ANALISIS DE CONGLOMERADOS

Dr. Cs. Gustavo Ruiz Aranibar[1]

✉ *ruizaranibargustavo@gmail.com.bo*

RESUMEN

El análisis de conglomerados implica agrupar objetos, sujetos o variables, con características similares en grupos. La semejanza o disimilitud de los objetos se mide por un índice particular de asociación. Se consideran los tipos de métodos que agrupan variables basadas en la estructura de correlación de variables.

En algunos estudios geológicos es conveniente agrupar muestras similares en las que se han realizado muchas mediciones y medir el grado de similitud entre los grupos. Utilizando el coeficiente de correlación o la función de distancia, la matriz resultante suele ser demasiado grande para la interpretación directa. El análisis de conglomerados, es una técnica desarrollada por sicólogos, es un método de búsqueda de relaciones en una gran matriz simétrica. Las variables o grupos de variables especificados pueden usarse entonces para agrupar las muestras por función de distancia.

PALABRAS CLAVE

Coeficiente de correlación, dendograma, distancia, taxonomía, similitud.

ABSTRACT

Cluster analysis involves grouping objects, subjects or variables, with similar characteristics into groups. Similarity or dissimilarity of objects is measured by a particular index of association. Types of methods that cluster variables based on correlation structure of variables.

In some geologic studies it is desirable to group together similar samples on which many measurements have been made, and to measure the degree of similarity between the groups. Using either a coefficient, the matching coefficient, or the distance function, the resulting matrix is usually too large for direct interpretation. Cluster analysis, a technique developed by psychologists, is a method of searching for relationships in a large symmetrical matrix. Specified variables or groups of variables can then be used in clustering the samples by distance function.

KEYWORDS

Correlation coefficient, dendogram, distance, taxonomy, similarity.

1. INTRODUCCIÓN.

El Análisis de Conglomerados (AC) también conocido como Cluster Analysis o Taxonomía Numérica, es una técnica estadística multivariable, cuya finalidad es dividir un conjunto de objetos en grupos de forma que los objetos en un mismo grupo sean muy similares entre sí (cohesión interna del grupo) y de los objetos de conglomerados diferentes sean distintos (aislamiento externo del grupo); éste permite agrupar los elementos o variables de un archivo de datos en función del parecido o similitud existente entre ellos, buscando agrupar elementos (o variables) y tratando de lograr la máxima homogeneidad entre los grupos y la mayor diferencia entre los ellos, es una técnica descriptiva, teórica y no inferencial.

[1] Se agradece a la UAGRM por la beca otorgada con fondos del IDH, para cursar y culminar exitosamente el Doctorado en Ciencias de la Educación Superior. Especializado en Estadística e Informática

El AC permite clasificar las unidades de análisis en grupos homogéneos de tal manera que las unidades pertenecientes a uno de los grupos o conglomerados serán lo más parecidas entre sí, aunque muy diferentes respecto a los otros grupos o dicho de otra manera es la tarea de agrupar un conjunto de objetos de tal manera que los miembros del mismo grupo sean más similares, en algún sentido u otro, siendo la tarea principal de la minería de datos exploratorios, técnica común en el análisis de datos estadísticos.

2. OBJETIVOS DEL ANÁLISIS DE CONGLOMERADOS.

El principal objetivo es agrupar objetos (personas, empresas, productos, etc.) en conglomerados, de forma que cada objeto es muy parecido a los que hay en el conglomerado con respecto a algún criterio de selección predeterminado.

Se expone las dos decisiones sobre las que se apoya esta técnica de análisis:

1. Elección de una medida de proximidad entre los individuos.

2. Elección de un criterio a partir del cual agrupar a los individuos o unidades de análisis (secciones censales, países, ciudades,etc.) en los conglomerados.

Para lo cual se debe:

- Plantear el problema a resolver por un AC.

  En el planteamiento del problema de conglomerados se debe considerar las variables en la que se basara el agrupamiento. Haciendo notar que la inclusión de una o más variables irrelevantes distorsiona una solución de agrupamiento, que podría ser útil o no.

  En esencia, el conjunto de las variables elegidas debe describir la semejanza entre los objetos en términos relevantes para el problema de investigación; en la investigación exploratoria, el investigador debe valerse de su juicio e intuición, se aconseja utilizar para los conglomerados que se utilicen un número de muestras mayores a 100.

- Establecer medidas de semejanza y de distancia entre los objetos a clasificar en función del tipo de datos analizados.

- Analizar algunos de los métodos de clasificación propuestos en la literatura, debido a la existencia de diferentes métodos tales como los jerárquicos aglomerativos, el algoritmo de las k-medias, y otros, que permiten determinar el número de grupos.

- Interpretar los resultados obtenidos.

  Como técnica de agrupación de variables, el AC es similar al análisis factorial; pero, mientras que la factorización es más bien poco flexible en algunos de sus supuestos (linealidad, normalidad, variables cuantitativas, etc.) y siempre estima de la misma manera la matriz de distancias, la aglomeración es menos restrictiva en sus supuestos (no exige linealidad, ni simetría, permite variables categóricas, etc.) y admite varios métodos de estimación de la matriz de distancias.

3. CONCEPTOS GENERALES DEL AC.

La taxonomía es la ciencia de la clasificación de los seres, elementos de una de las ciencias naturales, las describe, denomina y clasifica ordenadamente atendiendo a sus afinidades y relaciones.

El AC tuvo su origen cuando se utilizó en antropología por Driver y Kroeber en 1932 e introducido a la sicología por Zubin en 1938 y Robert Tryon en 1939, fue utilizado por Cattell en 1943 para la clasificación de la personalidad sicológica basada en teoría de rasgos.

El AC es una técnica usada para clasificar objetos o casos en grupos relativamente homogéneos llamados conglomerados. Los objetos de cada conglomerado tienden a ser similares entre sí y diferentes de los objetos de otros conglomerados. Como técnica de agrupación de casos, el AC es similar al análisis discriminante. Sin embargo, mientras que el análisis discriminante efectúa la clasificación tomando como referencia un criterio o variable dependiente (los grupos de clasificación), el AC permite detectar el número óptimo de grupos y su composición únicamente a partir de la similitud existente entre los casos; además, el AC no asume ninguna distribución específica para las variables.

Tanto el AC como el análisis discriminante se interesan en la clasificación. Sin embargo, el análisis discriminante requiere de un conocimiento previo del conglomerado o la pertenencia al grupo de cada objeto o caso incluido, para desarrollar la regla de clasificación.

En el AC, todo un conjunto de relaciones interdependientes, no distingue entre variables dependientes e independientes, sino que examina las relaciones interdependientes entre el conjunto completo de variables. Los objetos en un grupo son relativamente similares en términos de estas variables y diferentes de los objetos de otros grupos. Cuando se usa de esta manera, el AC es la contrapartida del análisis factorial, ya que no reduce el número de variables sino de objetos, a los que agrupa en un número mucho menor de conglomerados.

El método jerárquico es idóneo para determinar el número óptimo de conglomerados existente en los datos y el contenido de los mismos. El método de K medias permite procesar un número ilimitado de casos, pero sólo permite utilizar un método de aglomeración, y es este método, el que se describirá y aplicará en el presente trabajo.

Existe un notable contraste entre el AC con el análisis de varianza, la regresión, el análisis discriminante y el análisis factorial; los cuales se fundamentan en un razonamiento estadístico amplio. Aunque muchos de los procedimientos de conglomeración tienen propiedades estadísticas importantes, debe reconocerse fundamentalmente su sencillez.

Los siguientes estadísticos y conceptos se asocian con el AC.

- Calendario de aglomeración: este programa brinda información sobre objetos o casos que se combinan en cada etapa del proceso de conglomeración jerárquica.

- Centroide del conglomerado: es la media de los valores de las variables de todos los objetos o casos de un conglomerado particular.

- Centros del conglomerado: son los puntos de partida en la conglomeración no jerárquica. Los conglomerados se construyen en torno a estos centros.

- Pertenencia al conglomerado: indica el conglomerado al que corresponde cada objeto o caso.

- Dendograma: conocido como gráfica de árbol, es un medio gráfico para presentar los resultados de la conglomeración. Las líneas verticales representan conglomerados que están unidos. La posición de la línea en la escala, indica las distancias en las que se unen los conglomerados. El diagrama de árbol muestra los grupos que se forman al crear conglomerados de observaciones en cada paso y sus niveles de similitud. El nivel de similitud se mide en el eje vertical (alternativamente se puede mostrar el nivel de distancia) y las diferentes observaciones se especifican en el eje horizontal, como se observará en su aplicación.

Distancias entre los centros de los conglomerados: estas distancias indican cuán separados están los pares individuales de conglomerados. Los que están muy separados son distintos y, por lo tanto, son deseables. El objetivo principal del AC es definir la estructura de los datos colocando las observaciones más parecidas de los datos, en conglomerados de acuerdo a las distancias obtenidas de la matriz de distancia.

4. ANÁLISIS DE CONGLOMERADOS.

La clasificación es la colocación de objetos en grupos más o menos homogéneos, de tal manera que se revela la relación entre grupos. Este es el punto fuerte en especial de los taxonomistas, que intentan deducir el linaje de las criaturas vivas de sus características y semejanzas. La taxonomía es altamente subjetiva y depende de las habilidades de los taxonomistas, desarrollados a través de años de experiencia. En este sentido, el campo es análogo en muchos aspectos a la geología. En geología, un grupo de investigadores se han vuelto insatisfechos con la subjetividad y buscan nuevas técnicas de clasificación que incorporen las capacidades masivas de manejo de información creando una base de datos en la computadora. Estos trabajadores se llaman taxonomistas numéricos y son responsables de muchos de los avances en la clasificación numérica. En el AC se tiene que:

a. Los conglomerados resultantes deben mostrar un alto grado de homogeneidad interna (dentro del conglomerado) y un alto grado de heterogeneidad externa (entre conglomerados).

b. Gráficamente, los objetos dentro de los conglomerados estarán muy próximos, y los diferentes conglomerados muy alejados.

c. El AC permite la inclusión de múltiples variables para llevar a cabo la agrupación de objetos.

5. FUNDAMENTOS TEÓRICOS PARA EL AC

En la actualidad, la taxonomía numérica es el centro de una controversia entre los biólogos, al igual que el acrimonioso debate entre los sicólogos que se arremolinaron alrededor del análisis factorial en los años 1930 y 1940. Como en esa disputa, las técnicas de la taxonomía numérica han sido exageradamente promovidas por

algunos practicantes. Además, afirmaron que una taxonomía numéricamente derivada representaría mejor la filogenia de un grupo de organismos que cualquier otro tipo de clasificación. Esto, por supuesto, no puede ser demostrado. En la actualidad, los fundamentos teóricos del AC son incompletos, se sabe poco de las propiedades estadísticas de los métodos taxonómicos numéricos y no se dispone de pruebas de significación. Muchos de los métodos de taxonomía numérica son importantes en la investigación geológica, especialmente en la clasificación de los invertebrados fósiles y el estudio de la paleoecología.

Si se tiene una colección de objetos que desea organizar en una clasificación jerárquica como en biología, estos objetos se denominan "unidades taxonómicas operativas". En cada objeto, se realiza una serie de medidas que constituyen el conjunto de datos.

Teniéndose n objetos, con medidas de m características, el conjunto de datos forma una matriz de observaciones. A continuación, se calculará una medida de semejanza o similitud entre cada par de objetos. Se han utilizado los coeficientes de semejanza, como ser el coeficiente de correlación , y la distancia euclidiana estandarizada que es el coeficiente de distancia, se observara que, por definición, los elementos de la diagonal principal de esta matriz son nulos.

Las fórmulas que se utilizan para el AC son:

$$X_{ij} \quad \forall\ i = 1,\ldots,n \ \ y \ \ j = 1,\ldots,m$$

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n} - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}$$

$$Z_i = \frac{X_i - \overline{X}}{S}$$

$$d_{ij} = \sqrt{\frac{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}{m}}$$

$$r_{ij} = \frac{\operatorname{cov}(x, y)}{s_x s_y}$$

$$r_{ij} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2} * \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$-1 \le r_{ij} \le 1$$

Dónde:

n = número de observaciones.

m = número de variables de cada observación.

X_{nm} = matriz de observaciones.

X_{mn} = matriz transpuesta de observaciones.

X_{ik} = k-ésima medida en el objeto i.

X_{jk} = k-ésima medida sobre el objeto j.

S = desviación estándar.

Z_i = variable normalizada.

d_{ij} = distancia entre el objeto i y el objeto j.

r_{ij} = coeficiente de correlación entre dos columnas i y j.

r_{mm} = matriz de coeficientes de correlación.

6. MEDIDAS DE DISTANCIA.

Las distancias entre los centros de los conglomerados, indican que cuánto más separados están los pares individuales de conglomerados y son distintos, por lo tanto, son deseables. En la medición de las distancias

se utilizan diferentes medidas, especialmente medidas para variables cuantitativas, las más utilizadas son:

1. Distancia euclídeana y distancia euclídea al cuadrado
2. Distancia métrica de Chebychev
3. Distancia de Manhattan
4. Distancia de Minkowski
5. Distancia de Mahalanobis

Todas estas distancias no son invariantes a cambios de escala por lo que se aconseja estandarizar los datos si las unidades de medida de las variables no son comparables.

7. MÉTODOS DE CLASIFICACIÓN.

Entre los muchos tipos de métodos que existen cabe destacar los siguientes:

- Repartición: tienen un número de grupos, g fijado de antemano, como objetivo y agrupa los objetos para obtener los g grupos. Comienzan con una solución inicial y los objetos se reagrupan de acuerdo con algún criterio de optimalidad.
- Métodos tipo Q: son similares al análisis factorial y utilizan como información la matriz XX' utilizando las variables como objetos y los objetos como variables.
- Procedimientos de localización de modas: agrupan los objetos en torno a modas con el fin de obtener zonas de gran densidad de objetos separados unos de otros por zonas de poca densidad.
- Métodos que permiten solapamiento: permiten que los grupos tengan elementos en común.
- Método de Ward, tiene tendencia a formar conglomerados más compactos y de igual tamaño.

EL Método jerárquico se caracteriza porque en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo, este método es utilizado en el ejemplo de aplicación. Es un método aglomerativo, que comienza con n conglomerados de un objeto. En cada paso del algoritmo se recalculan las distancias entre los grupos existentes y se unen los dos grupos más similares o menos disimilares. El algoritmo acaba con un conglomerado conteniendo todos los elementos. Para determinar qué grupos se unen o se dividen, se utiliza una función objetivo o criterio que, en el caso de los métodos aglomerativos recibe el nombre de enlace.

8. INTERPRETACIÓN DE RESULTADOS

Una distancia baja indica que los dos objetos son similares o están muy cerca o juntos, una gran distancia indica disimilitud. Comúnmente, la matriz de datos originales se estandariza antes de calcular las mediciones de distancias, entonces estos nuevos valores tienen un promedio nulo y una desviación estándar la unidad, esto asegura que cada variable es ponderada igualmente, de lo contrario, la distancia será influenciada más fuertemente por la variable que tiene la mayor magnitud.

Se han desarrollado varias técnicas de agrupamiento, en este trabajo se desarrollará una técnica simple de agrupamiento, llamado el método ponderado par-grupo con promedios aritméticos, luego se señalará algunas modificaciones útiles a este esquema.

9. CONSTRUCCIÓN DE UN DENDOGRAMA.

El dendograma ó gráfico en forma de árbol,

es una herramienta visual para ayudar a decidir el número de conglomerados que podrían representar mejor la estructura de los datos. Se utiliza el dendograma para observar cómo se forman los conglomerados en cada paso y para evaluar los niveles de similitud (o distancia) de los conglomerados que se forman.

La decisión acerca de la agrupación final, se la conoce como cortar el dendograma. Cortar el dendograma es similar a trazar una línea vertical a lo largo del dendograma para especificar la agrupación final, si el dendograma está orientado horizontalmente mediante una línea horizontal sucede lo contrario, también se pueden comparar agrupaciones finales en los dendogramas para determinar cuál de ellas tiene más sentido para los datos. En la determinación del número final de conglomerados a formar o regla de parada, no hay un procedimiento determinado, decidiéndolo el investigador en la fase de interpretación de los datos.

Para una comprensión en la construcción de un dendograma, teniéndose una matriz de correlaciones, que es simétrica de los coeficientes de similitud, entre seis objetos supuestos, identificados como A,B,..,F, para las filas y para las columnas, como se muestra a continuación:

Tabla Nº 1
Matriz de correlaciones de seis variables

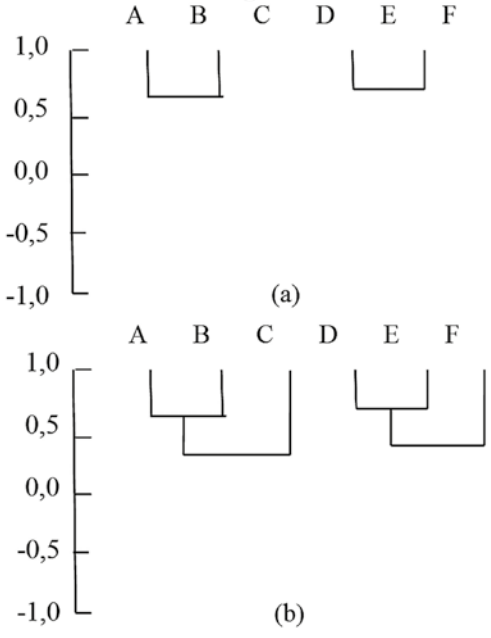
A	B	C	D	E	F
1,00	(0,57)	0,12	-0,65	-0,62	-0,39
(0,57)	1,00	(0,46)	-0,79	-0,72	-0,72
0,12	0,46	1,00	-0,58	-0,61	-0,52
-0,65	-0,79	-0,58	1,00	(0,66)	(0,41)
-0,62	-0,72	-0,61	(0,66)	1,00	0,40
-0,39	-0,72	-0,52	0,41	0,40	1,00

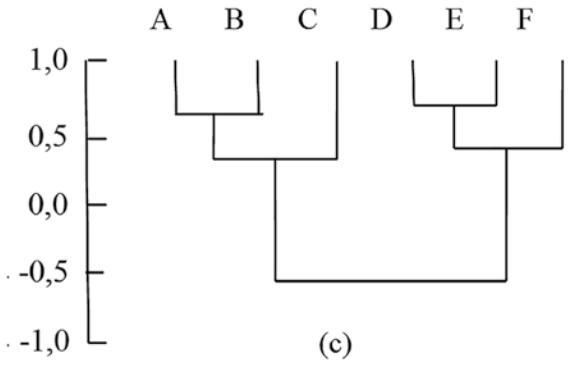
Fuente: Obtienido de Davis C. John. Statistics and Data Analysis in Geology.

El primer paso en la agrupación por el método de grupo de pares, es encontrar mutuamente las correlaciones más altas en la matriz para formar los centros de los conglomerados. La más alta correlación en cada columna de la matriz de la tabla 1, se muestra entre paréntesis. Los objetos A y B forman parejas mutuamente altas, porque A se asemeja mucho a B y B se asemeja más a A. Sin embargo, C y B no forman una pareja mutuamente alta, porque aunque C se asemeja mucho a B, B se parece más a A que a C. Para calificarlo como un par mutuamente alto, el coeficiente r_{ij} debe ser el coeficiente más alto en sus respectivas columnas.

Se puede indicar la semejanza entre los pares mutuamente altos en un diagrama como el de la Fig. 1. En la gráfica (a). el objeto A está conectado a B a un nivel de 0,57, indicando el grado de su similitud mutua. De la misma manera, D y E están conectados. Este es el primer paso en la construcción de un dendograma, o diagrama de árbol, que es la forma más común de mostrar los resultados de la agrupación.

Figura Nº 1
(a) Dendograma con grupos iníciales.
(b) Conexión de los objetos restantes a los grupos
(c) Conexión final de dos grupos, completando el dendograma





Fuente: Elaboración del autor

A continuación, la similitud de la matriz debe recalcularse, tratando a los elementos agrupados como un solo elemento. Hay varios métodos para hacer esto. La técnica simple que se está considerando, las nuevas correlaciones entre todos los grupos y los objetos no agrupados se recalculan mediante el cálculo aritmético simple. Así, a nueva correlación entre el grupo AB y el objeto C es igual a la suma de las correlaciones de los elementos comunes a AB y C, dividido entre 2. La tabla 2, contiene los resultados de estos nuevos cálculos.

Tabla Nº 2
Matriz de correlaciones promedio entre dos conglomerados y dos variables

AB	C	DE	F
1,00	(0,29)	-0,70	-0,55
(0,29)	1,00	-0,59	-0,52
-0,70	-0,59	1,00	(0,41)
-0,55	-0,52	(0,41)	1,00

Fuente: Elaboración del autor

El procedimiento de agrupamiento se repite; los pares mutuamente altos son buscados y agrupados. En este ciclo, el objeto C se une al grupo AB y el objeto F se une al grupo de (Fig. 1. b). Entonces el proceso continúa hasta que todos los racimos se unan. La matriz final de similitudes será una matriz de 2x2 entre los dos agrupamientos restantes, como se muestra en la tabla 3. Esto indica que el grupo ABC tiene una semejanza de -0,59 con el grupo DEF. Así el dendrograma puede ser completado. (Fig. 1. c)

Tabla Nº 3
Matriz de correlaciones promedio entre dos grupos

ABC	DEF
1,00	-0,59
0,59	1,00

Fuente: Elaboración del autor

Las características esenciales de este método de análisis de conglomerados pueden resumirse en la forma siguiente:

a. El coeficiente de correlación se utiliza como medida de similitud.
b. Las similitudes más altas se agrupan en primer lugar.
c. Dos objetos sólo pueden conectarse si tienen correlaciones mutuamente más altas entre sí.
d. Después de que se agrupan dos objetos, se promedian sus correlaciones con todos los demás objetos.

Una modificación obvia de este esquema es incorporar alguna otra medida de similitud. Aunque se han propuesto muchas medidas, sólo dos se utilizan ampliamente; el coeficiente de correlación y el coeficiente de distancia.

Como era de esperar, una distancia baja indica que los dos objetos son similares, o "estar cerca o juntos", ya que una gran distancia indica disimilitud. Comúnmente, la matriz de datos originales se estandarizan antes de calcular mediciones de distancia. Esto asegura que cada variable es ponderada igualmente, de lo contrario, la distancia será influenciada más fuertemente por la variable que tiene la mayor magnitud. Así por ejemplo, se puede medir tres ejes perpendiculares sobre una colección de muestras. Si se mide dos de los

ejes en centímetros y el tercero en milímetros, el tercer eje tendrá proporcionalmente diez veces la influencia sobre el coeficiente de distancia de las otras dos variables.

Se han desarrollado varias técnicas de agrupamiento: una consideración de todas las posibles variaciones y sus méritos relativos que están fuera del alcance de este trabajo, recomendándose el texto Benzécri J. P. señalada en la bibliografía para mayor conocimiento. Se describe la técnica de agrupamiento, llamado el método ponderado par-grupo con promedios aritméticos.

10. CONSIDERACIONES ENTRE EL ANÁLISIS DE CONGLOMERADOS, EL ANÁLISIS DE COMPONENTES PRINCIPALES Y EL ANÁLISIS FACTORIAL.

El AC es una técnica estadística clasificadora, pero, en realidad, es una técnica que, como el ACP o como el AF, pretenden representar una realidad en la que no se consigue visualizar, una realidad cuya representación original es multidimensional y es imposible que la podamos ver en su estado puro.

En realidad tanto el ACP, el AF y el AC son técnicas que tratan de representar una nube de puntos originales situada en un espacio de tantas dimensiones que es imposible visualizar. Y cada una de ellas, también, pueden ser usadas como métodos clasificatorios, como métodos para crear subpoblaciones, subgrupos o subsubgrupos.

Las diferencias fundamentales entre ellas es la forma de presentación que utilizan y la forma de resolver el problema de no visualización de la nube de puntos originales. El ACP y el AF construyen una nube de puntos de la misma naturaleza pero en menor número de dimensiones perdiendo una parte de la información original. En cambio, el AC crea una representación distinta a la de la nube de puntos. Crea otro tipo de representación, cambia la forma, no lo hace mediante una nube de puntos, lo hace mediante un dendograma, pero cada una de estas opciones tiene sus ventajas y sus desventajas.

El ACP y el AF respetan el tipo de representación de una nube de puntos, pero al reducir dimensiones se pierde información y esto es un problema, especialmente si la pérdida es importante. El AC respeta la nube de puntos originales, no reduce dimensiones y no se pierde información, pero sí, se cambia el mecanismo de representación. Esta se representa mediante un dendograma. Se puede decir que en el ACP y el AF se hace una representación figurativa y en el AC se hace una representación abstracta.

En el AC se define una noción de distancia entre puntos, se necesita elegir una distancia, una medida que cuantifique distancias entre los individuos dentro de la nube de puntos originales, y aquí aparece el primer problema del AC, porque existen muchas distancias propuestas.

En el AC la distancia euclídea es la más utilizada, que calcula la distancia en línea recta entre los puntos en el espacio o en el hiperespacio de la nube de puntos originales, siendo esta distancia en realidad una aplicación del teorema de Pitágoras.

La distancia Mahalanobis es de mucho prestigio en estadística, se trata de una distancia que toma en cuenta las distancias que hay entre cada una de las variables y las relativiza respecto a la dispersión que tiene cada una de estas variables originales.

Estos temas puede el lector profundizar, consultando la bibliografía señalada en las revistas Varianza Nº 10, 12 y 15, del IETA, Carrera de Estadística - FCPN de la UMSA.

11. TRABAJO COMPUTACIONAL

El AC de un pequeño conjunto de datos es relativamente simple, se vuelve arduo cuando (n) el tamaño de la muestra es grande, como también (m) el número de variables. Además, las rutinas gráficas para construir dendogramas se vuelven muy complejas, por estas razones, se desarrollo el programa computacional utilizado (8), que permite calcular en base a la matriz de observaciones, la matriz estandarizada, la matriz de distancias, la agrupación media ponderada de grupos por pares y la construcción del dendograma; también, se tiene la opción de realizar estos cálculos determinando la matriz de correlación, haciéndose notar que los resultados de la agrupación media ponderada de grupos por pares y la construcción del dendograma son diferentes en estos dos casos. El dendograma resultante con el programa computacional es un diagrama de árbol que muestra los grupos que se forman al crear conglomerados de variables en cada paso y sus niveles de similitud.

11. CONCLUSIONES

Una característica de este trabajo, es que necesita poca explicación. Las muestras observadas con sus respectivas variables constituyen la matriz de observaciones, las cuales son estandarizadas, cuando las variables medidas no son directamente comparables, cuyos valores son independientes en las unidades empleadas, son adimensionales, se caracterizan por tener una media igual a cero y una desviación estándar igual a la unidad, de esta manera se permite comparar los datos procedentes de diferentes muestras o poblaciones. El coeficiente de correlación indica mayor similitud a valores absolutos altos, mientras que el coeficiente de distancia indica mayor similitud a la menor distancia. Por lo tanto, las correlaciones deben estar vinculadas o conectadas a un valor alto, y los coeficientes de distancia deben estar unidos a valores bajos.

La matriz de distancia parece ser menos susceptible a cambios drásticos entre los diferentes métodos de agrupamiento. Sin embargo, no hay pruebas estadísticas disponibles para este método de agrupación, ni se ha desarrollado ninguna teoría estadística y aplicada.

En el AC, se puede conglomerar variables que permitan identificar grupos homogéneos, en este caso, las unidades usadas para el análisis son las variables y las medidas de distancia se calculan para todos los pares de variables de acuerdo a los valores de la matriz de coeficientes de correlación, cuyos valores se usan como medida de semejanza (lo opuesto a la distancia) entre las variables.

La conglomeración jerárquica de variables ayuda a identificar variables que hacen una contribución única de datos, la conglomeración también puede usarse para reducir el número de variables en el conglomerado, llamada componente del conglomerado; a menudo se reemplaza un conjunto grande de variables con un conjunto de componentes de conglomerados con poca pérdida de información. Se conglomeran las variables, porque es más sencillo interpretar los componentes conglomerados que en el análisis de componentes principales. El gran beneficio del AC es que proporciona una forma de clasificar los objetos que es relativamente simple y directo, y presenta los

resultados de una manera familiar y fácil de entender.

Luego de obtener los resultados, se concluye que los métodos multivariantes del AC y el análisis factorial, ayudan a reducir la información proporcionada. De esta manera facilitar la toma de decisiones en diferentes estudios, permitiendo analizar fácilmente una serie de variables agrupándolas para poder simplificar los estudios de mercado, investigación de productos, publicidad, estudios sobre precios, etc.

12. APLICACIONES DEL ANÁLISIS DE CONGLOMERADOS

- Segmentación del mercado: por ejemplo, puede agruparse a los consumidores según los beneficios que buscan en la compra de un producto. Cada conglomerado estaría formado por consumidores que son relativamente homogéneos en términos de los beneficios que buscan. Este procedimiento se conoce como segmentación por beneficios.

- Entender la conducta de los compradores: el AC puede usarse para identificar grupos homogéneos de compradores. Luego se examina por separado la conducta de compras de cada grupo. El AC también se ha empleado para identificar las estrategias que usan los compradores de automóviles cuando buscan información externa.

- Identificar oportunidades de nuevos productos: al agrupar marcas y productos, es posible determinar conjuntos competitivos dentro del mercado. Las marcas del mismo conglomerado compiten mucho más entre sí que con las marcas de otros conglomerados. Una empresa puede comparar sus ofertas actuales con las de sus competidores para identificar posibles oportunidades de productos nuevos.

- Elegir mercados de prueba: al agrupar ciudades en conglomerados homogéneos, es posible elegir ciudades comparables para probar diversas estrategias de marketing y poder clasificar a los consumidores.

- Reducir los datos: el AC es útil como herramienta general de reducción de datos, para desarrollar conglomerados o subgrupos de datos que sean más fáciles de manejar que las observaciones individuales. El análisis multivariado posterior no se realiza en las observaciones individuales, sino en los conglomerados. Por ejemplo, para describir las diferencias en la conducta de uso del producto por parte de los consumidores, primero se dividiría a éstos en conglomerados luego, las diferencias entre los grupos se examinaría con el análisis discriminante múltiple.

En Geología, Minas y Metalurgia: se usa el AC para clasificar los minerales por su tamaño, pureza, explotación, etc., para formar grupos de pixels en imágenes digitalizadas enviadas por un satélite desde un planeta a la tierra para identificar los terrenos.

En Odontología, Medicina y Bioquímica: se usa el AC para clasificar las diferentes clases de dientes, seres vivos con los mismos síntomas y características patológicas, remedios, tipos de enfermedades, compuestos químicos, etc.

En Agronomía: para comparar los productos agrícolas ya sea de una misma o de diferentes

especies, tanto de semillas como del producto cosechado, rendimiento de producción, etc.

En la taxonomía: el AC se utiliza para agrupar especies naturales.

Problema. En base a la fuente de datos obtenida para fines ilustrativos, se tienen los datos del análisis petrográfico de veinte muestras mineralógicas con ocho variables cada una, con contenido de óxidos en rocas ígneas, se desea obtener la matriz estandarizada, la matriz de distancias (o también la matriz de correlación), la agrupación media ponderada de grupos por pares, el dendograma tanto para las muestras mineralógicas como para las variables.

Datos:

Cuadro Nº 1

Matriz de observaciones de 20 muestras con 8 variables

Nombre de la muestra	Nº de muestra	SiO_2	Al_2O_3	Fe_2O_3	FeO	MgO	CaO	Na_2O	K_2O
		1	2	3	4	5	6	7	8
Sienita	1	61,7	15,1	2	2,3	3,7	4,6	4,4	4,5
Sienita	2	58,3	17,9	3,2	1,7	1,5	3,7	5,9	5,3
Sienita	3	51,2	17,6	3,5	4,3	3,2	4,5	5,7	4,4
Monzonita	4	54,4	14,3	3,3	4,1	6,1	7,7	3,4	4,2
Diorita	5	58	15,7	0,7	2,8	5	10,9	3	3,2
Diorita	6	46,9	15,9	2,9	10	7	9,6	2,7	0,7
Diorita	7	58	*17,3*	2,2	3,8	2,2	4,3	4,3	4,1
Cuarzo diorita	8	55,5	16,5	2,7	4,6	6,7	6,7	3,2	2,5
Gabro	9	55,4	15,3	2,7	5,5	5,8	9,9	2,9	1,5
Gabro	10	55,9	13,5	2,7	5,9	6,5	8,9	2,4	1,7
Norita	11	47,2	14,5	1,6	13,8	5,2	8,1	3,1	1,2
Norita	12	48,2	18,3	1,3	6,1	10,8	9,4	1,3	0,7
Hiperesteno gabro	13	44,8	18,8	2,2	4,7	11,3	14,6	0,9	0,1
Hiperesteno gabro	14	47	14,1	0,8	15	16	2,3	0,4	1,7
Sienita	15	59,8	17,3	3,6	1,6	1,2	3,8	5	5,1
Cuarzo sienita	16	66,2	16,2	2	0,2	0,8	1,3	6,5	5,8
Sienita alterada	17	50	9,9	3,5	5	11,9	8,3	2,4	5
Monzonita	18	57,4	18,5	3,7	2,1	1,7	6,8	4,5	3,7
Monzonita	19	59,8	15,8	3,8	3,3	2,2	3,9	3	4,4
Diabase	20	52,2	18,2	3,3	4,4	4,7	6,5	4,6	1,9

Fuente: Elaboración Propia

Solución:

Cuadro Nº 2
Matriz de observaciones estandarizadas

	1	2	3	4	5	6	7	8
1	1,10	0,91	0,72	0,36	0,52	0,60	1,12	1,24
2	1,04	1,08	1,15	0,26	0,21	0,48	1,50	1,46
3	0,91	1,06	1,26	0,67	0,45	0,58	1,45	1,21
4	0,97	0,86	1,19	0,64	0,86	1,00	0,87	1,16
5	1,03	0,95	0,25	0,43	0,70	1,42	0,76	0,88
6	0,84	0,96	1,04	1,55	0,98	1,25	0,69	0,19
7	1,03	1,04	0,79	0,59	0,31	0,56	1,09	1,13
8	0,99	0,99	0,61	0,71	0,94	0,87	0,81	0,69
9	0,99	0,92	0,97	0,85	0,81	1,29	0,74	0,41
10	1,00	0,81	0,97	0,92	0,91	1,16	0,61	0,47
11	0,84	0,87	0,58	2,14	0,73	1,05	0,79	0,33
12	0,86	1,10	0,47	0,95	1,52	1,22	0,33	0,19
13	0,80	1,13	0,79	0,73	1,59	1,90	0,23	0,03
14	0,84	0,85	0,29	2,33	2,25	0,30	0,10	0,47
15	1,07	1,04	1,30	0,25	0,17	0,49	1,27	1,40
16	1,18	0,98	0,72	0,03	0,11	0,17	1,65	1,60
17	0,89	0,60	1,26	0,78	1,67	1,00	0,61	1,38
18	1,02	1,12	1,33	0,33	0,24	0,88	1,15	1,02
19	1,07	0,95	1,37	0,51	0,31	0,51	0,76	1,21
20	0,93	1,10	1,19	0,68	0,66	0,84	1,17	0,52

Fuente: Elaboración Propia

Cuadro Nº 3
Matriz de distancias

	1	2	3	4	5	6	7	8	9	10
1	0,000	0,258	0,265	0,288	0,385	0,664	0,132	0,318	0,463	0,463
2	0,258	0,000	0,201	0,415	0,597	0,805	0,257	0,531	0,622	0,638
3	0,265	0,201	0,000	0,303	0,552	0,632	0,222	0,423	0,489	0,502
4	0,288	0,415	0,303	0,000	0,389	0,490	0,305	0,274	0,306	0,295
5	0,385	0,597	0,552	0,389	0,000	0,559	0,416	0,275	0,344	0,367
6	0,664	0,805	0,632	0,490	0,559	0,000	0,613	0,406	0,273	0,263
7	0,132	0,257	0,222	0,305	0,416	0,613	0,000	0,320	0,439	0,445
8	0,318	0,531	0,423	0,274	0,275	0,406	0,320	0,000	0,230	0,217
9	0,463	0,622	0,489	0,306	0,344	0,273	0,439	0,230	0,000	0,088
10	0,463	0,638	0,502	0,295	0,367	0,263	0,445	0,217	0,088	0,000
11	0,747	0,889	0,724	0,649	0,661	0,297	0,678	0,534	0,489	0,472

12	0,671	0,880	0,754	0,546	0,465	0,376	0,673	0,360	0,358	0,332
13	0,819	0,998	0,877	0,641	0,559	0,474	0,827	0,541	0,434	0,440
14	1,052	1,236	1,086	0,947	0,992	0,716	1,037	0,817	0,873	0,811
15	0,262	0,100	0,211	0,381	0,592	0,780	0,251	0,519	0,592	0,604
16	0,332	0,229	0,396	0,581	0,672	0,959	0,365	0,635	0,772	0,781
17	0,553	0,691	0,581	0,332	0,570	0,580	0,601	0,459	0,491	0,443
18	0,281	0,255	0,224	0,291	0,483	0,635	0,248	0,418	0,429	0,459
19	0,280	0,305	0,267	0,279	0,541	0,641	0,241	0,422	0,472	0,462
20	0,351	0,434	0,290	0,278	0,446	0,420	0,308	0,269	0,257	0,288

Fuente: Elaboración Propia

Cuadro Nº 3
Matriz de distancias continuación

	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	0,747	0,671	0,819	1,052	0,262	0,332	0,553	0,281	0,280	0,351
2	0,889	0,880	0,998	1,236	0,100	0,229	0,691	0,255	0,305	0,434
3	0,724	0,754	0,877	1,086	0,211	0,396	0,581	0,224	0,267	0,290
4	0,649	0,546	0,641	0,947	0,381	0,581	0,332	0,291	0,279	0,278
5	0,661	0,465	0,559	0,992	0,592	0,672	0,570	0,483	0,541	0,446
6	0,297	0,376	0,474	0,716	0,780	0,959	0,580	0,635	0,641	0,420
7	0,678	0,673	0,827	1,037	0,251	0,365	0,601	0,248	0,241	0,308
8	0,534	0,360	0,541	0,817	0,519	0,635	0,459	0,418	0,422	0,269
9	0,489	0,358	0,434	0,873	0,592	0,772	0,491	0,429	0,472	0,257
10	0,472	0,332	0,440	0,811	0,604	0,781	0,443	0,459	0,462	0,288
11	0,000	0,544	0,704	0,659	0,880	1,007	0,744	0,777	0,758	0,591
12	0,544	0,000	0,286	0,663	0,860	0,989	0,552	0,729	0,730	0,534
13	0,704	0,286	0,000	0,872	0,970	1,136	0,629	0,804	0,847	0,640
14	0,659	0,663	0,872	0,000	1,226	1,316	0,825	1,171	1,079	0,971
15	0,880	0,860	0,970	1,226	0,000	0,293	0,666	0,205	0,223	0,414
16	1,007	0,989	1,136	1,316	0,293	0,000	0,827	0,451	0,469	0,598
17	0,744	0,552	0,629	0,825	0,666	0,827	0,000	0,612	0,556	0,546
18	0,777	0,729	0,804	1,171	0,205	0,451	0,612	0,000	0,222	0,270
19	0,758	0,730	0,847	1,079	0,223	0,469	0,556	0,222	0,000	0,350
20	0,591	0,534	0,640	0,971	0,414	0,598	0,546	0,270	0,350	0,000

Fuente: Elaboración Propia

Agrupación media ponderada de grupos por pares:

Columna 1 y 2 = observaciones combinadas, dentro el grupo

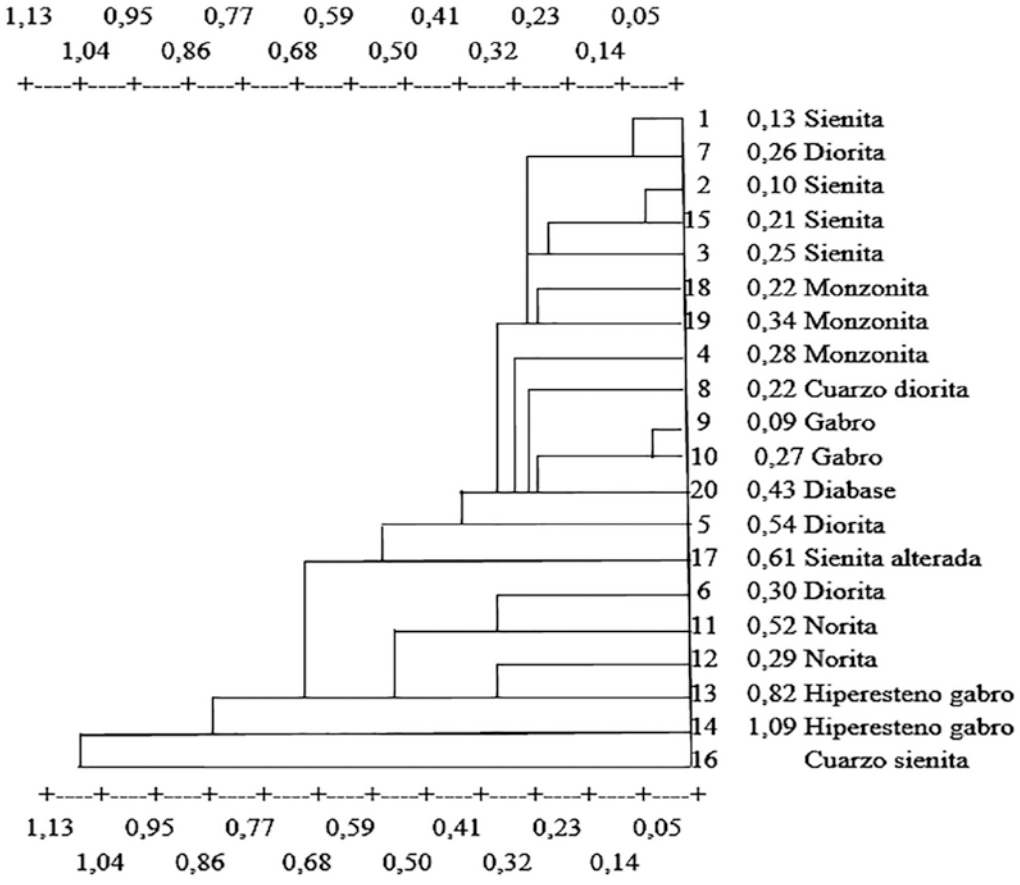
Columna 3 = nivel de correlación del agrupamiento.

Cuadro Nº 4
Agrupación media ponderada de grupos por pares

	1	2	3
1	1	7	0,1323
2	2	15	0,1003
3	9	10	0,0876
4	12	13	0,2862
5	2	3	0,2060
6	8	9	0,2235
7	18	19	0,2219
8	2	18	0,2462
9	6	11	0,2969
10	8	20	0,2709
11	1	2	0,2564
12	4	8	0,2825
13	1	4	0,3441
14	1	5	0,4283
15	1	17	0,5374
16	6	12	0,5244
17	1	6	0,6130
18	1	14	0,8201
19	1	16	1,0899

Fuente: Elaboración Propia

Figura Nº 3
Dendograma para las muestras mineralógicas



Fuente: Elaboración Propia

En este dendograma se observa los grupos que se forman entre las muestras de minerales, cuanto más cerca se encuentran, mayor es la similitud y cuanto más alejadas estén no se tendrá una similitud entre ellas, a pesar de que todas tienen cierta afinidad por el contenido de óxidos en cada muestra, lo cual conlleva a formar grupos afines. Para mayor comprensión de este dendograma, se debe leer de derecha a izquierda.

Trazando una línea vertical de corte imaginaria al nivel de similitud de aproximadamente 0,10; se tienen tres conglomerados, el primer conglomerado (extremo derecho) se compone de dos observaciones (las observaciones de filas 1 y 7 de la matriz de datos), el segundo conglomerado a la derecha se compones de dos observaciones 2 y 17 y el tercer conglomerado también se compone de dos observaciones 9 y 10.

Si se traza la línea vertical de corte imaginaria al nivel de similitud de aproximadamente 0,40 más a la izquierda, se tendrá cuatro conglomerados, y así sucesivamente. De manera que si se corta el dendograma mucho más a la izquierda, habrá menos conglomerados finales. Los tamaños de los conglomerados deben ser significativos, así en el dendograma resultante se observa tres conglomerados, generados como resultado de 6, 6, y 7 elementos, no teniendo sentido formar un conglomerado con un solo caso, como ocurre con los dos últimos.

En cuanto a la decisión sobre el número de conglomerados, no hay reglas exactas, ni rápidas, pero si existen algunos lineamientos:

a. Las consideraciones, conceptuales o prácticas pueden sugerir cierto número de conglomerados.

b. En los procedimientos de conglomeración jerárquica, puede usarse como criterio las distancias en las que se combinan los conglomerados.

Utilizando la matriz de correlación, se obtendrá el dendograma en función de las variables, luego de estandarizar la información, los cálculos y el gráfico, son los siguientes:

Cuadro Nº 5
Matriz de coeficientes de correlación

	1	2	3	4	5	6	7	8
1	1,000	0,075	0,176	-0,744	-0,759	-0,554	0,691	0,747
2	0,075	1,000	0,045	-0,332	-0,385	-0,009	0,302	-0,128
3	0,176	0,045	1,000	-0,413	-0,437	-0,149	0,434	0,412
4	-0,744	-0,332	-0,413	1,000	0,648	0,156	-0,633	-0,648
5	-0,759	-0,385	-0,437	0,648	1,000	0,410	-0,872	-0,593
6	-0,554	-0,009	-0,149	0,156	0,410	1,000	-0,579	-0,683
7	0,691	0,302	0,434	-0,633	-0,872	-0,579	1,000	0,732
8	0,747	-0,128	0,412	-0,648	-0,593	-0,683	0,732	1,000

Fuente: Elaboración Propia

Cuadro Nº 6
Matriz de distancias

	1	2	3	4	5	6	7	8
1	0,000	0,155	0,344	0,681	0,664	0,487	0,355	0,425
2	0,155	0,000	0,363	0,656	0,644	0,444	0,395	0,522
3	0,344	0,363	0,000	0,793	0,783	0,577	0,403	0,463
4	0,681	0,656	0,793	0,000	0,478	0,664	0,898	0,961
5	0,664	0,644	0,783	0,478	0,000	0,551	0,941	0,929
6	0,487	0,444	0,577	0,664	0,551	0,000	0,730	0,820
7	0,355	0,395	0,403	0,898	0,941	0,730	0,000	0,332
8	0,425	0,522	0,463	0,961	0,929	0,820	0,332	0,000

Fuente: Elaboración Propia

Agrupación media ponderada de grupos por pares:

Columna 1 y 2 = observaciones combinadas, dentro del grupo

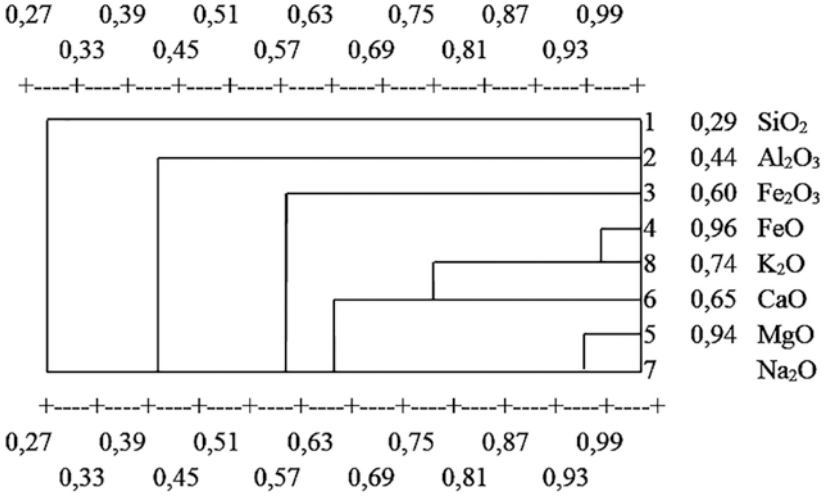
Columna 3 = nivel de correlación del agrupamiento

Cuadro Nº 7
Agrupación media ponderada de grupos por pares

	1	2	3
1	4	8	0,9607
2	5	7	0,9412
3	4	6	0,7420
4	4	5	0,6500
5	3	4	0,5978
6	2	3	0,4405
7	1	2	0,2921

Fuente: Elaboración Propia

Figura Nº 4
Dendograma para las variables



Fuente: Elaboración Propia

Los valores a lo largo del eje X representan las similitudes.

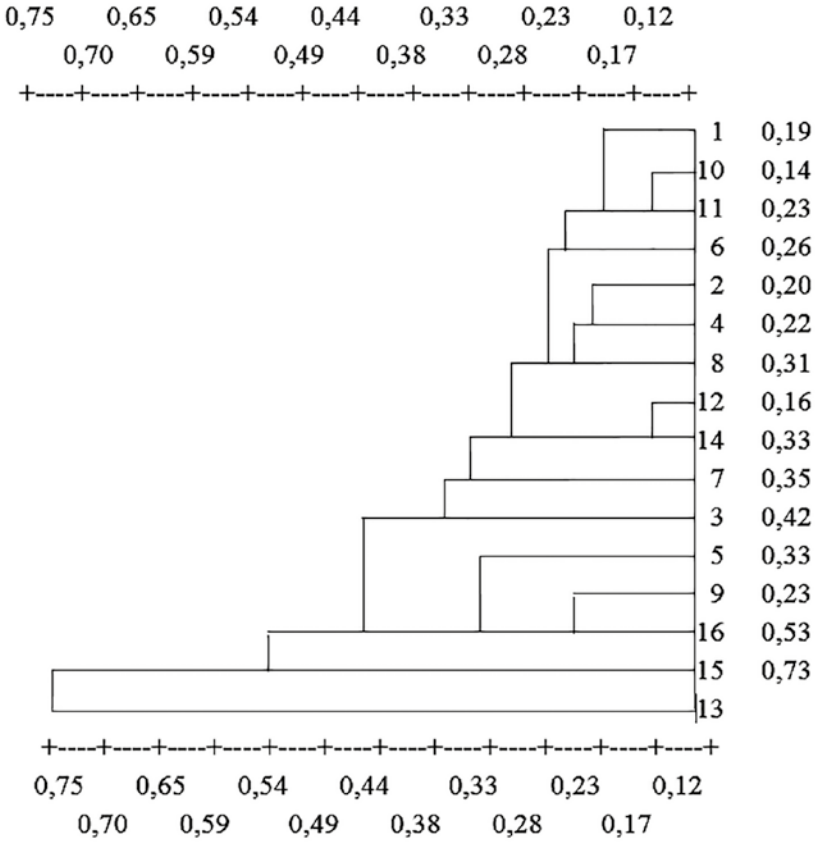
En este dendograma se observa los grupos que se forman entre las variables de los óxidos.

Problema. En base a la fuente de datos obtenida de la Sección de Bioestadística del S.S.U. de la UMSA para fines ilustrativos, se tienen algunos datos del análisis de sangre de diez y seis muestras con quince variables cada una, de pacientes que tienen diabetes.

La matriz de datos observados es de 16 por 15, y siendo los cálculos semejantes a los del problema anterior debido a que se utilizo el mismo programa computacional, solo se presentará los dendogramas respectivos de muestras y variables como se demuestra a continuación:

Figura Nº 5

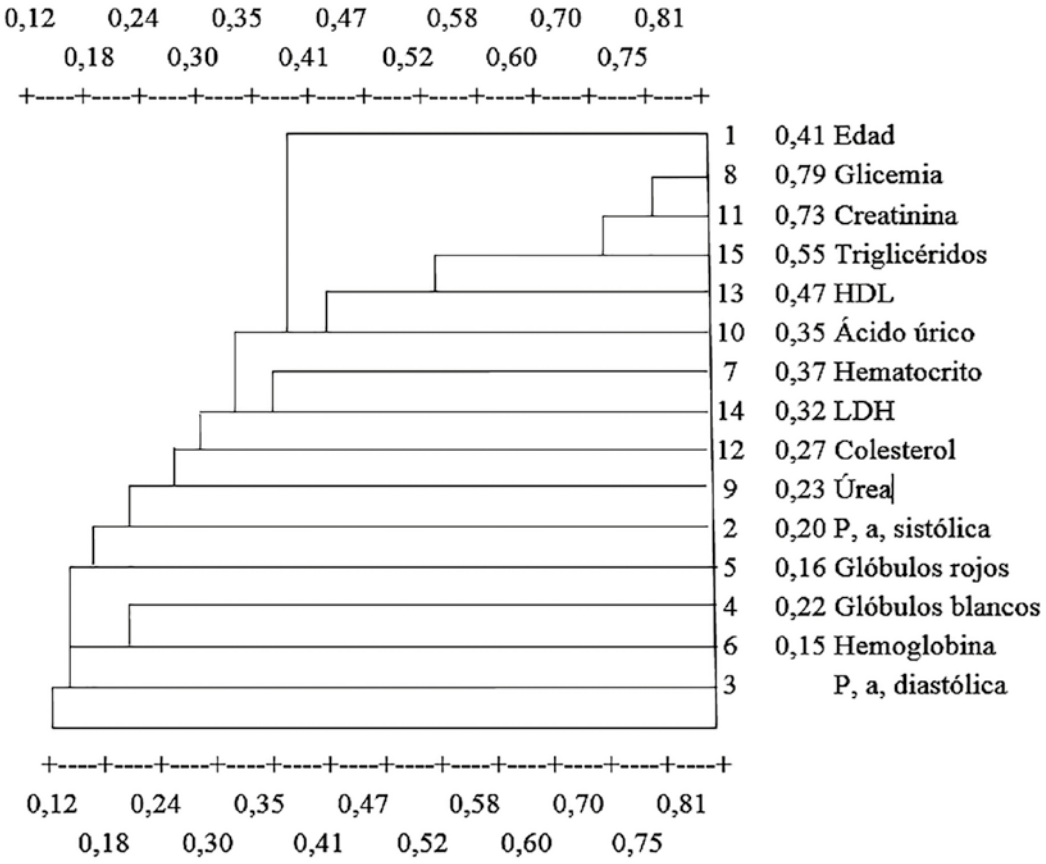
Dendograma para las variables



Fuente: Elaboración Propia

Figura Nº 6

Dendograma para las muestras de enfermos con diabetes



Fuente: Elaboración Propia

Colaboración.

Ing. Guillermo Salinas Reguerin
Ingeniero Comercial. Universidad Loyola
La Paz – Bolivia (2004)
Unidad de Bioestadística, SSU, UMSA

BIBLIOGRAFÍA

1. BENZÉCRI J. P. & Collaborateurs. L´Analyse Des Donnees, La Taxinomia. Ed. Dunod, Paris, Francia, 1973, pp. 615 – XIII.

2. DAGNELIE Pierre, Théorie et Méthodes Statistiques. (volume 1). Les Presses Agronomiques de Gembloux, Gembloux, Belgica, 1973 (2da. Edición), pp. 378 – IX.

3. DAGNELIE Pierre, Analyse Statistique á Plusieurs Variables. Les Presses Agronomiques de Gembloux, Gembloux, Belgica, 1975 (2da. Edición), pp. 362 – XIV.

4. DAVIS C. John. Statistics and Data Analysis in Geology. John Wiley & Sons, Inc., Toronto, Canada,1973, pp. 550 – VII.

5. KLOCKMANN F. y RAMDOHR P., Tratado de Mineralogia (Versión en aleman).

Editorial Gustavo Gili, Barcelona, España, 1961, pp. 736 – IX.

6. KRUMBEIN W. C. and GRAYBILL A. Franklin, An Introduction to Statistical Models in Geology. McGraw-Hill Book Company, Toronto,Canada, 1965, pp. 475 – XV.

7. RUIZ Aranibar Gustavo. Factores que Inciden en el Rendimiento Académico y Evaluación Docente. U.A.G.R.M. Santa Cruz – Bolivia. 2010, pp. 190 – IX.

8. RUIZ Aranibar Gustavo[2] . Libreria Científica de Programas Informáticos, La Paz -Bolivia.

" Todo trabajo de investigación es el resultado del esfuerzo que se realiza con: orden, voluntad, paciencia y estudio constante, para dar a conocer a nuestros semejantes, sabiendo que el final de los trabajos de investigación que se realizan, son el comienzo de las investigaciones que otros proseguirán en forma más profunda"

Dr. Cs Gustavo Ruiz Aranibar

[2] Calle 20 y Av. Ballivian, Nº 8035, Calacoto, La Paz – Bolivia, Tel. 591-22772162 Cel, 67111778
gustavoruiz432@hotmail.com.bo ruizaranibargustavo@gmail.com.bo Blog: Gustavo Ruiz Aranibar