondientes al censo realizado en Puerto Rico en el año 2000. Además de estas variables se calcularon cuatro nuevas (índices) que también se incluyeron en el estudio con el fin de aportar más información en estos aspectos. Los resultados del estudio podrían ser objeto de investigación por parte de expertos en sociología, economía y demografía, entre otras áreas del conocimiento que pudiesen establecer las causas de las condiciones en cada una de los grupos de municipios que se conformaron; y aun más importante, la toma de medidas por los entidades encargadas, encaminadas a mejorar estas condiciones en aquellos sectores donde se considere necesario.

1.2 Objetivos

- Analizar estadísticamente las relaciones entre las diferentes variables socioeconómicas seleccionadas para realizar la agrupación de los municipios de Puerto Rico.
- Determinar los factores que inciden en la clasificación de los municipios de Puerto Rico.
- Proponer una tipología de carácter socioeconómico de los municipios de Puerto Rico utilizando técnicas de análisis multivariado (Análisis de Componentes Principales, Análisis Factorial y Análisis de Conglomerados).
- Contrastar los conglomerados resultantes.

1.3 Resumen de Capítulos

El presente trabajo está distribuido en seis capítulos. En este capítulo se hace una introducción al tema de investigación y se plantean sus principales objetivos. En el segundo se presentan los conceptos teóricos en los cuales se basa el trabajo; en el Capítulo tres se presenta una descripción detallada de las variables que se consideraron, además de que se explica la forma en que se obtuvieron índices "Tasa de mortalidad", "índice de envejecimiento", "índice vital" e "Índice Gini", que posteriormente fueron incluidos en el estudio como nuevas variables; en el Capítulo cuatro se aplican las técnicas de análisis multivariado a los municipios en las 19 variables seleccionadas; se comparan y analizan resultados de distintas técnicas y finalmente se propone la clasificación de los municipios

de Puerto Rico en 5 grupos; en el Capítulo cinco se dan las conclusiones finales y en el capítulo seis se plantean las limitaciones que se tuvieron durante la elaboración de este trabajo y se proponen posibles trabajos futuros.

2 REVISIÓN DE LITERATURA

2.1 Análisis factorial

El análisis factorial (AF) es una técnica estadística de análisis multivariado cuyo objetivo principal es la reducción de un conjunto inicial de variables en un conjunto más pequeño de éstas sin pérdida de mucha información. Supongamos que para un determinado conjunto de *n* elementos hemos observado las variables $x_1, x_2, ..., x_p$, que generalmente se estandarizan cuando éstas tienen diferentes unidades de medidas para evitar que los resultados se vean afectados por aquellas variables que tienen valores atípicos. El análisis factorial representa cada una de estas variables iniciales como una combinación lineal de *m* factores (nuevas variables) con *m<p,* más un término relacionado con las perturbaciones no observadas. Se podría escribir el modelo como:

$$
\begin{aligned}
x_1 &= \lambda_{11} f_1 + \lambda_{12} f_2 + ... + \lambda_{1m} f_m + \varepsilon_1 \\
x_2 &= \lambda_{21} f_1 + \lambda_{22} f_2 + ... + \lambda_{2m} f_m + \varepsilon_2 \\
&\vdots \\
x_p &= \lambda_{p1} f_1 + \lambda_{p2} f_2 + ... + \lambda_{pm} f_m + \varepsilon_p
\end{aligned}
$$

Donde $f_1, f_2, ..., f_m$ son los factores que se supone tienen media cero, varianza uno y que son independientes entre sí. λ_{ij} es el factor de carga de la i-ésima variable en el j-ésimo factor. Los $\varepsilon_1, \varepsilon_2, ..., \varepsilon_p$ son los términos de error, cada uno de los cuales está relacionado con una variable original.

El modelo anterior se puede describir en forma matricial como:

$$X = \Lambda F + e$$

donde X es el vector $p \times 1$ de variables observadas, $\Lambda = [\lambda_{ij}]$ es una matriz $p \times m$ de cargas factoriales, F es un vector $m \times 1$ de factores y e es un vector $p \times 1$ de términos de error que se suponen no correlacionados entre sí, ni con los factores f; además con media igual a cero. Para la aplicación del método AF en la estimación de los factores no es necesario asumir alguna distribución particular para los datos, sin embargo para efectos de pruebas de hipótesis en el marco del AF se debe asumir la normalidad de éstos. Existen dos clases de modelos factoriales: el de factores ortogonales, cuando se asume que los factores no están relacionados y el de factores oblicuos, en el que los factores pueden estar relacionados entre sí [1].

"En el AF al dividir cada variable en dos partes (una debido a los factores comunes y otra debido a su propio y único factor) también se divide la varianza de los x_i en dos partes:

1. La comunalidad; que es la parte de la varianza que se debe a los factores comunes.
2. La especificidad o varianza específica, que es la parte de la varianza que se debe al factor único" [3].

Esto implica que las varianzas de las variables observadas pueden escribirse como:

$\sigma_i^2 = h_i^2 + \varphi_i^2 \; ; \; i = 1, 2, ..., p$ donde

$\sigma_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2 + \varphi_i^2 \; ; \; i = 1, 2, ..., p$ o de manera equivalente:

$h_i^2 = \sum_{j=1}^{m} \lambda_{ij}{}^2$ es la comunalidad para x_i y φ_i^2 es la especificidad para x_i.

$$\text{var}\,ianza = comunalidad + especificidad \quad [6]$$

2.1.1 **Modelo del factor principal o eje principal**

El método del factor principal es uno de las técnicas de análisis factorial. Este método es muy similar en varios aspectos al Análisis de Componentes Principales (El cual se expondrá en la siguiente sección), pero no actúa directamente en la matriz de varianzas- covarianzas sino en una estimación llamada en ocasiones matriz de covarianzas reducidas.

Cuando el factor de análisis se basa en la matriz de correlación de las variables manifiestas, dos métodos utilizados son:

Tome la comunalidad de una variable x_i como el cuadrado de los coeficientes de correlación múltiple de x_i con las otras variables observadas.

Tome la comunalidad de x_i como el mayor de los valores absolutos de los coeficientes de correlación entre el x_i y una de las otras variables [28]. Para más detalle sobre este método se pueden consultar las fuentes [6] y [28].

2.2 Análisis de componentes principales (ACP)

El análisis de componentes principales es otra técnica de análisis multivariado encaminada a reducir la dimensionalidad de los datos con la menor pérdida de información posible. Este método reduce un número de variables originales a un conjunto más pequeño de

combinaciones lineales de éstas que expliquen la mayor cantidad de variabilidad de los datos. Supóngase nuevamente que para un determinado conjunto de *n* elementos se han observado las variables $X_1, X_2, \ldots, X_p$, el objetivo es reducir el número de variables a k (*k*<*p*), donde estas representan los componentes principales sin la pérdida de mucha información contenida en las *p* variables originales.

De manera general se puede resumir el método afirmando que para hallar el primer componente principal, se encuentra el valor propio más grande de la matriz de varianzas-covarianzas y el vector propio normalizado asociado a este primer valor propio; los componentes de este vector son precisamente los coeficientes del primer componente principal. Igualmente para hallar el segundo componente principal, se hace el mismo procedimiento, con la diferencia que se toma el segundo valor propio mayor. Por tanto, lo que se debe hacer para hallar todos los componentes principales es hallar los valores propios y vectores propios normalizados asociados a la matriz de varianzas-covarianzas [1]y[4].

2.3 Validación de los modelos ACP , AF y Número de factores o componentes

2.3.1 Prueba de Bartlett e Índice Kaiser-Meyer-Olkin (KMO)

La Prueba de Significación Estadística, debido a Bartlett (1950) se puede utilizar de cualquiera de las dos maneras siguientes; en primer lugar, someter a prueba si la matriz de correlaciones es una matriz identidad. Si esta hipótesis nula de que la matriz de correlaciones es una matriz de identidad no puede ser rechazada, no sería sensato extraer

factores de la matriz. En segundo lugar, después de que cada factor se extrae, se puede analizar la matriz de correlaciones residuales para evaluar la información que se mantiene y en cada paso evaluar si esta matriz es una matriz identidad [2].

El KMO es una de las pruebas más utilizadas para analizar la viabilidad o no de la aplicación del método factorial. Se podría enunciar de la siguiente manera:

- Si KMO>0.7 existe alta intercorrelación entre las variables y por lo tanto es útil la aplicación del AF
- Si 0.5<KMO<0.7 el grado de intercorrelación es media y por lo tanto el AF es aplicable, pero sus resultados son menos útil que en el caso anterior
- Si KMO<0.5 la intercorrelacion entre las variables es baja y por lo tanto no es recomendable aplicar el AF [25].

2.3.2 Numero de factores o componentes

La elección del número de factores o componentes es un paso crucial en el proceso de reducción del número de variables pues con la elección de un número inadecuado puede ser que se pierda mucha información de los datos iniciales o por el contrario se utilice información que de cierto modo sobra en el análisis. Existen numerosas estrategias para la toma de esta decisión. En general, se pueden aplicar varios criterios, comparar sus resultados y elegir el número de componentes que más se ajuste a nuestro caso particular.

Valores propios

Un valor propio corresponde al número equivalente de variables a las que el factor representa. Por ejemplo, un factor asociado con un valor propio de 3.69 indica que el factor explica la varianza en los datos originales equivalente a 3.69 variables en promedio. Generalmente uno de los criterios más utilizados es mantener los factores que tengan valores propios mayores que uno. Otra regla utilizada frecuentemente es mantener los factores hasta el caso en el que un factor adicional represente menos de la varianza de una variable típica, es decir, menos de un valor propio. El diagrama de la varianza incremental o gráfica de sedimentación muestra el tamaño de los Valores propios y también puede ayudar en nuestra decisión. La idea de la prueba es que los factores a lo largo de la cola de la curva de variación en su mayoría representan el error aleatorio y por lo tanto hay que seleccionar la solución del número de factores justo antes de la nivelación de la curva [4].

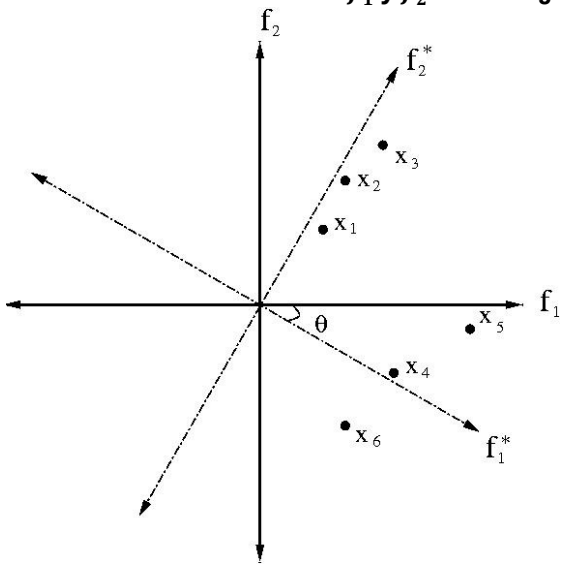
2.4 Rotación de factores

La rotación de los factores es un procedimiento cuyo propósito es redefinir los pesos factoriales de tal modo que se evite la mayor cantidad de pesos medios en las puntuaciones de cada variable en éstos y por el contrario tiendan a ser muy altas (cercanas a 1 o -1); y de esta forma ayudar a su interpretación [4] (Gráfica 2.1).

Dado que en el AF la solución inicial no necesariamente es única, es de suma importancia conseguir esta interpretación para que los resultados sean óptimos. La facilidad de la interpretación de los factores se da cuando se consigue lo que se conoce como estructura

simple (Thurstone). Esta estructura se caracteriza porque cada variable observada tiene en lo posible carga alta en un solo factor y cargas bajas en los factores restantes y además para cada factor hay en lo posible por lo menos una variable con carga alta en él [1].

Gráfica 2.1 Rotación de los factores f_1 y f_2 en un ángulo θ[15]



Hay que resaltar que en el proceso de rotación, no hay cambio en el número de factores ni en la cantidad de varianza explicada por éstos; lo que se hace es una redistribución de la varianza para tratar de facilitar su interpretación. La rotación no siempre arroja buenos resultados con respecto a tratar de equilibrar el total de varianza explicada por cada factor [4].

Existen varios tipos de rotación y uno de los más implementados es el método Varimax que corresponde a los métodos de rotación ortogonal en la que los ángulos entre los factores antes y después de dicha rotación se mantienen ortogonales.

2.5 Análisis de conglomerados

El objetivo principal del análisis de conglomerados (clúster análisis) es agrupar objetos de tal forma que los que pertenezcan a un mismo conglomerado sean lo más parecidos entre sí; y de que entre conglomerados haya diferencias significativas.

De acuerdo con [3], el análisis de conglomerados es una técnica de agrupación de personas u objetos en grupos desconocidos que se diferencia de otros métodos de clasificación, en que en el análisis de conglomerados el número y características de los grupos proceden de los datos y generalmente no se conocen antes de hacer al análisis. Para seleccionar los elementos que pertenecen a un mismo grupo se usan medidas de proximidad o similitud. Por ejemplo para agrupar elementos se usan medidas de distancia y para agrupar variables se utilizan medidas como la correlación entre éstas.

La aplicación del análisis de conglomerados se da en una gran diversidad de aéreas de investigación: en biología para clasificar seres de acuerdo a su especie, en psicología para clasificar personas de acuerdo a sus conductas, en medicina para asignar pacientes a diferentes categorías de diagnóstico dependiendo de los síntomas que presente, en geografía en estratificaciones territoriales [3].

El agrupamiento de objetos en una población se hace teniendo en cuenta una medida de distancia entre dichas observaciones. A continuación se exponen algunas las medidas de distancia más utilizadas.

Medidas de distancia (o disimilitud)

Distancia de Minkowsky

$$d_m(x,y) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{\frac{1}{m}}$$

Casos particulares:

a. Distancia Manhattan

$$d_1(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

b. Distancia Euclídea

$$d_2 = \sum_{i=1}^{p} (x_i - y_i)^2$$

Algunas medidas de similitud [18]**:**

a. Medida de correlación

$$r(x,y) = \frac{\sum_{i=1}^{M}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{M}(x_i - \overline{x})^2 \sum_{i=1}^{M}(y_i - \overline{y})^2}} = \frac{(x - \overline{x})'(y - \overline{y})}{\|x - \overline{x}\|\|y - \overline{y}\|}$$

Nota: $1 - r(x,y)$ puede ser considerado como una medida de disimilitud.

b. Medida de Tanimoto

$$S_T(x,y) = \frac{x'y}{\|x\|^2 + \|y\|^2 - x'y}$$

Los métodos de agrupación se dividen básicamente en jerárquicos y no jerárquicos (o de partición); los métodos jerárquicos a su vez pueden ser aglomérativos (AGNES) o divisivos (DIANA).

2.5.1 Método jerárquico aglomerativo

En el método aglomerativo se empieza con n agrupaciones, es decir cada observación constituye su propio grupo, luego se agrupan los dos conglomerados más cercanos y así sucesivamente, de tal forma que al final todas las observaciones pertenecen al mismo grupo. Si X es un conjunto de n observaciones y $d(C_i, C_j)$ es una función que mide la proximidad entre cualesquiera dos observaciones C_i y C_j ; entonces en términos generales, el procedimiento puede ser descrito como sigue, según González, en [11].

1. Primero cada elemento de X forma un pequeño conglomerado por sí mismo.
2. En el primer paso los dos objetos más cercanos o similares se unen usando una medida de disimilitud $d(C_i, C_j)$; esto es, se encuentra el valor más pequeño en la matriz de disimilitudes y se unen los correspondientes objetos.
3. En el segundo paso se tienen $N - 1$ conglomerados. Ahora se deben fusionar los grupos más cercanos utilizando algún método de enlace o vinculación.
4. En el paso t hay $N - (t - 1)$ conglomerados y se quieren unir los grupos más cercanos como en el paso anterior.
5. Se repite el procedimiento hasta que todos los elementos del vector estén en el mismo grupo.

Los métodos de enlace o vinculación mencionados en el paso 3 son medidas cuantitativas para unir los dos grupos más similares en el algoritmo de agrupación aglomerativo; entre los más populares están: el método de promedios aritméticos par-grupo (UPGMA), el método de vinculación completa de conglomerados (CLINK), el método de agrupación diferencia mínima de Ward y el método de promedios aritméticos ponderados par-grupo (WPGMA).

Una representación gráfica muy útil para mostrar los pasos efectuados en el proceso del análisis de conglomerados jerárquico es el dendograma, el cual está inmerso en la mayoría de programas estadísticos.

Ejemplo 1

TABLA 2.1 Datos del ejemplo 1

Obs.	$Peso(kg)$	$Altura(cm)$
O_1	17	93
O_2	50	155
O_3	98	181
O_4	83	176
O_5	70	169
O_6	63	152

Considérese el siguiente ejemplo hipotético para observar cómo se representa un método jerárquico mediante el dendograma. Los datos de la Tabla 2.1 corresponden a la estatura (en cm) y al peso (en kg) de seis personas. Como primera medida se calculan las distancias entre los objetos, en este caso utilizando la distancia Euclidiana; las cuales se muestran en

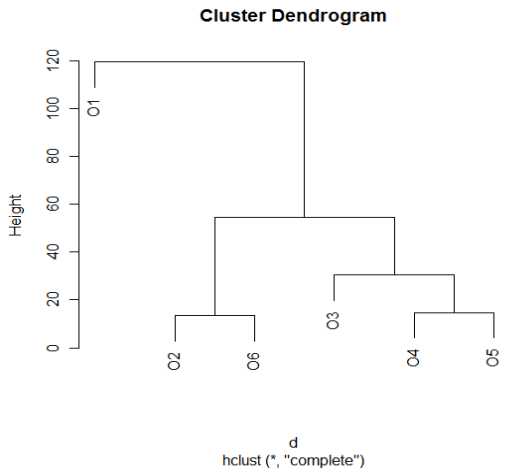
la Tabla 2.2. La gráfica 2.2 deja ver la manera en que se van agrupando los elementos de acuerdo a su proximidad hasta que al final las seis observaciones forman un solo conglomerado.

TABLA 2.2 Distancia Euclidea para las seis personas de ejemplo 1

$Obs.$	O_1	O_2	O_3	O_4	O_5	O_6
O_1	0					
O_2	70.235	0				
O_3	119.603	54.589	0			
O_4	106.042	39.115	15.811	0		
O_5	92.655	24.413	30.463	14.764	0	
O_6	74.813	13.341	45.453	31.241	18.385	0

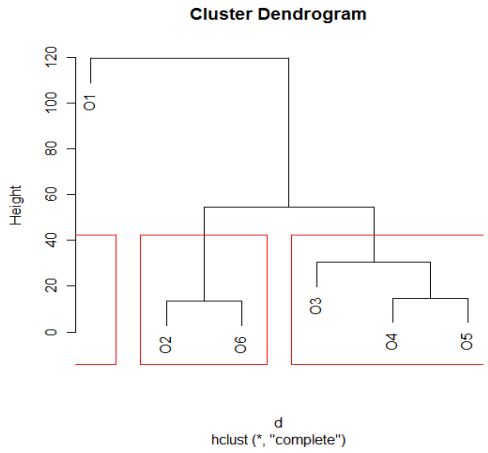
El dendograma muestra una disposición de los datos de tal forma que si se hace un corte horizontal a cierto nivel establecido se obtiene una agrupación de dichos datos.

Gráfica 2.2 Dendograma para el ejemplo 1



Si se supone que el número correcto de grupos que presentan estos datos es 3, se obtendría lo que se muestra en la gráfica 2.3, donde los conglomerados para este caso estarían formados de la siguiente manera: $\{O_1\}$, $\{O_2, O_6\}$ y $\{O_3, O_4, O_5\}$

Gráfica 2.3 Partición en tres grupos de ejemplo 1



2.5.2 Método jerárquico divisivo

Este método de agrupamiento jerárquico construye la jerarquía en orden inverso al método aglomerativo. En términos generales el algoritmo puede ser descrito como sigue, según González en [11]:

1. Inicialmente todos los elementos en X están en un mismo grupo.
2. En el primer paso se divide el conjunto de datos en dos grupos, para esto se busca el objeto para el cual la medida de disimilitud promedio es más grande que todos los demás. Este objeto de mayor disimilitud inicia un nuevo grupo llamado disidente.
3. Para cada objeto en el grupo más grande, se calcula el promedio de disimilitud con los objetos restantes y se compara dio de disimilitud de los objetos del grupo

disidente. El objeto en el grupo más grande con diferencia mayor en los cambios de lados, se mueve al grupo disidente. Repita los cálculos hasta que todas las diferencias sean negativas.

4. En el siguiente paso, dividir el grupo mayor, es decir el grupo de mayor diámetro. El procedimiento es el mismo que en el paso anterior.
5. En los pasos siguientes, se divide el mayor grupo siguiendo el procedimiento anterior.
6. El proceso continúa hasta que cada objeto forme un conglomerado con un solo elemento.

2.5.3 Métodos de partición (no jerárquicos)

Algoritmo de k-medias

El objetivo de este método es dividir un conjunto de datos en un número de conglomerados asignado inicialmente.

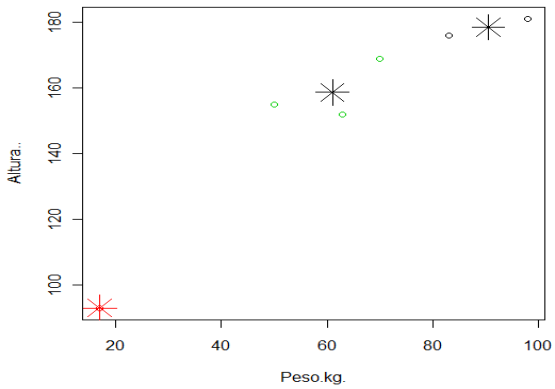
En términos generales el algoritmo procede de la siguiente manera [3]:

1. Se dividen los datos en K grupos iniciales. Los miembros de estos grupos puede ser especificados por el usuario o pueden ser seleccionado por un programa. de acuerdo con un procedimiento arbitrario.
2. Se Calculan las medias o centroides de cada conglomerado

3. Para un caso determinado, se calcula su distancia a cada centroide. Si el caso es más cercano al centro de gravedad de su propio grupo, se deja en ese grupo, de lo contrario, se vuelve a asignar al grupo cuyo centro de gravedad está más cerca de él.
4. Repita el paso 3 para cada caso
5. Repita los pasos 2, 3 y 4 hasta que no se vuelvan a asignar casos.

Es conveniente que se tomen diferentes asignaciones para los valores iniciales, dado que en muchas ocasiones, los resultados dependen de la elección y del orden [6].

Gráfica 2.4 Centroides aplicando K-medias



La gráfica 2.4 muestra la representación de los conglomerados aplicando el algoritmo de k- medias de la librería {clúster} del paquete estadístico R a los datos del Ejemplo 1, en esta gráfica se puede apreciar los tres conglomerados finales con sus respectivos centroides; después de que el método ha convergido. Cuyo resumen se muestra a continuación en la siguiente tabla:

TABLA 2.3 Clasificación de ejemplo 1 aplicando algoritmo k-medias

	Peso	Altura	Conglomerado
O1	17	93	3
O2	50	155	2
O3	98	181	1
O4	83	176	1
O5	70	169	2
O6	63	152	2

lo cual indica que los grupos aplicando este algoritmo estarían formados de la siguiente manera: $\{O_1\}$, $\{O_2, O_5, O_6\}$ y $\{O_3, O_4\}$.

Particionamiento alrededor de Medoides (PAM)

PAM funciona de manera análoga al algoritmo de k-medias. También se deben especificar el número de conglomerados de antemano y el objetivo es seleccionar una partición $\{C_1, C_2, \cdots C_k\}$ de la población y un conjunto de centros de grupos $\{c_1, c_2, \cdots c_k\}$ que minimice la expresión:

$$\sum_{i}^{k} \sum_{j \in C_i} d(x_j, c_i)$$

Donde d es alguna medida de distancia. Este algoritmo minimiza la suma de disimilitudes en cambio de la disimilitud promedio y es mucho más estable que el algoritmo k-medias; por otra parte PAM trabaja mucho más rápido en problemas de grandes dimensiones [13].

Además, "PAM es más robusto al ruido y los valores atípicos cuando se compara con el de k-medias" [14].

Para los datos del Ejemplo 1, al aplicar el algoritmo PAM se obtuvo los resultados de la Tabla 2.4:

TABLA 2.4 Clasificación de datos del ejemplo 1 con algoritmo PAM

	Peso	Altura	Conglomerado
O1	17	93	1
O2	50	155	2
O3	98	181	3
O4	83	176	3
O5	70	169	3
O6	63	152	2

Esto nos muestra la siguiente agrupación: $\{O_1\}$, $\{O_3, O_4, O_5\}$ y $\{O_2, O_6\}$.

Nótese que esta agrupación coincide de la obtenida con el método jerárquico y difiere de la obtenida por el método de las k-medias.

2.6 Número óptimo de conglomerados

La elección del número de grupos para asociar los elementos en estudio, generalmente no es un trabajo fácil y depende en gran medida del conocimiento que el investigador tenga sobre las características de éstos y la correcta aplicación de algunas técnicas de validación disponibles en la bibliografía.

Un criterio bastante utilizado compara la suma de cuadrados (SCDG) dentro de los grupos para todas las variables para G grupos con la de G+1 grupo.

$$F = \frac{SCDG(G) - SCDG(G+1)}{SCDG(G+1)/n-G-1} = \frac{(n-G)MS(G) - (n-G-1)MS(G+1)}{MS(G+1)} \qquad 2.1$$

Esto cociente suele compararse con una distribución F, que en ocasiones su uso no está justificado de acuerdo a los requerimientos de dicha distribución. Al respecto Hartigan propuso una regla empírica que suele dar buenos resultados y consiste en seguir dividiendo el conjunto de datos si la expresión 2.1 es mayor que 10. En la segunda igualdad $MS(G)$ representa la fila de totales del cuadrado medio de error de la tabla de Anova que es equivalente a $\frac{SCDG(G)}{(n-G)}$. Ms representa los cuadrados medios del conglomerado o del error según corresponda de la Tabla de Anova.

Otro criterio que suele utilizarse fue propuesto por Calinski y Harabasz (1974). Y consiste en seleccionar el número de grupos que maximicen la expresión:

$$CH = max \frac{tr(B)/G-1}{tr(W)/n-G} \qquad 2.2$$

donde W representa la variabilidad dentro de cada grupo y B la variabilidad entre los grupos. El numerador de esta expresión es la suma de los términos clúster Ms para todas las variables y el denominador la suma de la columna de error Ms [6].

Otros criterios de validación de conglomerados

Otros de los criterios bastante utilizados para elegir el número óptimo de conglomerados son las medidas internas de validación y las medidas de estabilidad. Entre las más utilizadas están:

Medidas internas de validación

Estos índices utilizan la información inherente al conjunto de datos; Entre las más usadas están:

- Conectividad
- Ancho de silueta
- Índice Dunn

Medidas de estabilidad

Estas medidas han presentado resultados eficaces en datos que están altamente correlacionados. Entre las más comunes están:

- Proporción promedio de no superposición (APN)
- Distancia promedio (AD)
- Distancia media entre los promedios (ADM)
- Figura de merito (FOM)

Al respecto Brock, V. Pihur, S. Datta, y S. Datta, en su artículo publicado en *Journal of statistical software,* hacen un detallado estudio de estas medidas y su implementación con el paquete estadístico R-project. Para más detalles ver [8] y [9].

Para la realización de los cálculos estadísticos necesarios en la investigación se utilizaron los paquetes R-project 2.12.2, SPSS 17.0 y Minitab 15.0.

3 METODOLOGÍA

En este capítulo se explica detalladamente la forma en que se realizó el estudio, explicando cada una de las variables que se consideraron y utilizando las expresiones adecuadas para el cálculo de los índices que también se incluyeron como variables.

3.1 Descripción de las variables

Para el estudio inicial se escogieron las siguientes 19 variables; las cuales se consideraron básicamente por dos razones: la primera porque en algunos estudios anteriores otros autores (Rasic [20], Poza [26]) incluyeron este tipo de variables en sus investigaciones y además fueron las variables de tipo socioeconómico para las que se encontró información detallada por municipio. Para cada una de ellas se recopiló la información en los 78 municipios que actualmente conforman Puerto Rico. Las variables se enumeran a continuación:

V_1 Por ciento de población graduada de escuela superior o más

V_2 Por ciento de población mayor de 65 años

V_3 Por ciento de población empleada en ventas y oficinista

V_4 Por ciento de población empleada como gerencial o profesional

V_5 Ingreso per cápita

V_6 Por ciento de familias sin al menos un vehículo disponible

V_7 Por ciento de población graduada de bachillerato o más

V_8 Por ciento de población bajo el nivel de pobreza

V_9 Por ciento de población empleada.

V_{10} Por ciento de población mayor de 16 años desempleada

V_{11} Densidad de población

V_{12} Tasa bruta de mortalidad por cada mil habitantes

V_{13} Por ciento de población empleada en agricultura, pesca y silvicultura

V_{14} Índice de envejecimiento

V_{15} Por ciento de hogares con mujer cabeza de hogar sin esposo presente

V_{16} Por ciento de hogares con abuelos que viven con sus nietos y son responsables de ellos

V_{17} Por ciento de población casada no separada

V_{18} Índice vital

V_{19} Índice Gini

A continuación se muestra como se obtuvieron los valores de las variables (índices) V12, V14, V18 y V19 . Para las restantes 15 variables los valores se tomaron directamente de los datos de la página de la Oficina del Censo [24], correspondientes al año 2000.

Tasa bruta de mortalidad (V12):

Esta tasa indica cuantas defunciones, por cada 1000 habitantes se produjeron durante el año natural en una cierta área. Es decir $V12 = \frac{F}{P} \times 1000$ donde F representa la cantidad de fallecimientos en el periodo de tiempo y P la población total [21].

Índice de envejecimiento (V14): Este índice mide el número de adultos mayores de 65 años por cada 100 niños y jóvenes (menores de 15 años); es decir: $V14 = \frac{p>65}{p<15} \times 100$

Índice vital (V18):

Se calculó como la razón entre el total de nacimientos y el total de defunciones durante el año 2000 en cada municipio.

Índice Gini (V19):

Este indicador es uno de los más utilizados para medir la distribución de los ingresos en determinada sociedad.

"Para una distribución de N rentas (X) con los valores ordenados en sentido creciente, esto es,

$x_1 \leq x_2 \leq \cdots \leq x_N$ el índice de Gini está dado por:

$$IG = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$

Donde p_i nos indica la proporcion que suponen respecto al total los i rentistas que perciben rentas más pequeñas, esto es, cuya renta es menor o igual que p_i. Y q_i nos indica la proporción sobre el valor total de la renta acumulada por los i primeros individuos.

El Índice de Gini es uno de los indicadores aplicados habitualmente en los estudios de desigualdad realizados por distintos organismos internacionales tales como Banco Mundial, Naciones Unidas o Social Watch"[22].

3.2 Procedimiento

Como primera parte del proceso se hizo un estudio exploratorio de los datos donde se analizó la existencia de datos atípicos y su posible influencia en los resultados del modelo elegido para reducir el número de variables.

Como segunda parte del proceso se validaron los métodos de análisis factorial y el análisis de componentes principales. Se compararon los resultados obtenidos y se escogió uno de ellos para reducir las 19 variables iniciales teniendo en cuenta la cantidad de varianza explicada por cada modelo en total y el valor de las comunalidades, la cual representa la cantidad de la varianza explicada para cada una de las variables consideradas en el estudio.

Posteriormente se aplicó el modelo de análisis de componentes principales, pues éste mostró mejores resultados para este caso particular. Para una mejor interpretación de los componentes se aplicó el método de rotación Varimax, pues algunas variables presentaron cargas similares en más de un componente. Se rotularon estas nuevas variables (componentes) teniendo en cuenta las variables iniciales que conformaban cada uno de ellos por su alta correlación.

Se tomaron estas cuatro nuevas variables como criterio de clasificación para aplicar algunos métodos de validación de conglomerados tendientes a encontrar el número óptimo de grupos en que se dividieron los 78 municipios.

Por los resultados obtenidos al aplicar los procedimientos se decidió clasificar en 5 conglomerados estos municipios, aplicando posteriormente el método de agrupación alrededor de los medoides PAM con el cual se establecen los cinco grupos de municipios.

Se hizo una descripción de cada conglomerado estableciendo intervalos de confianza del 95% alrededor del promedio de cada conglomerado para cada uno de estos grupos y se analizaron las variables más significativas en cada conglomerado. Para facilitar la comparación de las características de los 5 grupos se presenta gráficamente cada una de las 19 variables. Finalmente se obtienen las conclusiones, se plantean algunas de las limitaciones que se presentaron durante la investigación y se plantean algunos posibles trabajos futuros al respecto.

4 ANÁLISIS DE RESULTADOS

4.1 Análisis de datos

Antes de aplicar cualquier método multivariado es necesario hacer un estudio inicial de los datos para detectar posibles valores atípicos que puedan distorsionar los resultados obtenidos.

Una herramienta a la hora de analizar la presencia de datos atípicos es la distancia de Mahalanobis la cual sigue aproximadamente una distribución ji-cuadrado para muestras relativamente grandes, con grados de libertad igual al número de variables presentes en el estudio (Johnson y Wichern). Por lo general una observación se considera una observación como posible dato atípico multivariante si su distancia al centroide de los datos es mayor que el valor de la distribución ji-cuadrado con un valor de significancia especificado y que en la mayoría de los casos se toma como $\alpha = 0.001$. Para este propósito podemos aplicar una regresión tomando como variable dependiente una variable que no sea de interés y como variables independientes las variables consideradas de interés en el estudio [1]. Con este propósito en nuestro caso se aplicó una regresión tomando como variable dependiente el número del municipio al organizarlos alfabéticamente (Tabla 4.1) y las 19 variables consideradas anteriormente como explicativas. Los resultados al utilizar el programa SPSS con las respectivas distancias de Mahalanobis para cada municipio se muestra en la Tabla 4.1

El valor de la distribución ji-cuadrado con 19 grados de libertad a un nivel α = 0.001 es 43.82; lo que indicaría que los municipios de Culebra, Guaynabo y San Juan podrían ser posibles datos atípicos. Para corroborar si estos datos son influyentes se analizó la distancia Cooks (DC) cuyos resultados también se muestran en la Tabla 4.1. Generalmente una observación con una DC>1 se considera potencialmente influyente, si una observación tiene un DC<1 no requiere ninguna discusión y una observación con una DC<0.5 merece algo de cuidado. Más específicamente cualquier observación con DC>$F_{0.5,p,n-p}$ se considera un valor influencial [5]. En el análisis, el valor de F que se obtuvo fue $F_{0.5,19,60}$ = 0.976166 por lo tanto no hay ninguna observación que pudiese ser influyente en el estudio; San Juan y San Sebastián con DC de 0.136 y 0.145 respectivamente merecen un poco de atención pues sus valores DC son mayores de 0.1, con valores muy cercanos a éste. Para saber si estos municipios influyen de manera drástica en los resultados del estudio, se consideró el modelo de ACP (dado que éste será el que se aplique después de hacer un análisis que se expondrá más adelante) excluyéndolos (estos resultados se muestran en las Tablas A1-A4 del apéndice) y se obtuvieron resultados similares; por lo que en el modelo se incluyeron todos los municipios.

TABLA 4.1 Distancias de Mahalanobis(DM) y Cooks(D_cook) para los 78 municipios

	Municipio	*DM*	*D_cook*		*Municipio*	*DM*	*D_cook*
M1	Adjuntas	24.257	0.050	M40	Juncos	4.815	0.001
M2	Aguada	19.102	0.075	M41	Lajas	12.319	0.000
M3	Aguadilla	6.938	0.015	M42	Lares	24.199	0.004
M4	Aguas Buenas	15.007	0.018	M43	Las Marías	33.856	0.000
M5	Aibonito	9.979	0.006	M44	Las Piedras	14.963	0.000
M6	Añasco	16.862	0.035	M45	Loíza	35.255	0.002
M7	Arecibo	7.765	0.010	M46	Luquillo	10.317	0.001
M8	Arroyo	27.894	0.028	M47	Manatí	9.113	0.003
M9	Barceloneta	32.896	0.085	M48	Maricao	35.486	0.000
M10	Barranquitas	10.624	0.016	M49	Maunabo	22.338	0.009
M11	Bayamón	24.145	0.008	M50	Mayagüez	26.173	0.000
M12	Cabo Rojo	13.282	0.021	M51	Moca	15.839	0.000
M13	Caguas	10.164	0.010	M52	Morovis	11.081	0.001
M14	Camuy	14.585	0.005	M53	Naguabo	9.935	0.003
M15	Canovanas	10.030	0.022	M54	Naranjito	12.723	0.006
M16	Carolina	20.383	0.009	M55	Orocovis	22.482	0.020
M17	Cataño	32.055	0.019	M56	Patillas	13.279	0.016
M18	Cayey	6.443	0.003	M57	Peñuelas	11.543	0.000
M19	Ceiba	36.592	0.000	M58	Ponce	12.578	0.003
M20	Ciales	13.424	0.005	M59	Quebradillas	12.454	0.012
M21	Cidra	14.002	0.003	M60	Rincón	24.314	0.019
M22	Coamo	6.198	0.003	M61	Río Grande	11.656	0.018
M23	Comerío	25.714	0.021	M62	Sabana Grande	5.460	0.006
M24	Corozal	17.324	0.007	M63	Salinas	7.503	0.004
M25	Culebra	62.213	0.057	M64	San Germán	12.893	0.028
M26	Dorado	26.823	0.056	M65	San Juan	45.294	0.137
M27	Fajardo	13.846	0.001	M66	San Lorenzo	14.653	0.025
M28	Florida	24.872	0.015	M67	San Sebastián	38.710	0.145
M29	Guánica	18.836	0.015	M68	Santa Isabel	39.944	0.044
M30	Guayama	11.458	0.000	M69	Toa Alta	17.433	0.018
M31	Guayanilla	13.015	0.001	M70	Toa Baja	16.013	0.038
M32	Guaynabo	56.162	0.180	M71	Trujillo Alto	14.654	0.020
M33	Gurabo	17.743	0.003	M72	Utuado	14.737	0.035
M34	Hatillo	19.000	0.001	M73	Vega Alta	13.953	0.024
M35	Hormigueros	28.139	0.000	M74	Vega Baja	4.310	0.004
M36	Humacao	7.059	0.001	M75	Vieques	40.893	0.001
M37	Isabela	6.152	0.000	M76	Villalba	15.593	0.028
M38	Jayuya	16.961	0.002	M77	Yabucoa	11.167	0.034
M39	Juana Díaz	16.134	0.002	M78	Yauco	8.991	0.021

4.2 Análisis factorial y análisis de componentes principales

En esta parte del estudio se aplicaron dos de los métodos más utilizados para reducir la dimensionalidad de los datos, el ACP y el AF en el cual se usó el algoritmo del factor principal o eje principal.

4.2.1 Validación de los modelos

Para verificar la validez de los modelos: como primera medida se analizó la matriz de correlación de las 19 variables consideradas en el estudio y se pudo apreciar la alta correlación entre varias de ellas: por ejemplo la variable "por ciento de población mayor de 65 años (V2)" está altamente correlacionada con las variables "Tasa bruta de mortalidad (V12)", "Índice de envejecimiento (V14)" e "Índice Vital (V18)" (Tabla A19 del Apéndice) lo que sugiere su agrupación.

Se usaron además algunas pruebas como el Índice de Kaiser-Meyer-Olkin (KMO) y la prueba de esfericidad de Bartlett mencionados en el capítulo anterior.

Las Tablas 4.2 y 4.3 muestran los valores del KMO obtenidos con las variables iniciales para los modelos factorial y de componentes principales respectivamente. El valor del KMO en ambos casos es 0.764, lo cual implica que es viable la aplicación de los métodos. Igualmente la prueba de Bartlett con un valor de 1662.37 tanto para el AF como para el ACP y un valor de p muy próximo a cero permite rechazar la hipótesis nula de que la matriz de correlaciones es una matriz identidad. Al rechazar esta hipótesis se concluyó que las variables están correlacionadas. Si no se llegara a rechazar la hipótesis nula se concluiría que el modelo es inadecuado.

TABLA 4.2 Validación para ACP

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.764
Bartlett's Test of Sphericity	Approx. Chi-Square	1662.387
	Df	171
	Sig.	<.0001

TABLA 4.3 Validación para AF

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.764
Bartlett's Test of Sphericity	Approx. Chi-Square	1662.387
	Df	171
	Sig.	<.0001

Para elegir cuál de los modelos se usó finalmente para reducir el número de variables se compararon el total de la varianza explicada y las comunalidades (porcentaje de cada variable explicada por el modelo) para cada variable en los dos casos. El total de la varianza explicada para el método del factor principal es 74.581% mientras que para el ACP es 79.476% (Los detalles se muestran en las Tablas 4.4 y 4.5).

TABLA 4.4 Varianza total explicada por ACP

Component	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.931	41.744	41.744	4.852	25.535	25.535
2	3.813	20.066	61.809	4.105	21.605	47.139
3	2.196	11.555	73.365	3.544	18.655	65.795
4	1.161	6.111	79.476	2.599	13.681	79.476

Extraction Method: Principal Component Analysis.

TABLA 4.5 Varianza total explicada por AF

Factor	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.712	40.589	40.589	5.003	26.332	26.332
2	3.610	19.002	59.591	3.381	17.796	44.128
3	1.994	10.497	70.088	3.378	17.780	61.908
4	.854	4.493	74.581	2.408	12.673	74.581

Extraction Method: Principal Axis Factoring.

1. En las Tablas 4.6 y 4.7 se observa que en 15 de las 19 variables el ACP explica una mayor variabilidad (ésta está representada por la comunalidad) que el AF. Además en la variable “Abuelos responsables de sus nietos” el AF sólo reproduce un 26.6% de su variabilidad, contra un 52% que explica el ACP para la misma variable.

TABLA 4.6 Varianza explicada por el modelo AF para cada variable

Variables	Comunalidades	Variables	Comunalidades
Grad_esc_superior	0.779	Mayor de 16 desempleada	0.522
Pob_mayor de 65	0.842	Densidad población	0.658
Trab_Ventas y oficinista	0.659	Tasa bruta mortalidad	0.765
Trab_Geren y Profesionales	0.753	Trab_agric_pesc_silv.	0.564
Íngreso per cápita	0.921	Indice_envejecimiento	0.891
Sin vehículo disponible	0.51	Mujer_cabeza_hogar	0.966
Graduado_Bachillerato o más	0.946	Abue_resp_ de sus nietos	0.266
Pob._bajo_nivel Pobreza	0.946	casados_no separados	0.89
Poblacion_Empleada	0.79	Indice vital	0.87
		Indice Gini	0.633

TABLA 4.7 Varianza explicada por el modelo ACP para cada variable

Variables	Comunalidades	Variables	Comunalidades
Grad_esc_superior	0.803	Mayor de 16 desempleada	0.591
Pob_mayor de 65	0.876	Densidad población	0.736
Trab_Ventas y oficinista	0.704	Tasa bruta mortalidad	0.831
Trab_Geren y Profesionales	0.808	Trab_agric_pezc_silv	0.740
Ingreso per cápita	0.910	Indice_envejecimiento	0.904
Sin vehículo disponible	0.638	Mujer_cabeza_hogar	0.904
Graduado_Bachillerato o más	0.921	Abue_resp_ de sus nietos	0.520
Población_bajo_nivel Pobreza	0.914	casados_no separados	0.892
Población_Empleada	0.811	Indice vital	0.894
		Indice gini	0.704

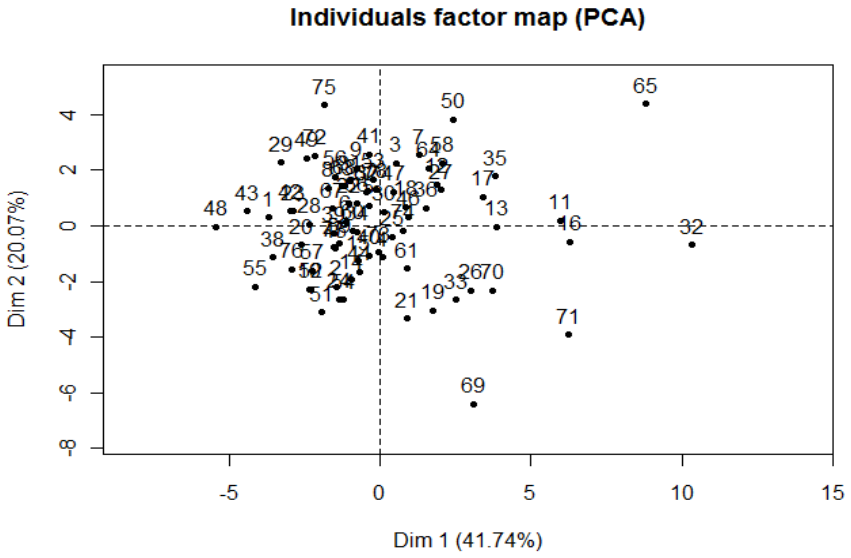
Aunque con cualquiera de los dos métodos se obtienen resultados adecuados en lo que respecta a la variabilidad explicada, según lo discutido anteriormente se eligió y se trabajó en base a los resultados obtenidos con el ACP.

La Tabla 4.7 muestra que las variables *población graduada de bachillerato, población bajo nivel de pobreza, ingreso per cápita, índice de envejecimiento y mujer cabeza de hogar* están muy bien explicadas por el modelo ya que éste reproduce más del 90% de la variabilidad de cada una de ellas; en contraste con esto, las variables *abuelos responsables de sus nietos y población mayor de 16 desempleada* son las variables cuya varianza explicada por el modelo es menor de 60%.

Para visualizar mejor los resultados de los componentes principales se ha utilizado la librería FactoMineR del paquete estadístico R expuesto por (Husson, Le y Pages) en [7]. Las Gráfica 4.1 muestra el mejor plano para representar los municipios en términos de su

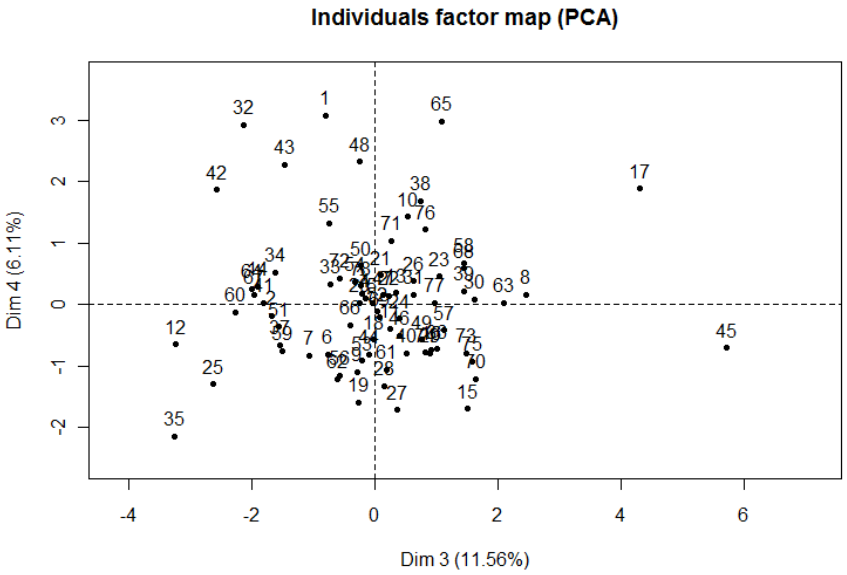
dispersión. La varianza de los datos representada en este plano corresponde a la suma de los Valores propios correspondientes para estos dos componentes; es decir: 41.74% + 20.07% = 61.81% del total de la varianza.

Gráfica 4.1 Diagrama de dispersión de los municipios en los componentes 1 y 2



Se puede observar el comportamiento extremo de algunos objetos (municipios) en el plano formado por los dos primeros componentes; por ejemplo el punto 65 que corresponde al municipio de San Juan, el cual parece ser un municipio inusual ya que sus resultados son extremos tanto para el primero como para el segundo componente. El punto 32 (Guaynabo) y el 48 (Maricao) tienen rendimientos diferentes con respecto al segundo componente pues están opuestos por el eje principal de variabilidad. Igualmente se puede representar en un plano los componentes 3 y 4 (Gráfica 4.2) y obsérvese por ejemplo que los municipios identificados con los puntos 45 (Loíza) y 17 (Cataño) tienen un comportamiento inusual con respecto a estos componentes.

Gráfica 4.2 Diagrama de dispersión de los municipios en los componentes 3 y 4



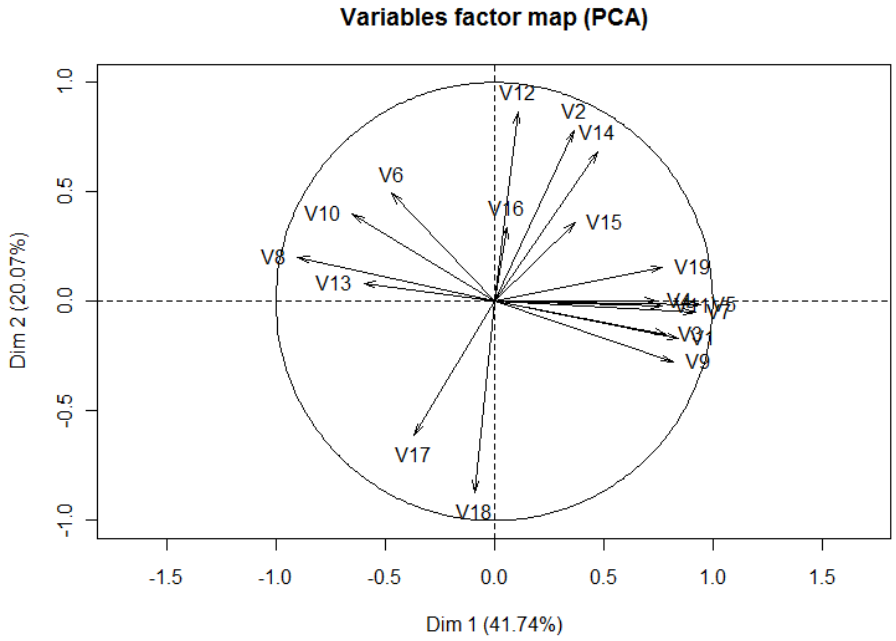
En general algunos de los municipios parecen estar muy alejados del origen de los componentes. Para tener una idea de cuáles son estos municipios, se puede calcular la distancia desde el centro de la nube hasta cada uno de estos datos (utilizando el comando round(res.pcainddist,2) de la librería FactoMineR de R) expuesto en [7], los resultados se muestran en la Tabla 4.8 donde confirmamos que municipios como Guaynabo (M32) y San Juan (M65) efectivamente son los municipios más alejados.

TABLA 4.8 Distancia de cada municipio al origen

M5	M40	M47	M74	M18	M46	M22	M78	M62	M66	M36	M61	M37
1.56	1.68	1.77	1.78	1.81	1.86	1.9	2.07	2.17	2.23	2.38	2.4	2.41
M77	M30	M4	M6	M34	M44	M31	M59	M73	M53	M3	M39	M15
2.42	2.44	2.46	2.49	2.5	2.51	2.55	2.59	2.6	2.64	2.72	2.83	2.94
M56	M63	M57	M14	M20	M7	M27	M60	M54	M67	M28	M52	M64
2.95	3.05	3.21	3.25	3.28	3.32	3.35	3.37	3.51	3.51	3.59	3.6	3.6
M41	M9	M72	M2	M21	M58	M10	M49	M24	M23	M33	M76	M13
3.71	3.74	3.75	3.76	3.8	3.85	3.86	3.94	4	4.01	4.05	4.1	4.11
M8	M68	M12	M19	M51	M38	M29	M26	M50	M42	M1	M55	M70
4.18	4.2	4.29	4.38	4.41	4.58	4.59	4.79	5.04	5.07	5.12	5.25	5.28
M43	M25	M35	M45	M75	M11	M17	M16	M48	M69	M71	M65	M32
5.75	5.88	6.02	6.2	6.39	6.41	6.47	6.64	6.95	7.19	7.54	10.64	11.26

Cuando existe un gran número de variables y objetos los resultados no son fáciles de observar en este tipo de gráfica, para esto podemos representar cada variable como un vector usando sus coeficientes de correlación con los dos componentes representados en el plano como coordenadas (Gráfica 4.3). Las variables están representadas en un círculo de radio 1; de hecho, hay que señalar que estas dos componentes son ortogonales (en el sentido de que su coeficiente de correlación es igual a 0) y que una variable no puede estar fuertemente relacionada con dos componentes ortogonales simultáneamente [7]. En las Gráficas 4.3 y 4.4 la coordenada c_1 del punto terminal $C = (c_1, c_2)$ de cualquiera de esas variables (vectores) representa el coeficiente de correlación de dicha variable con el componente que está en el eje horizontal y la coordenada c_2 representa el coeficiente de correlación de la variable con el componente que está en el eje vertical. Se observa por ejemplo que la variable V8 en la gráfica 4.3 tiene coordenada aproximadamente de -0.9 en el primer componente y de aproximadamente 0.2 en el segundo componente.

Gráfica 4.3 Coeficientes de correlación entre las variables y los componentes 1 y 2

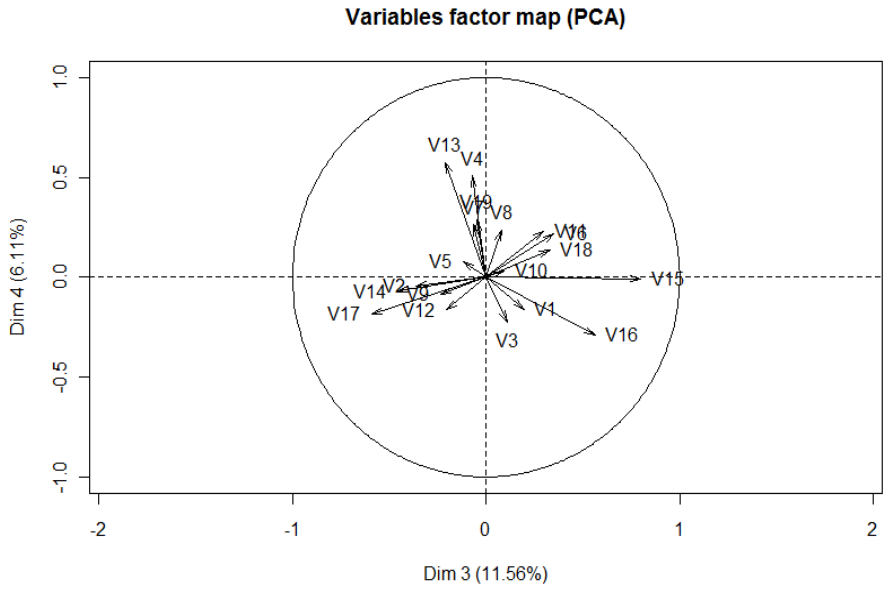


La Gráfica 4.3 muestra la representación de las variables en el plano formado por los dos primeros componentes en el cual se pueden observar las variables que están más altamente correlacionadas con cada componente. La variable V5 está altamente correlacionada positivamente con el primer componente, lo que indica que la variable "Ingreso per cápita " aporta mucha información para la formación de este primer componente; igualmente la variable V8 "por ciento de población bajo nivel de pobreza" ya que está altamente correlacionada negativamente con este mismo factor; de la misma forma la variable V12, "tasa bruta de mortalidad", está altamente correlacionada positivamente con el segundo componente lo que indicaría que ésta aporta bastante información para la construcción de este factor. Nótese que la variable V18 "índice vital" tiene una alta correlación negativa con el segundo componente.

Esta representación del conjunto de variables además permite una rápida visualización de las relaciones positivas o negativas entre las variables y la presencia de grupos de variables que están estrechamente relacionadas. Por ejemplo V8 y V9 están correlacionadas negativamente, lo que indica que cuanto más es el porcentaje de la población que cuenta con un empleo, menor es el índice de ésta que se encuentra bajo el nivel de pobreza establecido, o viceversa. Por su parte están positivamente correlacionadas; V8, V10 y V6, lo que indica por ejemplo que cuanto mayor es la proporción de población que está desempleada mayor es la cantidad de familias que no cuentan con un vehículo a su disposición y también mayor es la cantidad de personas bajo el nivel de pobreza.

De la misma manera se puede obtener una representación de las variables para el tercer y cuarto componente (Gráfica 4.4) donde se puede observar por ejemplo que la variable V13 está negativamente correlacionada con el tercer componente y positivamente correlacionada con el cuarto componente. Análisis similares se pueden realizar para las demás variables.

Gráfica 4.4 Coeficientes de correlación entre las variables y los componentes 3 y 4



4.2.2 Número de factores o componentes

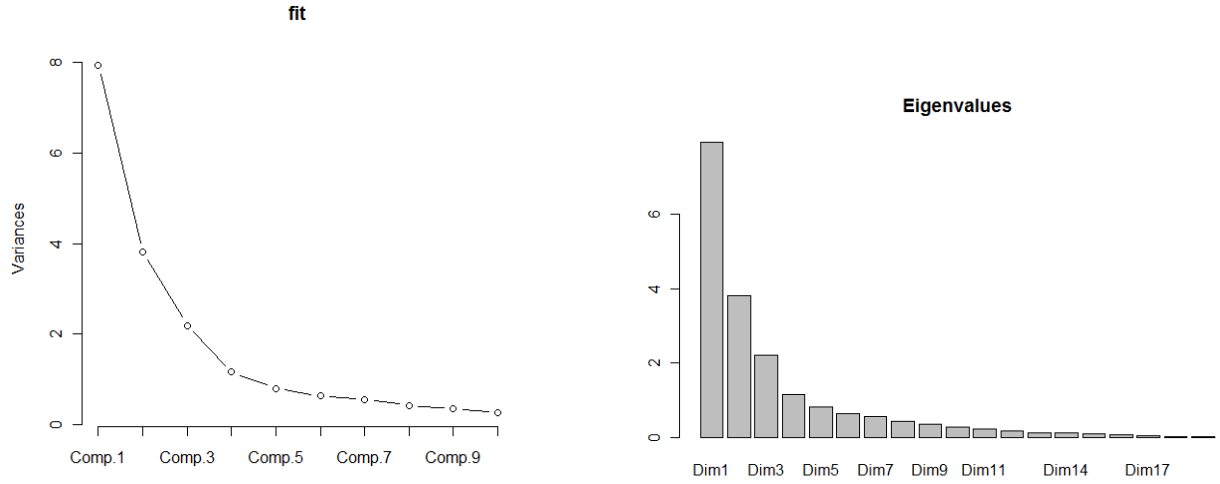
Para elegir el número de factores apropiado para el estudio, se consideraron los siguientes criterios:

a) La gráfica de sedimentación (Gráfica 4.5a), la cual sugiere 4 como el número adecuado de factores a extraer ya que es a partir de este número que la curva de sedimentación se estabiliza.

b) Complementariamente la Gráfica 4.5b muestra los valores de los Valores propios que corresponden igualmente a la varianza explicada para cada factor y es razonable pensar en 4 factores según dicha gráfica.

Gráfica 4.5 a) Gráfica de sedimentación y b) valores propios asociados a cada componente

a) b)



c) Otro de los criterios utilizados es mantener los factores cuyos valores propios son mayores que 1, la Tabla A5 (Columna Total de los valores propios) del apéndice muestra que son cuatro los factores que cumplen con este criterio.

d) Otra forma es el de mantener los factores hasta el punto en el que un factor adicional represente menos de la varianza de una variable típica, es decir, menos de un valor propio. En nuestro caso tenemos que una variable típica representa aproximadamente 100%/19 = 5.2633%; es decir mantendríamos 4 factores según los resultados de la Tabla A5 (Columna % de varianza) del Apéndice.

Se han extraído los cuatro componentes los cuales explican cerca del 80% de la varianza total de las variables iniciales (Tabla A5 del apéndice). El primer componente explica aproximadamente el 42% de la varianza total siendo el más importante; el segundo factor explica aproximadamente el 20%, el tercer componente el 11.5% mientras que el cuarto componente explica el 6.11% de la varianza total. Se puede apreciar que hay algunas

variables cuyos pesos iniciales son muy parecidos en varios componentes (Tabla A6 del Apéndice) y para dar una mejor interpretación a estos factores se aplicó la rotación Varimax cuyos resultados se muestran en la Tabla A7 del Apéndice.

4.2.3 Rotulación de los Componentes

Para rotular los componentes se procedió a analizar qué tipo de variables mostraban mayor correlación con ellos, utilizando los resultados de la tabla A7 del apéndice. A continuación se presenta la configuración de cada componente con su rotulación respectiva.

Componente 1

El primer componente tiene cargas positivas altas en las variables "Por ciento graduado de escuela superior o más"(V1), "Empleados en ventas y oficinistas"(V3), y "Población empleada"(V5), y cargas altas negativas en las variables "Familias sin vehículos disponibles(V6), "Población bajo el nivel de pobreza(V8)", "Empleados en agricultura pesca y silvicultura(V13)" y "Población mayor de 16 años desempleada(V10)"; a este componente se le ha llamado "**Condiciones de desarrollo socioeconómico básico"**

Según las cargas de las variables mostradas en la Tabla A6 del apéndice este componente expresado como índice estará dado por:

$$C1 = .743V1 + .741V3 + .630V9 - .650V6 - .856V8 - .614V10 - .799V13$$

De esta ecuación podemos resaltar que las variables V6, V8, V10, y V13, al tener una relación inversa con el factor, cuanto mayor es su valor, mayor es el impacto negativo sobre este componente y por ende sobre las condiciones económicas básicas de un municipio.

Componente 2

El segundo componente tiene cargas positivas altas en las variables "Empleados como gerenciales y profesionales(V4)", "Ingreso per cápita(V5)" , "Por ciento graduado de bachillerato o más(V7)" , "Densidad de población(V4)" e "Índice Gini(V19)". A este factor se le ha llamado "**Condiciones de desarrollo socioeconómico superior**"

Por lo tanto tenemos que:

$$C2 = .869V4 + .684V5 + .805V7 + .679V11 + .726V19$$

Aquí se puede resaltar que V4 es la variable que más peso tiene sobre el factor. Además dado que la relación entre las variables y el componente es directa, cuanto mayores sean los valores de las variables mayor será su impacto sobre dicho componente.

Componente 3

El tercer componente tiene cargas positivas altas en las variables "Población mayor de 65 años" (V2), "Tasa bruta de mortalidad por cada mil habitantes" (V12), "Índice de envejecimiento(V14)" y carga alta negativa en la variable "Índice vital"(V18). A este componente se le ha llamado "**Condiciones de edad avanzada**"

Expresado como índice:

$$C3 = .908V2 + .879V12 + .894V14 - .932V18$$

Aquí hay que resaltar que el hecho de que las variables V2, V12, y V14 estén relacionadas positivamente con el factor quiere decir que cuanto mayor sea el valor que éstas asuman, mayor será el impacto sobre el factor y que el hecho de que la variable V18 esté relacionada negativamente con este componente quiere decir que cuanto mayor sea el

valor del índice vital (lo que se da si el número de nacimientos es mucho mayor que el número de defunciones) menor será su impacto sobre las condiciones de edad avanzada.

Componente 4

El último componente tiene cargas positivas altas en las variables "Mujer cabeza de hogar sin esposo presente(V15)", "Abuelos que viven con sus nietos y son responsables de ellos(V16)", y tiene carga negativa alta en la variable "Por ciento de casados no separados(V17)". A este componente se le ha denominado "**Estructura no convencional de la familia**" que expresado como índice sería:

$$C4 = .917V15 + .685V16 - .815V17$$

Se puede decir entonces que la variable V15 es la que mayor impacto tiene sobre este componente. Además el hecho de que la variable V17 este en relación inversa con el factor quiere decir en términos prácticos que cuanto mayor sea el porcentaje de parejas de casados que no se separen menor serán las condiciones no convencionales de la familia (que sería lo ideal en cualquier sociedad)

4.3 Análisis de agrupación

El siguiente paso fundamental es proponer la clasificación de los 78 municipios en conglomerados de tal forma que a cada grupo pertenezcan los que tengan características socioeconómicas similares y que estas características entre grupos sean diferentes. Para

este análisis se usaron como variables de clasificación los cuatro componentes resultantes al aplicar el análisis de componentes principales discutidos en la sección anterior.

4.3.1 Determinación del número de conglomerados

La determinación del número de conglomerados es una de las tareas más difíciles ya que no existe una regla general que indique el número óptimo de grupos en que se deba particionar determinado número de elementos. En este sentido un amplio conocimiento de las características de los datos por parte del investigador y el buen uso de algunos criterios disponibles son la clave para lograr buenos resultados al respecto. Estos criterios para elegir el número de conglomerados no deben aplicarse ciegamente, pues en ocasiones nos pueden conducir a soluciones que no son adecuadas. En esta sección se aplicaron diferentes criterios comparando sus resultados y al final se tomó la decisión que se consideró más acertada para nuestro caso particular de agrupar los 78 municipios.

Mediante el uso de las Tablas de ANOVA obtenidas al aplicar un método de particionamiento hasta un número máximo de 10 conglomerados, se obtuvieron los valores de H y CH aplicando las expresiones 2.1 y 2.2 respectivamente. Por ejemplo usando las Tablas A8 y A9 del Apéndice que corresponde al análisis de varianza para un número de conglomerados igual 2 y 3 respectivamente se obtuvo:

que el contraste para ver si convenía pasar de dos grupos a tres fue:

$$H = \frac{76(3.466) - 75(2.787)}{2.787} = 19.515$$

Donde el valor 3.466 corresponde a la suma de la columna del cuadrado medio de error para 2 grupos y el valor 2.787 a la suma del cuadrado medio del error para 3 grupos.
y para el cálculo del valor del criterio CH según la ecuación 2.2 se encontró que:

$$CH = \frac{19.734 + 16.289 + 6.66 + 6.854}{.500 + .592 + .849 + .844} = 17.96$$

De la misma forma se calcularon los demás valores de H y CH y sus resultados se muestran en la Tabla 4.9; para esto se utilizaron las Tablas de ANOVA usando el método de las k-medias mediante el paquete estadístico SPSS (Tablas A8-A16 del Apéndice).

TABLA 4.9 Valores de H y CH para diferentes número de grupos

	G=2	G=3	G=4	G=5	G=6	G=7	G=8	G=9	G=10
C1	.812	.952	.912	.433	.449	.438	.324	.344	.276
C2	1.008	.492	.467	.484	.423	.437	.474	.413	.318
C3	.726	.383	.534	.448	.483	.475	.411	.418	.386
C4	.920	.960	.495	.577	.404	.352	.317	.275	.231
TOT.	3.466	2.787	2.408	1.942	1.759	1.702	1.526	1.45	1.296
H		19.51	12.80	18.75	8.59	3.41	9.18	4.67	9.19
CH	17.78	17.96	21.39	20.61	18.32	18.23	17.93	18.85	18.73

En la fila correspondiente a H de la Tabla 4.9, se observa que cuando pasamos de 5 a 6 conglomerados este índice tiene un valor menor que 10; indicándonos que 5 es el número óptimo de conglomerados para nuestro caso, según el criterio de Hartigan. Según el criterio CH expuesto en la ecuación 2.2 lo ideal es hacer una partición de los municipios en 4 conglomerados como primera medida o 5 conglomerados con un valor bastante cercano.
La Tabla 4.10 resume los valores óptimos de las medidas internas de validación aplicadas a los diferentes métodos y cuya salida se muestra en la Tabla 4.16 en el apéndice.

TABLA 4.10 Valores Óptimos según medidas internas de Validación

Valores óptimos			
	Valor	Método	Conglomerados
Conectividad	5.0246	Jerárquico	2
Dunn	0.4908	Jerárquico	2
Silueta	0.5178	Jerárquico	2

La Gráfica 4.6, la cual se elaborada utilizando la librería clValid del programa R-project, muestra estos resultados de una manera más clara para los métodos de particionamiento jerárquico, el método de k-medias y el método PAM. Según estos resultados se sugiere aplicar un método jerárquico con dos conglomerados de acuerdo a los índices de Conectividad, Dunn y Silueta, pues de acuerdo a lo expuesto por Brock et al en [8]; el primero de ellos se debe minimizar, mientras que los otros dos índices toman un valor óptimo en su máximo.

Gráfica 4.6 Medidas internas de validación

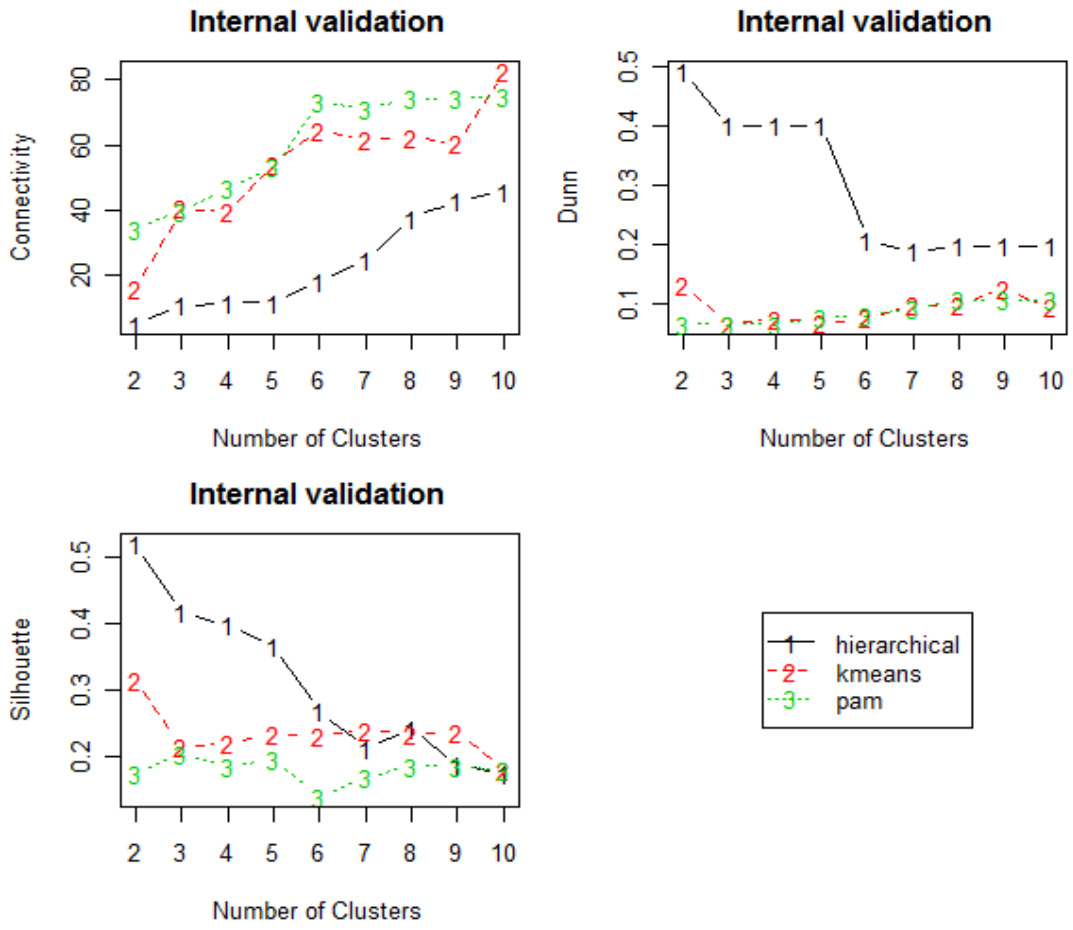
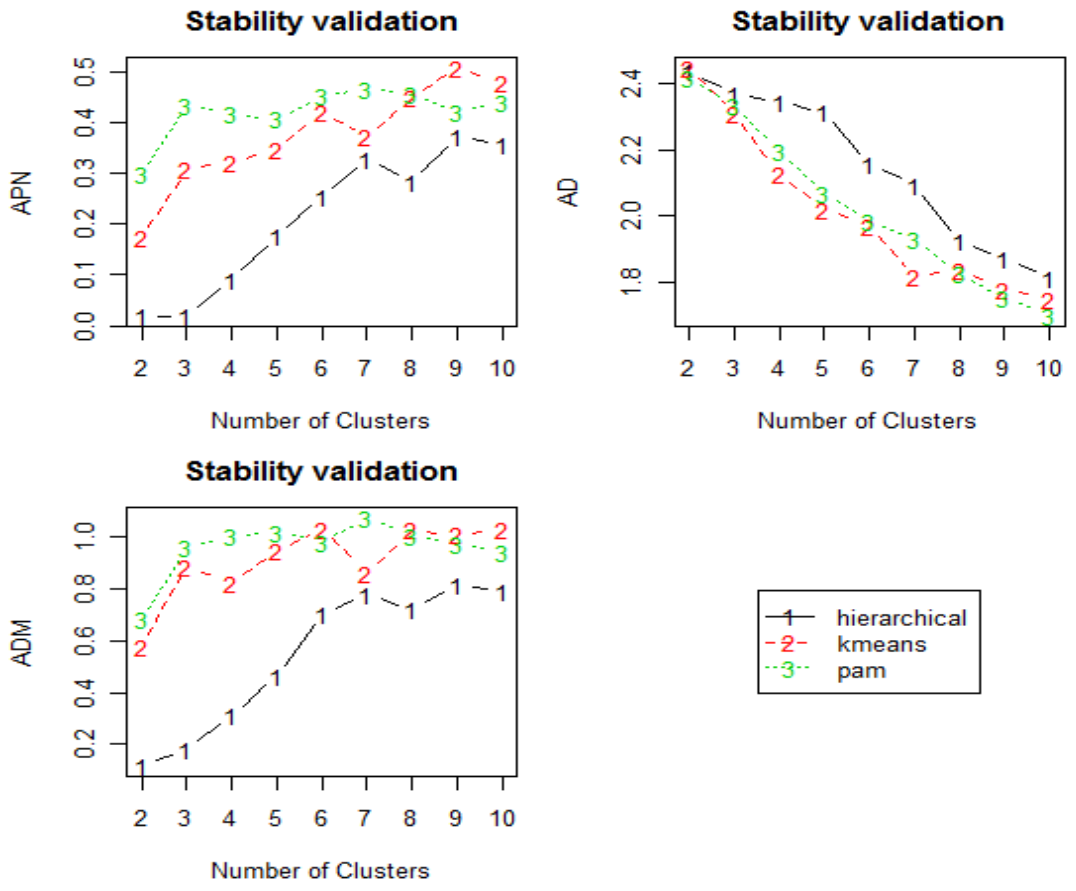


TABLA 4.11 Valores Óptimos según medidas de estabilidad para validación

Valores óptimos			
	Valor	Método	Conglomerados
APN	0.0188	Jerárquico	2
AD	1.6991	Pam	10
ADM	0.1167	Jerárquico	2
FOM	0.9502	Jerárquico	10

De la misma forma se analizaron las medidas de estabilidad cuyos resultados se resumen en la Tabla 4.12 y la Gráfica 4.7 nos sugieren aplicar un método jerárquico en tres de los casos y el algoritmo PAM con 10 conglomerados en un caso.

Gráfica 4.7 Medidas de estabilidad para validación



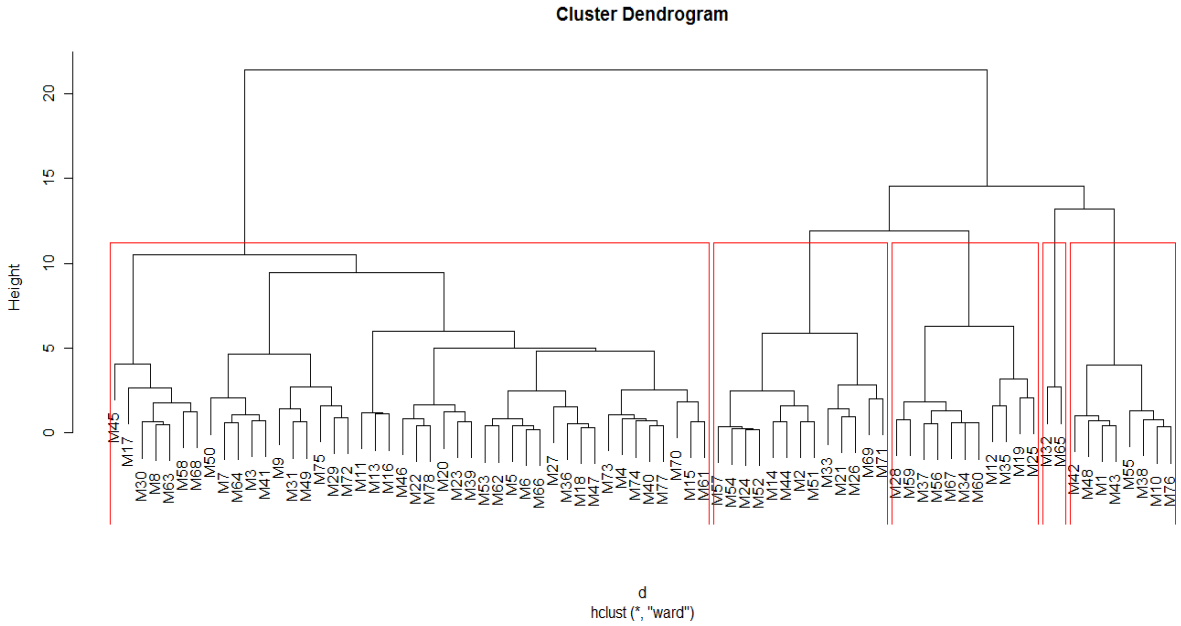
Teniendo en cuenta estos resultados se ha considerado: primero con respecto al número sugerido por estas medidas, que para nuestro caso no es conveniente dividir los 78 municipios solamente en 2 conglomerados, y con respecto al método de agrupamiento a utilizar se han considerado algunas recomendaciones de autores citados en la bibliografía

sobre usar preferiblemente algoritmos de particionamiento pues cumplen con ciertas propiedades que los hacen más adecuados. Podemos utilizar los métodos jerárquicos y sus representaciones gráficas (Dendogramas) con el objetivo de observar algunas características de los datos y como ayuda para elegir el número apropiado de conglomerados, pero no para tomar la clasificación ahí definida como la definitiva, pues tienen la desventaja de que su rigidez evita que se pueda corregir lo que ya se ha hecho; además que el dendograma correspondiente a un determinado método jerárquico no es único, por tal razón su estructura de agrupamiento mostrada no representa con certeza las verdaderas distancias entre las observaciones [18].

La Gráfica 4.8 muestra el dendograma correspondiente de los municipios, el cual se construyó usando la distancia Euclídea y el método jerárquico de Ward; en el cual se observa una posible partición de éstos después de hacer un corte con 5 grupos.

Nótese por ejemplo que los municipios de San Juan (M65) y Guaynabo (M32), que anteriormente se había comentado eran puntos extremos, están en el mismo conglomerado. También obsérvese que bajo este algoritmo hay un conglomerado que contiene más de la mitad de los municipios.

Gráfica 4.8 Dendograma correspondiente a los 78 municipios

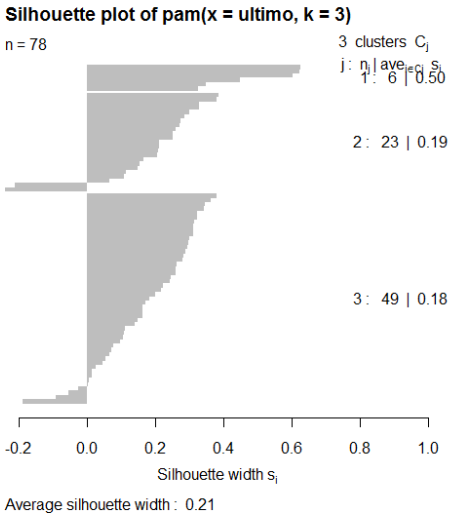


4.3.2 **Aplicación del método de partición PAM**

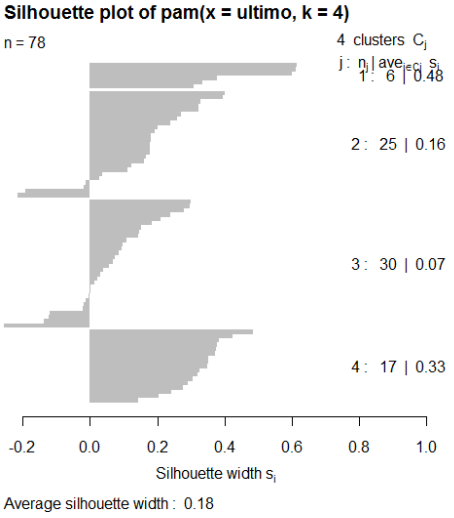
Dado lo mencionado anteriormente y teniendo en cuenta que el comportamiento de los dos métodos de particionamiento que se han analizado PAM Y k-medias tienen rendimientos muy similares según Gráficas 4.6 y 4.7, se optó finalmente por aplicar el método de particionamiento alrededor de los medoides PAM puesto que generalmente suele dar mejores resultados que el método de las k-medias[18]. Además cuando se analizaron los datos se encontró que algunos de ellos resultaron ser atípicos, y en las apreciaciones hechas en la sección 2.5.3 se mencionó la ventaja de utilizar el algoritmo PAM en estos casos. Para esto se utilizó el comando PAM en la librería {cluster} del paquete estadístico R, el cual provee como criterio para establecer el número de grupos el promedio del ancho de silueta (gráficas 4.9-4.12). Se ha considerado solo hasta seis grupos pues los valores del

ancho de silueta para un número más grande de grupos tienden a descender. En ellas se puede observar básicamente que este método nos sugiere una agrupación en 3 o 5 conglomerados pues recuerdese que este valor debe ser maximizado; confirmando nuestros análisis anteriores de que 5 es el número óptimo de grupos.

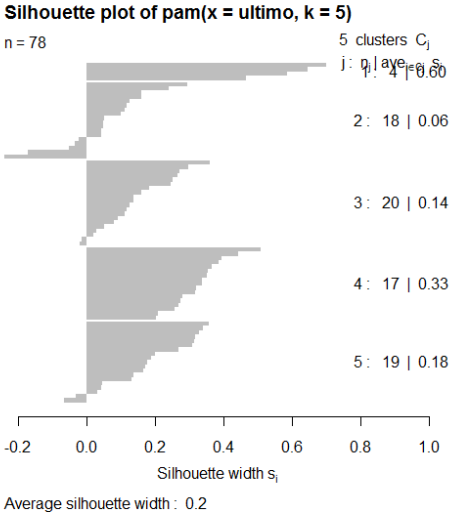
Gráfica 4.9 Ancho de silueta para 3 grupos



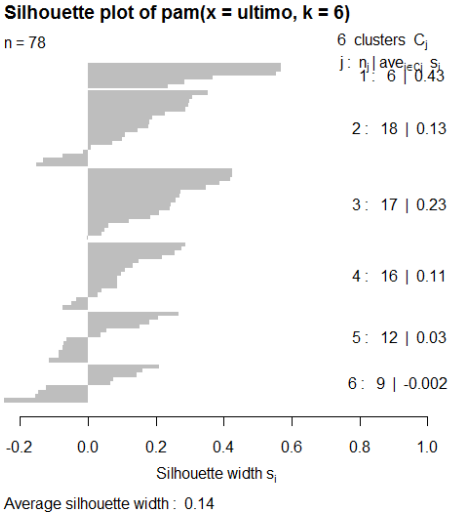
Gráfica 4.10 Ancho de silueta para 4 grupos



Gráfica 4.11 Ancho de silueta para 5 grupos



Gráfica 4.12 Ancho de silueta para 6 grupos



En la Tabla 4.12 se muestran los municipios con su respectiva clasificación de grupo después de aplicar el algoritmo PAM, la Gráfica 4.13 muestra el número de municipios en cada uno de estos conglomerados y en la Gráfica 4.18 el mapa de Puerto Rico con la clasificación de los municipios.

Gráfica 4.13

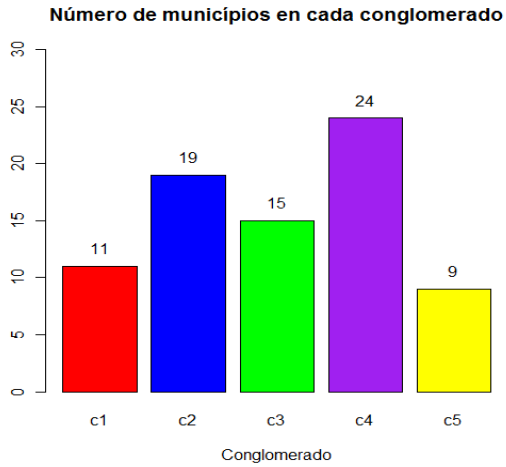
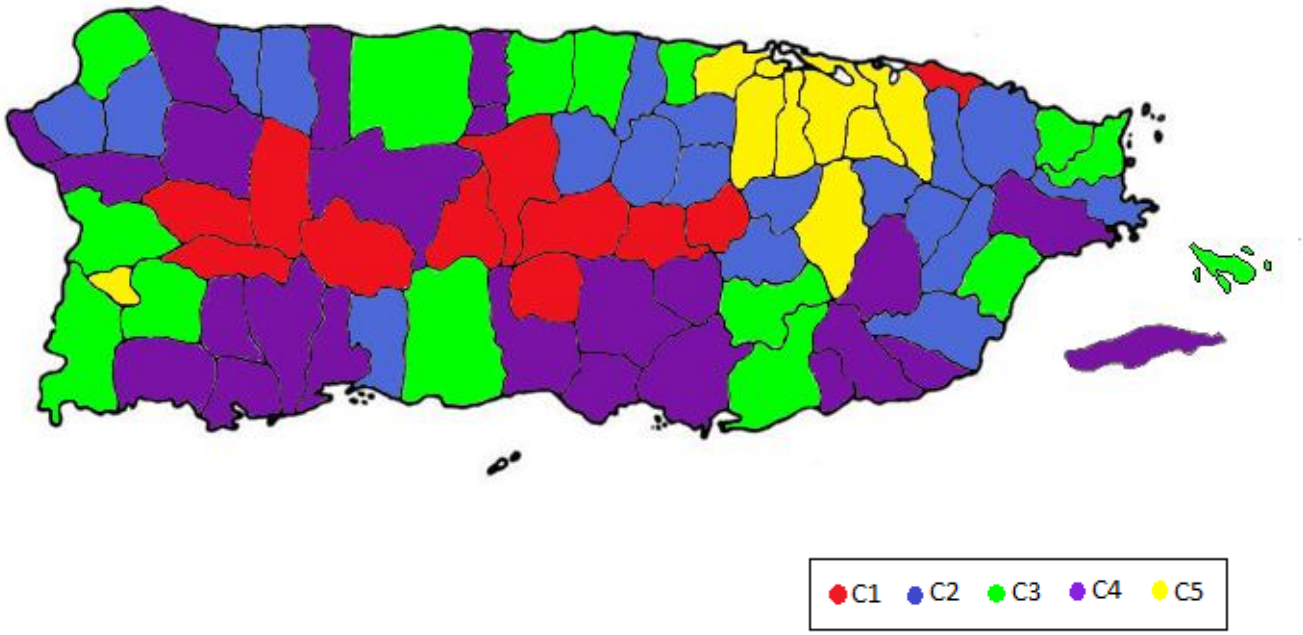


TABLA 4.12 Distribución de los municipios en los 5 conglomerados

Conglomerado 1	Conglomerado 2	Conglomerado 3	Conglomerado 4		Conglomerado 5
Adjuntas	Aguada	Aguadilla	Aibonito	Naguabo	Bayamón
Barranquitas	Aguas Buenas	Arecibo	Añasco	Patillas	Caguas
Ciales	Camuy	Cabo Rojo	Arroyo	Rincón	Carolina
Comerío	Canovanas	Cayey	Barceloneta	Saban Grande	Cataño
Jayuya	Ceiba	Culebra	Coamo	Salinas	Guaynabo
Lares	Cidra	Dorado	Florida	San Lorenzo	Hormigueros
Las Marías	Corozal	Fajardo	Guánica	San Seb.	San Juan
Loíza	Gurabo	Guayama	Guayanilla	Santa Isabel	Toa Baja
Maricao	Juncos	Humacao	Hatillo	Utuado	Trujillo Alto
Orocovis	Las Piedras	Luquillo	Isabela	Vieques	
	Moca	Manatí	Juana Díaz	Yauco	
	Morovis	Mayaguez	Lajas		
	Río Grande	Ponce	Maunabo		
	Toa Alta	San Germán			
	Vega Alta	Vega Baja			
	Yabucoa				

Gráfica 4.14 Mapa de los municipios de Puerto Rico distribuidos en los 5 conglomerados



4.3.3 Descripción de los conglomerados

En la Tabla 4.13 se muestran los cuatro componentes con cada una de las variables que están altamente correlacionadas con éstos (ya sea positivamente o negativamente) lo que nos ayuda a la interpretación del análisis que se muestra a continuación.

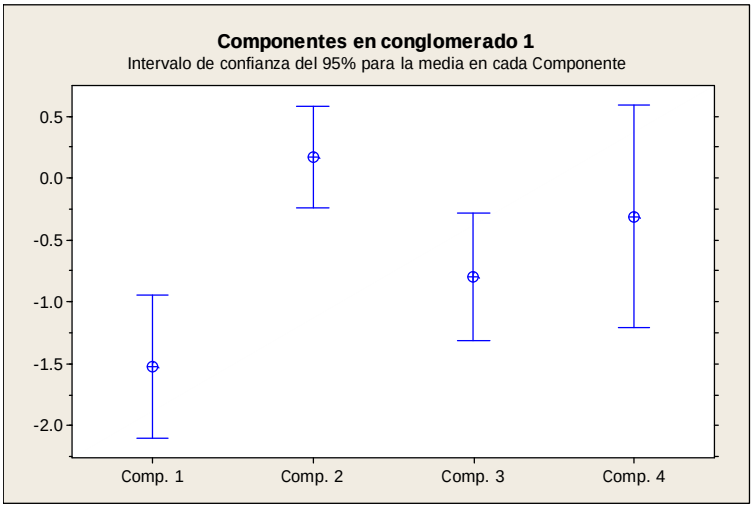
TABLA 4.13 Cargas altas de las variables en cada componente

VARIABLES	*Condiciones de desarrollo socioeconómico básico*	*Condiciones de desarrollo socioeconómico superior*	*Condiciones de edad avanzada*	*Estructura no convencional de la familia*
Grad_esc_sup	.743			
Trab_Vent_ofic	.741			
Pob_Empleada	.737			
Pob_sin veh dispon.	-.650			
Pob._bajo_Pob	-.856			
Pob.>16 des.	-.614			
Trab_agr_pes_sil	-.799			
Trab_Ger Prof.		.869		
Íngr percápita		.684		
Grad_Bach.más		.805		
Densidad_Pob.		.679		
Indice Gini		.726		
Pob>65			.908	
Tasa bruta_mort.			.879	
Indice_envej.			.894	
Indice vital			-.932	
Muj_cab_hog				.917
Ab_resp_ nietos				.685
Cas_no sep.				-.815

Se estimaron intervalos de confianza del 95% alrededor de la media de las ponderaciones de cada componente en los 5 conglomerados resultantes. Es importante anotar que los componentes están normalizados con media cero y desviación estándar uno. Por tanto, un componente particular será significativo en determinado conglomerado si este intervalo no contiene el cero. Se hace una descripción de las características más importantes en cada uno de los conglomerados y la información completa sobre las variables que se tuvieron en cuenta para el estudio se puede apreciar en las gráficas de la 4.24 a la 4.42.

Conglomerado 1

Gráfica 4.15



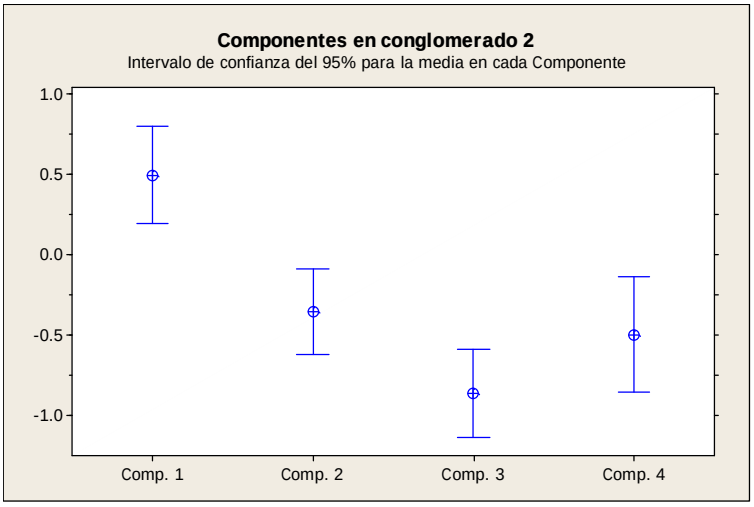
Básicamente este conglomerado se caracteriza por tener promedios bajos en los componentes "Condiciones de desarrollo socioeconómico básico" y "Condiciones de edad avanzada". A este grupo pertenecen los municipios con el nivel económico más bajo. Como podemos observar las variables que están altamente correlacionadas con el primer componente se destaca que menos del 50% de la población que pertenece a estos municipios se graduó de escuela superior; el 63.9% de su población aparece registrado bajo el nivel de pobreza y tiene el mayor porcentaje de personas cuya labor económica está relacionada con la agricultura, la pesca y la silvicultura comparado con los demás grupos. En cuanto a las variables que están relacionadas con el componente "Condiciones de desarrollo socioeconómico superior" para este grupo de municipios, se destaca que su ingreso per cápita es el más bajo comparado con los otros grupos con $ 5,066.8 y su densidad de población es la más baja, con 576 habitantes por milla cuadrada. En el componente "Condiciones de edad avanzada", cuyas variables están relacionadas con las condiciones de

edad avanzada, este grupo tiene uno de los porcentajes más bajos en personas mayores de 65 años con tan solo 9.24%, también tiene el índice de envejecimiento más bajo con un promedio de entre 34 y 35 adultos mayores de 65 años por cada 100 menores de 15 años. Para el último componente "Estructura no convencional de la familia" se destaca que el porcentaje de mujeres que son responsable del hogar es del 18.83% siendo uno de los más bajos y que el porcentaje de población casada no separada es uno de los más altos.

Conglomerado 2

Este grupo de municipios se caracteriza por tener promedios bajos en los componentes "Condiciones de desarrollo socioeconómico superior", "Condiciones de edad avanzada" y "Estructura no convencional de la familia" y promedio relativamente alto en el componente "Condiciones de desarrollo socioeconómico básico".

Gráfica 4.16

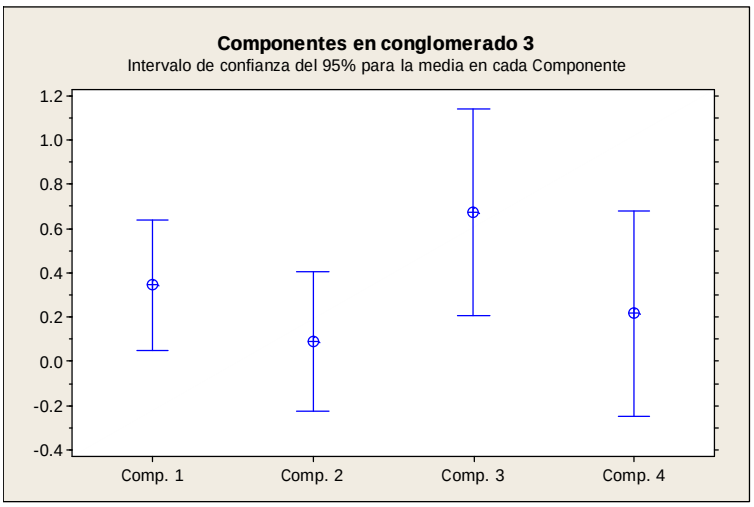


Para las variables relacionadas con el componente “Condiciones de desarrollo socioeconómico básico” se destaca que su tasa de promedio de desempleo es del 7.71%, siendo ésta una de las más bajas comparadas con los otros grupos. Para las variables que conforman el componente “Condiciones de desarrollo socioeconómico superior” en este grupo se aprecia que en promedio el 13.66% de la población de estos municipios se graduó de bachillerato siendo este un porcentaje relativamente bajo. En cuanto a las variables del componente “Condiciones de edad avanzada”, estos municipios tienen el porcentaje más bajo de su población que son mayores de 65 años con sólo el 8.97%, también tiene el índice más bajo en la tasa de mortalidad con un valor de 6.16 durante el año 2000, indicando esto que hubo aproximadamente 6 defunciones por cada 1000 habitantes que pertenecen a este grupo. En las variables que conforman el componente “Estructura no convencional de la familia” este grupo tiene el porcentaje más bajo en las mujeres que son cabeza de hogar y también el porcentaje más bajo de abuelos que son responsables de sus nietos, además tiene el mayor porcentaje de población casada no separada.

Conglomerado 3

Tiene promedios altos en los componentes “Condiciones de desarrollo socioeconómico básico” y “Condiciones de edad avanzada”. En lo relacionado con las variables del primer de estos componentes, el 58.78% de la población se graduó de bachillerato, siendo uno de los más altos; su porcentaje de población bajo nivel de pobreza es 48.79%, uno de los más bajos comparado con los demás conglomerados.

Gráfica 4.17

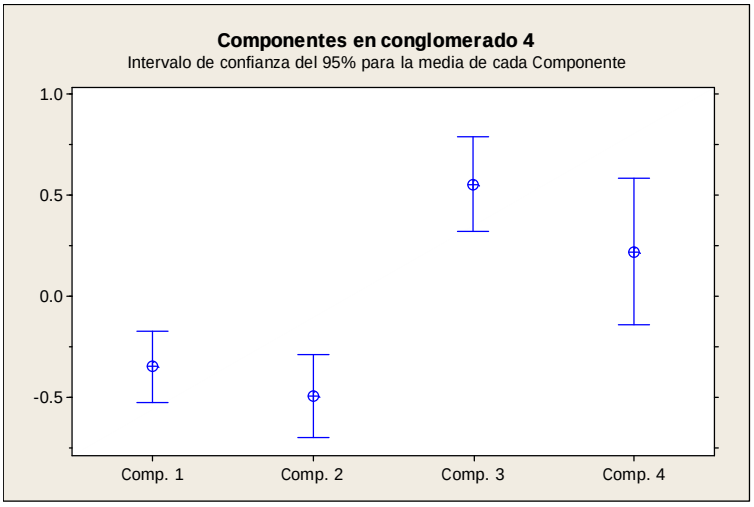


En cuanto al componente "Condiciones de desarrollo socioeconómico superior", en este conglomerado el ingreso per cápita promedio es de $ 7,695 siendo este el segundo más alto de los 5 grupos, el 16.28% de su población obtuvo por lo menos un título de bachillerato, siendo el segundo más elevado con relación a los otros conglomerados y el 25.6% tiene sus empleos en el área gerencial o profesional. Para las variables correspondientes al componente "Condiciones de edad avanzada", se destaca que este grupo tiene el índice de envejecimiento más alto con 50.45, lo que indica que hay aproximadamente una persona mayor de 65 años por cada dos niños menores de 15 años. Para el último componente se observa que tiene uno de los porcentajes más altos en la variable "abuelos que viven con sus nietos y son responsables de ellos" con 15.76%.

Conglomerado 4

Se caracteriza básicamente por tener promedios relativamente bajos en los componentes "Condiciones de desarrollo socioeconómico básico" y "Condiciones de desarrollo socioeconómico superior" y además un promedio alto en el componente "Condiciones de edad avanzada". Con respecto a las variables correspondientes al primer componente presenta uno

Gráfica 4.18

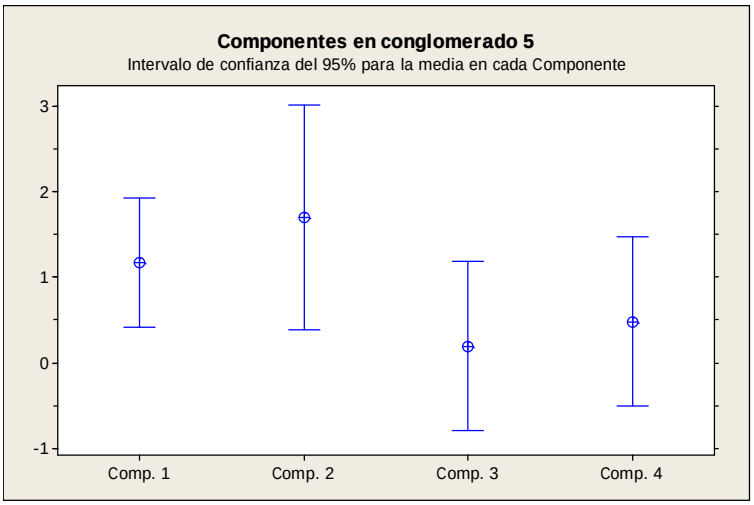


de los porcentajes más altos de personas que están bajo el nivel de pobreza con el 56.78%, superado sólo por el conglomerado 1, figura con el nivel de desempleo más alto con un 9.76% comparado con los demás conglomerados. En el componente "Condiciones de desarrollo socioeconómico superior" se destaca que la variable "trabajadores como gerenciales o profesionales" presenta uno de los porcentajes más bajos con el 22.29% y su densidad de población también es una de las más bajas con 680.4 habitantes por milla cuadrada. En el componente "Condiciones de edad avanzada" la variable "índice vital" presenta el valor

más bajo con 1.93 indicando que nacen aproximadamente 2 personas por cada persona que fallece. Para el último componente este grupo de municipios presenta uno de los porcentajes más altos en la variable "abuelos que viven con sus nietos y son responsables de ellos".

Conglomerado 5

Gráfica 4.19



En este conglomerado son significativos los componentes "Condiciones de desarrollo socioeconómico básico" y "Condiciones de desarrollo socioeconómico superior" y de hecho aquí están los municipios que según los resultados tienen los mejores indicadores económicos. En el primer componente se puede apreciar que las variables relacionadas con el bienestar económico tienen los porcentajes más altos, se destaca la marcada diferencia en la variable "población bajo nivel de pobreza" con 10 puntos porcentuales por debajo del siguiente conglomerado. Para las variables correspondientes al componente "Condiciones de desarrollo socioeconómico superior" sobresale la diferencia en las variables "ingreso per cápita" (Gráfica 4.32), "por ciento de población graduada de bachillerato" (Gráfica 4.35)

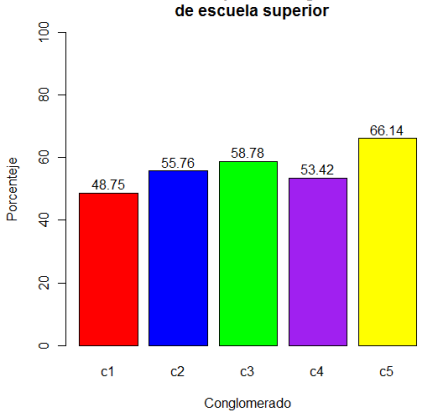
y la más notable de todas es la variable “densidad de población” que tiene un promedio de 4,239 habitantes por milla cuadrada, aproximadamente 4 veces mayor que la densidad de población del conglomerado que le sigue.

En los componentes “Condiciones de edad avanzada” y “Estructura no convencional de la familia”, sobresalen las variables “índice de envejecimiento” y “mujer cabeza de hogar” con puntajes de 52.35 y 22.78%, lo que indica que en el primer caso hay más de 50 personas mayores de 65 años por cada 100 menores de 15 años y en el segundo caso que en ese porcentaje de hogares tienen a una mujer como jefe de hogar sin que su esposo esté presente.

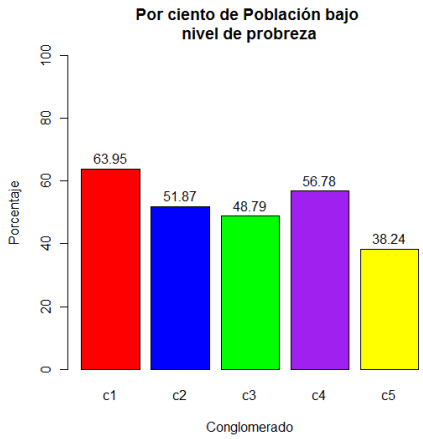
Como se mencionó anteriormente las siguientes gráficas presentan la información completa de los cinco conglomerados en cada una de las 19 variables que se tomaron en el estudio.

- **Variables que conforman el componente “Condiciones de desarrollo socioeconómico básico”**

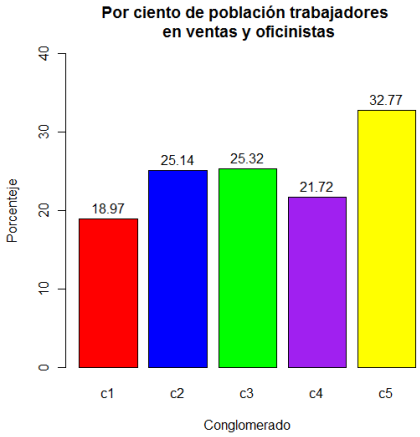
Gráfica 4.20



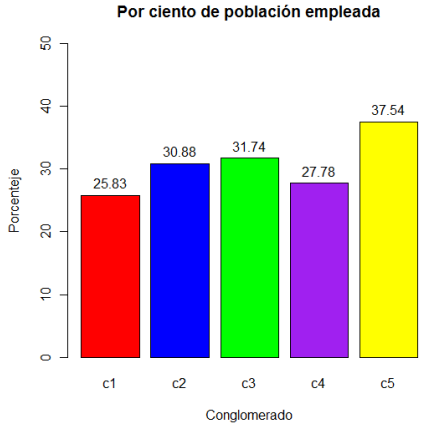
Gráfica 4.21



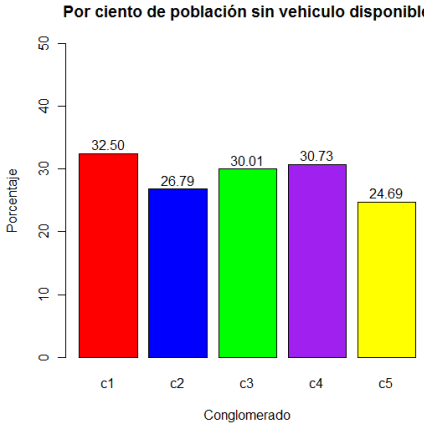
Gráfica 4.22



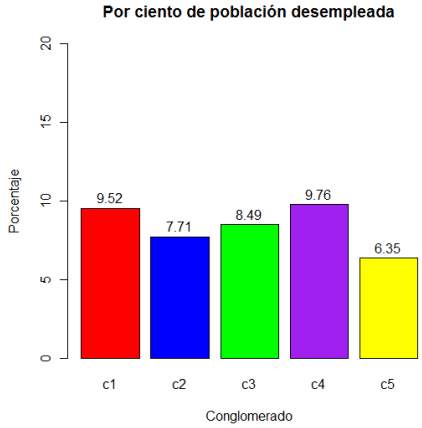
Gráfica 4.23



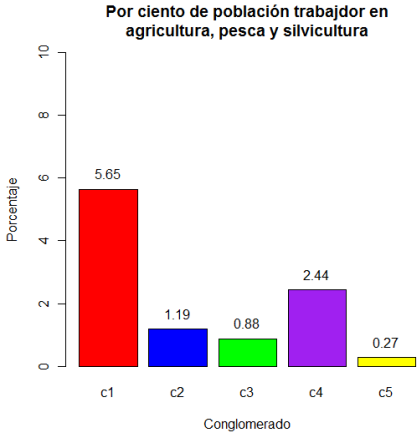
Gráfica 4.24



Gráfica 4.25

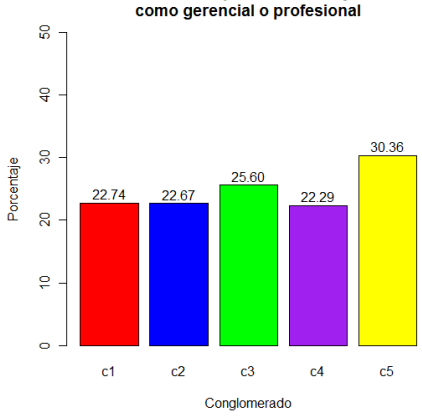


Gráfica 4.26

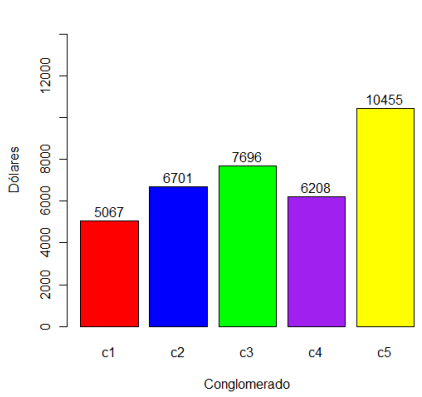


- **Variables que conforman el componente "Condiciones de desarrollo socioeconómico superior "**

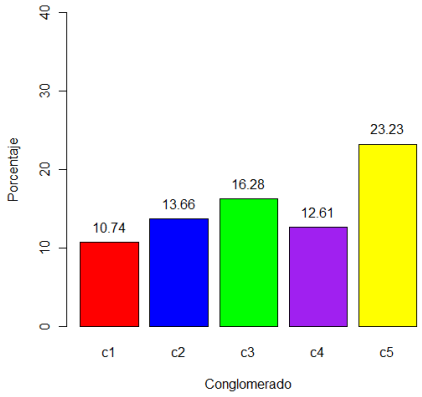
Gráfica 4.27



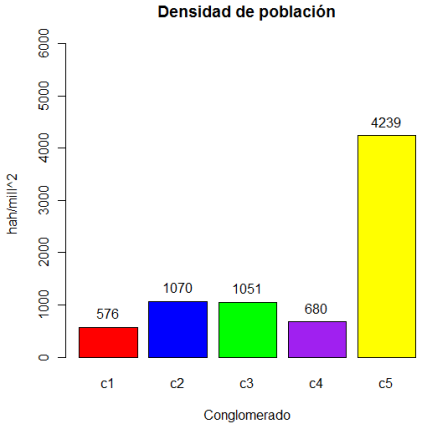
Gráfica 4.28



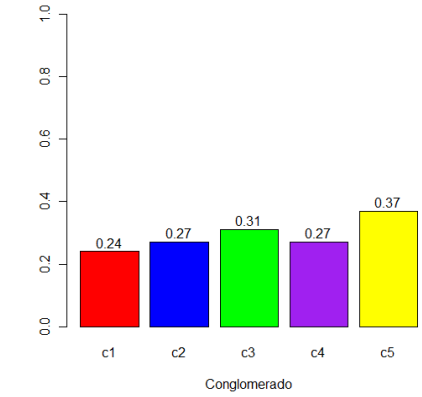
Gráfica 4.29



Gráfica 4.30

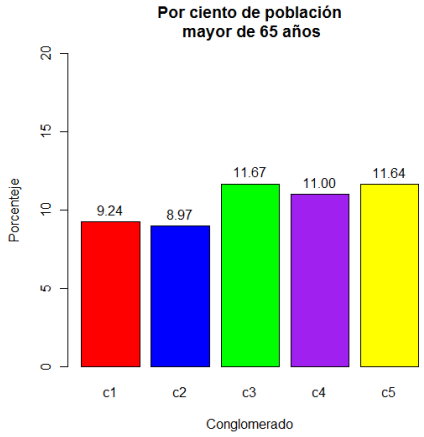


Gráfica 4.31

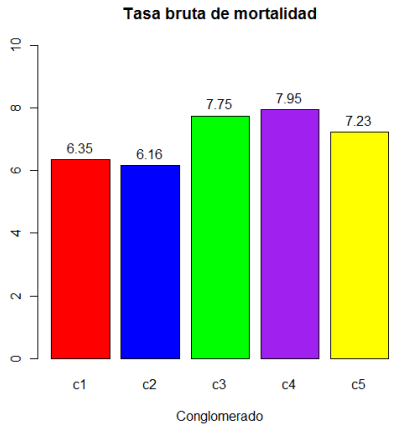


- **Variables que conforman el componente “Condiciones de edad avanzada”**

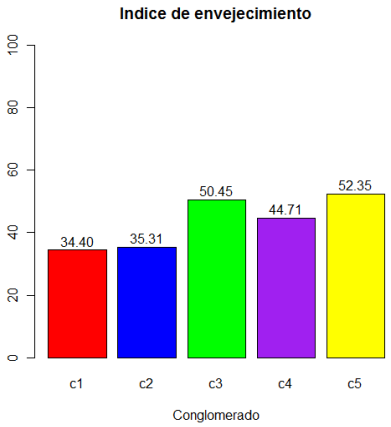
Gráfica 4.32



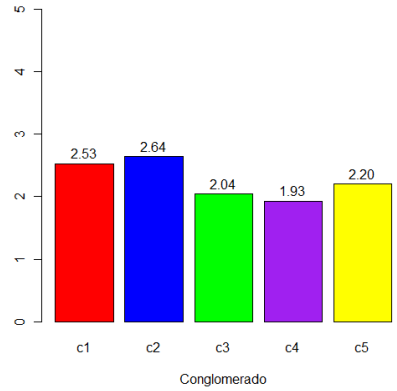
Gráfica 4.33



Gráfica 4.34

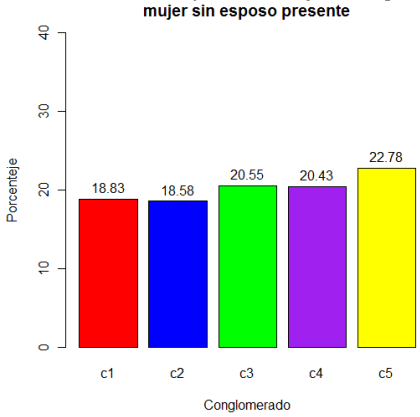


Gráfica 4.35

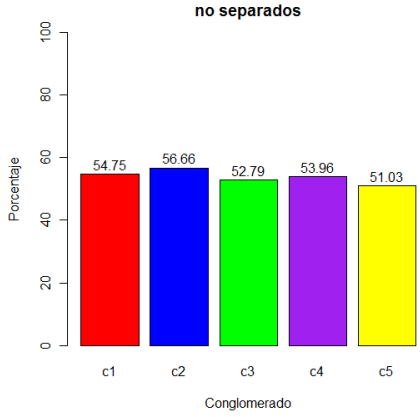


- **Variables que conforman el Componente “Estructura no convencional de la familia”**

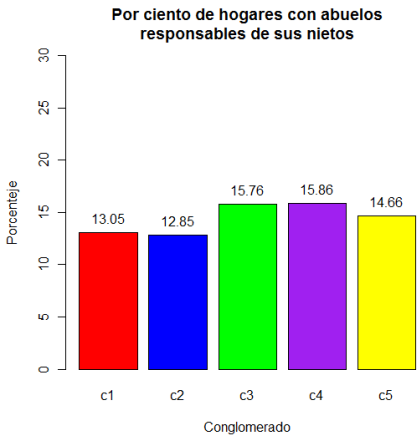
Gráfica 4.36



Gráfica 4.37



Gráfica 4.38



5 CONCLUSIONES

Mediante el uso del análisis de componentes principales se obtuvo cuatro grupos diferentes de variables. El componente “Condiciones de desarrollo socioeconómico básico” está formado por aquellas variables que miden las condiciones socioeconómicas básicas; el componente “Condiciones de desarrollo socioeconómico superior” formado por variables que miden las condiciones socioeconómicas de un desarrollo más elevado; el componente “Condiciones de edad avanzada” contiene variables que miden las características de la población envejeciente y las variables del componente “Estructura no convencional de la familia” miden condiciones no tradicionales de estructura del hogar. Con los resultados de este análisis de componentes principales se aplicó un análisis de clasificación, el cual permitió dividir los 78 municipios en 5 conglomerados con algunas características que sobresalen en cada uno:

- En el conglomerado 1 se encuentran los municipios con las condiciones económicas más bajas y tiene como característica principal un bajo desarrollo en las variables relacionadas con el nivel de educación y actividad económica. Geográficamente estos municipios se encuentran ubicados en el centro de la Isla, con la excepción del municipio de Loíza que está localizado en el noreste.
- El conglomerado 2 se caracteriza principalmente por sus valores bajos en las variables que miden las condiciones de edad avanzada y las que miden la estructura

no convencional del hogar; muestra promedios relativamente adecuados en las condiciones socioeconómicas básicas pero presenta índices bajos en algunas de las variables relacionadas con las condiciones socioeconómicas de desarrollo más elevado. Podríamos decir en términos generales que este es un grupo de municipios de clase media comparado con los demás grupos.

- El conglomerado número 3 se caracteriza por tener un promedio alto en las variables relacionadas con las condiciones de edad avanzada, como se puede apreciar en las gráficas de dichas variables. Este grupo tiene unos de los índices más elevados en el porcentaje de personas mayores de 65 años, la tasa de mortalidad e índice de envejecimiento. Además de un bajo puntaje en la variable "índice vital". Sus condiciones en las variables relacionadas con el nivel de educación y las características del empleo son adecuadas.
- El cuarto conglomerado presenta índices relativamente bajos en los componentes uno y dos, caracterizándose por tener promedios bajos en los niveles de educación y por tanto promedios bajos también en la población con empleos profesionales. Además, presenta promedios altos en las variables relacionadas con la estructura no convencional del hogar, como en el caso de la variable "abuelos responsables de sus nietos". Es de resaltar que este grupo de municipios presenta el índice más alto en la tasa de mortalidad.

- Los municipios que conforman el conglomerado cinco, el cual cuenta con las mejores condiciones económicas, están ubicados básicamente en el área

metropolitana de la isla a excepción del municipio de Hormigueros que está ubicado en el área oeste. Sin embargo, como cualquier región, tiene algunas dificultades, específicamente en materia de la distribución de ingresos, la cual está determinada por el Índice Gini que presenta el nivel más alto comparado con los demás grupos, lo que en términos prácticos significa que en estos municipios es donde se presenta mayor desigualdad en la distribución de los ingresos entre sus habitantes. También el índice de mortalidad, aunque no presentó el valor más alto, se encuentra entre los grupos con mayor tasa de mortalidad.

- En los conglomerados 3 y 5 hay mayor proporción de personas mayores de 65 años comparado con los otros conglomerados.

- Los conglomerados 1 y 2 tienen las tasas de mortalidad más bajas, lo que indica que en estos grupos es menor el número de defunciones por cada 1000 habitantes, además estos mismos grupos tienen los índices de envejecimiento más bajos, lo que indica que son los conglomerados donde hay menor cantidad de personas mayores de 65 años con respecto a la población menor de 15.
- Los grupos 3 y 4 son los que tienen un menor valor en el índice vital, lo que muestra que la razón entre el número de nacimientos y las defunciones es menor que en los demás grupos.
- En los grupos 3 y 5 es donde hay mayores porcentajes de mujeres cabeza de hogar.

- Las personas que se casan tienden a separarse menos en los grupos 1 y 2 que en los demás grupos; además en estos mismos grupos el porcentaje de abuelos que viven con sus nietos y son responsables de ellos es más bajo.

6 LIMITACIONES Y TRABAJOS FUTUROS

Limitaciones

- A la fecha de culminación de esta tesis aún no se había publicado información de las variables incluidas en este estudio a nivel municipal para el Censo realizado en Puerto Rico en 2010, lo que impidió la creación del perfil correspondiente a este año y la posterior comparación de los perfiles de los años censales 2000 y 2010.

- Se quiso incluir en el estudio otro tipo de variables que aportaran información a la condición social y económica, específicamente en el campo de la tecnología y seguridad, pero no fue posible encontrar información de tales variables para todos los municipios.

Trabajo futuro

- Realizar un perfil socioeconómico usando datos del censo 2010 y hacer la comparación respectiva con los resultados aquí obtenidos.
- Incluir otro tipo de variables que aporten más información en el campo socioeconómico, en la creación de este nuevo perfil.
- Se recomienda a estudiosos de los campos de sociología, economía, demografía y de otras áreas de las ciencias sociales que utilicen los hallazgos aquí obtenidos para realizar investigaciones sobre posibles causas y consecuencias de los componentes encontrados.

7 REFERENCIAS

[1] TENKO, R. y GEORGE, A. An Introduction to Applied Multivariate Analysis. New York, Taylor & Francis Group, 2004.

[2] BRUCE T. Exploratory and confirmatory factor analysis. Washington, American Psychological Association, 2002

[3] AFIFI, A. y CLARK V. Computeraided Multivariate Analysis. 3nd Ed. Los Angeles , Chapman & Halucrc, 1997.

[4] KACHIGAN, S. Multivariate Statistical Analysis: A Conceptual Introduction. 2da Ed. Radius Press, 1991

[5] ACUÑA, E. Análisis de Regresión. Mayagüez, Departamento de Matemáticas, Universidad de Puerto Rico-Recinto Universitario de Mayagüez, 2007

[6] PEÑA, Daniel. Análisis de datos multivalentes, España, Ed. Mc Graw Hill, 2002

[7] HUSSON, F., LE, S., PAGES, J. Exploratory Multivariate Analysis by Example Using R. New York, CRC Press Taylor & Francis Group. 2011

[8] CIValid: An R package for cluster validation por BROCK, G., PIHUR V., DATTA S., y DATTA S. Journal of Statistical Software, 25(4), March 2008.

[9] BROCK, G., PIHUR V., DATTA S., y DATTA S.: An R package for cluster validation. University of Louisville. 2011

[10] LE, S., JOSSE, J., HUSSON, F. FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software, 25(1). 2008.

[11] GONZÁLEZ, M. A Comparison in Cluster Validation Techniques. Tesis. Mayagüez, Puerto Rico. Universidad de Puerto Rico, Facultad de Artes y Ciencias, 2005

[12] JOHNSON, R y WICHERN D. Applied Multivariate Statistical Analysis. 6ta Ed. New Jersey, Pearson/Prentice Hall. 2007

[13] WIT E. y MCCLURE J. Statistics for Microarrays: Design, Analysis, and Inference**.** England, Wiley & Sons Ltd. 2004

[14] TAI-HOON, K. y HOJJAT A. Advances in Computer Science and Information Technology. Japan, Springer, 2010.

[15] DIAZ, L**.** Estadistica Multivariada: Inferencia y Métodos. Bogotá, Universidad nacional de Colombia. 2002

[16] CUADRAS, C. Nuevos métodos de análisis multivariante. Barcelona, CMC Editions , España. 2008

[17] CASTRILLÓN, O. Uso de técnicas multivariadas y modelos estadísticos para el análisis del desempeño académico de los estudiantes de cálculo I. Tesis. Mayagüez, Puerto Rico. Universidad de Puerto Rico, Facultad de Artes y Ciencias, 2007

[18] ACUÑA, E. Unsuperviser Classification Clustering (clase 15). Mayagüez, Puerto Rico. Universidad de Puerto Rico, Facultad de Artes y Ciencias.

[19] RULLAN, J. Situación de salud en Puerto Rico, Secretario de Salud; Indicadores Básicos 2000, Organización Panamericana de la Salud.

[20] RASIC, I. Methods of multivariate analysis to uncover socio-economic differences among spatial-economics entities. 45th Congress of the European Regional Science Association, 2006.

[21] GOMEZ M. Elementos de Estadística descriptiva. Costa Rica, EUNED. 1997

[22] Introducción a la Estadística Económica por Pérez Rigoberto "et al". España, Departamento de Economía Aplicada Universidad de Oviedo. 2011

[23] Manual de Control Estadístico de Calidad: Teoría y Aplicaciones. Ed. III serie. Verdoy, Pablo "et al". Universat Jaume I. Publicacions , 2006.

[24] Oficina del censo, Junta de Planificación de Puerto Rico, http://www.censo.gobierno.pr/, consultado en enero de 2011.

[25] ÁLVAREZ, R. Estadística Multivariante y no Paramétrica con SPSS. España, Ed. Díaz de Santos, 1995.

[26] POZA, C. Análisis estadístico multivariante por Comunidades Autónomas: diferencias y similitudes. España, Jean Monnet European Studies Center, Universidad Antonio de Nebrija. 2005

[27] ALDENDENFER, M. y BLASHFIELD, R. Cluster Análisis: Quantitative Aplications in the Social Sciences. Newbury Park, California, Sage Publications. 1884

[28] EVERITT, B. y HOTHORN, T. An Introduction to Aplied Multivariate Analysis whit R. New York, Springer, 2011

APÉNDICE

TABLA A1

Pesos de variables sin municipio de Culebra

	Componentes					Componentes			
Variable	1	2	3	4	Variable	1	2	3	4
V1	0.713	0.474	0.053	0.266	V11	0.383	0.683	0.034	0.334
V2	0.03	0.228	0.905	0.092	V12	0.065	-0.07	0.896	0.198
V3	0.764	0.359	0.039	0.146	V13	0.787	0.004	0.037	0.327
V4	0.16	0.89	0.079	0.008	V14	0.161	0.284	0.891	0.024
V5	0.592	0.726	0.2	0.04	V15	0.137	0.212	0.035	0.912
V6	0.717	0.147	0.093	0.509	V16	0.085	0.174	0.085	0.687
V7	0.48	0.824	0.126	0.022	V17	0.071	0.361	0.314	0.814
V8	0.835	0.479	0.063	0.039	V18	0.107	0.046	0.933	0.106
V9	0.719	0.548	0.026	-0.16	V19	0.316	0.735	0.274	0.077
V10	0.589	0.435	0.167	0.149					

Tabla A2

Pesos de variables sin municipio de Guaynabo

	Componentes					Componentes			
Variable	1	2	3	4	Variable	1	2	3	4
V1	0.769	0.054	0.336	0.288	V11	0.424	0.043	0.684	0.312
V2	0.059	0.915	0.18	0.062	V12	0.103	0.876	0.031	0.215
V3	0.746	0.015	0.335	0.168	V13	0.762	-0.04	0.093	0.351
V4	0.244	0.082	0.834	0.016	V14	0.196	0.904	0.226	0.047
V5	0.758	0.233	0.524	0.083	V15	0.096	0.017	0.286	0.898
V6	-0.67	0.138	-0.01	0.399	V16	0.071	0.097	0.177	0.716
V7	0.593	0.119	0.73	0.053	V17	0.078	0.291	0.444	0.778
V8	0.905	0.079	0.298	0.035	V18	0.13	0.927	0.048	0.107
V9	0.801	0.046	0.346	0.184	V19	0.376	0.278	0.608	0.145
V10	0.637	0.181	0.313	0.149					

Tabla A3

Pesos de variables sin municipio de San Juan

	Componentes					Componentes			
Variable	1	2	3	4	Variable	1	2	3	4
V1	0.779	0.057	0.353	0.235	V11	0.55	0.172	0.494	0.345
V2	0.037	0.913	0.177	0.058	V12	0.109	0.868	0.078	0.218
V3	0.782	0.013	0.266	0.182	V13	0.819	0.057	0.153	0.264
V4	0.248	0.044	0.864	0.004	V14	0.185	0.905	0.205	0.059
V5	0.686	0.182	0.632	0.009	V15	0.141	0.021	0.139	0.936
V6	0.632	0.089	0.261	0.387	V16	0.04	0.098	0.138	0.667
V7	0.568	0.078	0.765	0.032	V17	0.049	0.235	-0.23	0.873
V8	0.876	0.079	0.367	0.002	V18	0.131	0.924	0.046	0.113
V9	0.762	0.032	0.419	0.213	V19	0.359	0.214	0.697	0.109
V10	0.652	0.18	0.312	0.173					

Tabla A4

Pesos de variables sin municipio de San Germán

	Componentes					Componentes			
Variable	1	2	3	4	Variable	1	2	3	4
V1	0.752	0.438	0.051	0.265	V11	0.39	0.682	0.044	0.343
V2	0.046	0.209	0.921	0.055	V12	0.084	0.047	0.876	0.231
V3	0.74	0.355	0.022	0.168	V13	0.802	0.039	0.052	0.303
V4	0.207	0.869	0.091	-0.01	V14	0.189	0.272	0.894	0.024
V5	0.626	0.687	0.215	0.015	V15	0.109	0.226	0.02	0.915
V6	0.643	0.153	0.118	0.44	V16	0.084	0.196	0.107	0.674
V7	0.503	0.807	0.122	0.028	V17	0.085	0.373	0.291	0.813
V8	0.854	-0.42	0.085	0.002	V18	0.109	0.035	-0.93	0.121
V9	0.733	0.481	0.053	0.197	V19	0.314	0.728	0.258	0.097
V10	0.609	0.409	0.16	0.161					

TABLA A5

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	7.931	41.744	41.744	7.931	41.744	41.744	4.852	25.535	25.535
2	3.813	20.066	61.809	3.813	20.066	61.809	4.105	21.605	47.139
3	2.196	11.555	73.365	2.196	11.555	73.365	3.544	18.655	65.795
4	1.161	6.111	79.476	1.161	6.111	79.476	2.599	13.681	79.476
5	.805	4.239	83.715						
	⋮	⋮	⋮						

Tabla A6

Cargas de las variables en cada componente

Variable	Componentes 1	2	3	4	Variable	Componentes 1	2	3	4
V1	.843	-.168	.196	-.162	V11	.770	-.019	.297	.233
V2	.364	.782	-.361	-.044	V12	.109	.866	-.205	-.163
V3	.788	-.149	.108	-.224	V13	-.597	.084	-.211	.576
V4	.739	.004	-.073	.507	V14	.473	.684	-.457	-.069
V5	.943	-.015	-.118	.080	V15	.368	.362	.798	-.007
V6	-.472	.495	.350	.219	V16	.058	.339	.564	-.289
V7	.919	-.048	-.064	.264	V17	-.367	-.610	-.592	-.185
V8	-.900	.202	.079	.237	V18	-.092	-.871	.329	.135
V9	.820	-.275	-.236	-.087	V19	.771	.155	-.047	.289
V10	-.649	.401	.093	.031					

Tabla A7

Cargas de las variables en cada componente después de la rotación

	Componentes					Componentes			
Variable	1	2	3	4	Variable	1	2	3	4
V1	.743	.432	-.039	.252	V11	.392	.679	-.043	.345
V2	.052	.211	.908	.069	V12	-.090	-.049	.879	.218
V3	.741	.350	.018	.180	V13	-.799	.042	-.051	-.311
V4	.211	.869	.090	-.010	V14	.182	.265	.894	-.038
V5	.630	.684	.211	.024	V15	.107	.227	.015	.917
V6	-.650	-.151	.127	.420	V16	.082	-.186	.092	.685
V7	.507	.805	.121	.032	V17	.083	-.373	-.284	-.815
V8	-.856	-.417	-.083	-.011	V18	.115	.037	-.932	-.104
V9	.737	.478	.050	-.189	V19	.319	.726	.252	.107
V10	-.614	-.406	.165	.148					

Tabla A8

ANOVA para 2 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor score 1	15.303	1	0.812	76	18.85	0
REGR factor score 2	0.425	1	1.008	76	0.422	0.518
REGR factor score 3	21.804	1	0.726	76	30.023	0
REGR factor score 4	7.116	1	0.92	76	7.738	0.007

Tabla A9

ANOVA para 3 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor score 1	2.812	2	0.952	75	2.955	0.058
REGR factor score 2	20.067	2	0.492	75	40.822	0
REGR factor score 3	24.146	2	0.383	75	63.083	0
REGR factor score 4	2.512	2	0.96	75	2.617	0.08

Tabla A10

ANOVA para 4 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor score 1	3.18	3	0.912	74	3.488	0.02
REGR factor score 2	14.142	3	0.467	74	30.267	0
REGR factor score 3	12.489	3	0.534	74	23.379	0
REGR factor score 4	13.446	3	0.495	74	27.14	0

Tabla A11

ANOVA para 5 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor score 1	11.341	4	0.433	73	26.171	0
REGR factor score 2	10.424	4	0.484	73	21.556	0
REGR factor score 3	11.067	4	0.448	73	24.682	0
REGR factor score 4	8.719	4	0.577	73	15.109	0

Tabla A12

ANOVA para 6 grupos

	Cluster		Error		F	Sig.
	Mean Square		Mean Square			
REGR factor score 1	8.932	5	0.449	72	19.885	0
REGR factor score 2	9.308	5	0.423	72	22.003	0
REGR factor score 3	8.439	5	0.483	72	17.458	0
REGR factor score 4	9.583	5	0.404	72	23.72	0

Tabla A13

ANOVA para 7 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor score 1	7.651	6	0.438	71	17.468	0
REGR factor score 2	7.66	6	0.437	71	17.519	0
REGR factor score 3	7.217	6	0.475	71	15.204	0
REGR factor score 4	8.662	6	0.352	71	24.576	0

Tabla A14

ANOVA para 8 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor sore 1	7.763	7	0.324	70	23.982	0
REGR factor score 2	6.257	7	0.474	70	13.192	0
REGR factor score 3	6.889	7	0.411	70	16.76	0
REGR factor score 4	7.83	7	0.317	70	24.702	0

Tabla A15

ANOVA para 9 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor score 1	6.659	8	0.344	69	19.367	0
REGR factor score 2	6.066	8	0.413	69	14.698	0
REGR factor score 3	6.022	8	0.418	69	14.419	0
REGR factor score 4	7.255	8	0.275	69	26.405	0

Tabla A16

ANOVA para 10 grupos

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
REGR factor	6.375	9	0.289	68	22.083	0
REGR factor	6.247	9	0.306	68	20.446	0
REGR factor	5.862	9	0.357	68	16.441	0
REGR factor	5.956	9	0.344	68	17.307	0

Tabla A17

Medidas internas de validación aplicadas a los 78 municipios

		Número de conglomerados								
		2	3	4	5	6	7	8	9	10
Jerárquicos	Connectivity	5.025	10.633	11.466	11.716	17.977	24.849	37.486	42.877	45.520
	Dunn	0.491	0.400	0.400	0.400	0.207	0.189	0.199	0.199	0.199
	Silhouette	0.518	0.416	0.397	0.366	0.268	0.212	0.242	0.188	0.175
kmeans	Connectivity	16.109	40.543	39.364	54.133	64.412	61.554	62.387	60.489	82.497
	Dunn	0.132	0.067	0.075	0.069	0.077	0.098	0.098	0.127	0.095
	Silhouette	0.315	0.215	0.221	0.233	0.232	0.239	0.233	0.236	0.181
pam	Connectivity	34.292	39.698	46.995	53.635	73.346	71.389	74.236	74.253	75.087
	Dunn	0.067	0.067	0.067	0.078	0.081	0.094	0.107	0.107	0.107
	Silhouette	0.176	0.206	0.185	0.197	0.141	0.168	0.186	0.187	0.181

Tabla A18

Medidas de estabilidad aplicadas a los 78 municipios

		Número de conglomerados								
		2	3	4	5	6	7	8	9	10
Jerarquicos	APN	0.0188	0.0191	0.0912	0.1763	0.2549	0.329	0.2837	0.3728	0.3579
	AD	2.4361	2.3689	2.3438	2.3114	2.1563	2.0927	1.9242	1.8727	1.8148
	ADM	0.1167	0.179	0.3096	0.464	0.7035	0.7773	0.7218	0.8137	0.7926
	FOM	1.0036	1.0073	0.9994	0.9998	0.9839	0.9618	0.9611	0.9512	0.9502
kmeans	APN	0.1727	0.3077	0.3227	0.3472	0.4221	0.3736	0.4507	0.5089	0.4788
	AD	2.4462	2.3102	2.1258	2.0203	1.9707	1.8192	1.8392	1.7797	1.7503
	ADM	0.5781	0.8864	0.8204	0.9461	1.0303	0.8602	1.0283	1.009	1.0295
	FOM	0.9978	1.004	1.0059	1.01	1.0035	0.9967	0.9915	0.9791	0.9802
pam	APN	0.2987	0.4351	0.4177	0.4077	0.4525	0.4675	0.4573	0.4213	0.4417
	AD	2.4174	2.3308	2.1931	2.0682	1.9829	1.9287	1.8253	1.752	1.6991
	ADM	0.6816	0.9592	1.0019	1.0183	0.9796	1.0737	1.0049	0.9704	0.9438
	FOM	1.001	1.0017	1.0019	1.0031	1.0075	0.9951	1.0065	1.0022	1.0037

Tabla A19

Matriz de Correlaciones para las 19 variables

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18	V19
V1	1.00	.07	.68	.59	.77	-.38	.79	-.86	.67	-.47	.61	-.07	-.67	.21	.38	.11	-.27	.08	.50
V2	.07	1.00	.13	.30	.36	.10	.29	-.22	.19	-.01	.14	.72	-.08	.92	.11	.21	-.37	-.76	.38
V3	.68	.13	1.00	.41	.63	-.50	.66	-.71	.62	-.60	.64	.00	-.59	.24	.36	-.03	-.27	.04	.49
V4	.59	.30	.41	1.00	.72	-.22	.87	-.56	.55	-.46	.57	.01	-.18	.34	.19	-.05	-.27	-.02	.61
V5	.77	.36	.63	.72	1.00	-.45	.89	-.87	.82	-.59	.67	.07	-.50	.47	.22	.03	-.26	-.10	.83
V6	-.38	.10	-.50	-.22	-.45	1.00	-.44	.57	-.61	.45	-.24	.20	.22	-.02	.19	.18	-.36	-.23	-.31
V7	.79	.29	.66	.87	.89	-.44	1.00	-.77	.73	-.55	.70	.04	-.41	.39	.27	-.08	-.31	-.05	.74
V8	-.86	-.22	-.71	-.56	-.87	.57	-.77	1.00	-.87	.64	-.59	.04	.63	-.34	-.20	-.05	.10	-.04	-.57
V9	.67	.19	.62	.55	.82	-.61	.73	-.87	1.00	-.61	.57	-.11	-.44	.30	.04	-.09	.03	.07	.54
V10	-.47	-.01	-.60	-.46	-.59	.45	-.55	.64	-.61	1.00	-.49	.23	.41	-.10	.01	.10	-.03	-.27	-.41
V11	.61	.14	.64	.57	.67	-.24	.70	-.59	.57	-.49	1.00	-.01	-.38	.23	.53	.02	-.50	.04	.56
V12	-.07	.72	.00	.01	.07	.20	.04	.04	-.11	.23	-.01	1.00	-.03	.66	.24	.15	-.41	-.89	.19
V13	-.67	-.08	-.59	-.18	-.50	.22	-.41	.63	-.44	.41	-.38	-.03	1.00	-.18	-.31	-.12	.20	-.02	-.29
V14	.21	.92	.24	.34	.47	-.02	.39	-.34	.30	-.10	.23	.66	-.18	1.00	.03	.04	-.32	-.76	.45
V15	.38	.11	.36	.19	.22	.19	.27	-.20	.04	.01	.53	.24	-.31	.03	1.00	.55	-.81	-.12	.29
V16	.11	.21	-.03	-.05	.03	.18	-.08	-.05	-.09	.10	.02	.15	-.12	.04	.55	1.00	-.41	-.07	.04
V17	-.27	-.37	-.27	-.27	-.26	-.36	-.31	.10	.03	-.03	-.50	-.41	.20	-.32	-.81	-.41	1.00	.35	-.40
V18	.08	-.76	.04	-.02	-.10	-.23	-.05	-.04	.07	-.27	.04	-.89	-.02	-.76	-.12	-.07	.35	1.00	-.13
V19	.50	.38	.49	.61	.83	-.31	.74	-.57	.54	-.41	.56	.19	-.29	.45	.29	.04	-.40	-.13	1.00